Can we really trust generative models in healthcare? A systematic review of uncertainty quantification in generative AI for medical imaging

Computer Methods and Programs in Biomedicine 285 (2026) 109560

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

# Can we really trust generative models in healthcare? A systematic review of uncertainty quantification in generative AI for medical imaging

Alen Shahini [a], Silvia Seoni [a], Alessandra Manzin [b], Martina Oria [a,b], Anjan Gudigar [c], U Raghavendra [c], Abdulkadir Sengur [d], Rajendra Acharya [e,f], Massimo Salvi [a,*]

[a] Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy
[b] National Metrology Institute of Italy (INRiM), Turin, Italy
[c] Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India
[d] Electrical-Electronics Engineering Department, Technology Faculty, Firat University, Elazig, Turkey
[e] Centre for Health Research, University of Southern Queensland, Australia
[f] School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, Australia

ARTICLE INFO

*Keywords:*
Uncertainty quantification
Medical imaging
Generative models
Monte Carlo methods
Deep learning
Systematic review

ABSTRACT

*Background and Objective*: Generative AI is transforming medical imaging research through synthesis, enhancement, and reconstruction of clinical images. While these advances show promise in addressing data scarcity and supporting diagnostics, clinical adoption remains limited due to challenges in assessing the trustworthiness of generated content. This study aims to systematically evaluate the integration of uncertainty quantification (UQ) methods within generative models for medical imaging to enhance result reliability. *Methods*: A systematic review following PRISMA guidelines was conducted, analyzing studies from January 2018 to December 2025 that combined generative models with UQ techniques in medical imaging. The search strategy covered major medical and computer science databases, with studies evaluated against predefined inclusion criteria focusing on implementation methodology and performance metrics. *Results*: From the analysis of 41 eligible studies, 29 focused on radiology, 8 on microscopy, and 4 on optical coherence tomography. Across this heterogeneous body of evidence, integrating UQ was frequently associated with improved performance or with more informative reliability assessment, including reported gains in reconstruction quality, segmentation accuracy, and anomaly detection. Notably, 56% of studies (n=23) were published in 2025, indicating rapid field growth. *Conclusions*: UQ integration represents a crucial advancement toward trustworthy generative AI systems in medical imaging. Key priorities identified include standardizing uncertainty metrics, developing computationally efficient frameworks, and embedding uncertainty awareness within generation processes. These findings suggest that UQ methods can enhance the clinical reliability of generative AI applications in medical imaging.

## 1. Introduction

Generative artificial intelligence (GenAI) models have advanced to the point where they can produce synthetic medical images that are virtually indistinguishable from authentic clinical scans. These models leverage deep learning (DL) architectures such as Generative Adversarial Networks (GANs) [1], Variational Autoencoders (VAEs) [2], diffusion models (DMs) [3], and transformers [4] to generate synthetic medical images that can augment limited datasets, enhance image quality, and facilitate novel diagnostic approaches.

The applications of GenAI in medical imaging span multiple modalities and clinical scenarios (Fig. 1). Image reconstruction and denoising are among the most established applications, where generative models address the fundamental challenge of acquiring high-quality medical images while minimizing radiation exposure [5] or scan time [6]. Cross-modal synthesis enables the generation of one imaging modality from another, potentially reducing the need for multiple scans and associated costs [6,7]. Domain translation allows models trained on one scanner or protocol to be applied to different acquisition settings, addressing the heterogeneity challenges inherent in medical imaging [8]. Additionally, data augmentation and simulation [9] through generative models provides valuable synthetic training data to

* Corresponding author.
*E-mail address:* massimo.salvi@polito.it (M. Salvi).

https://doi.org/10.1016/j.cmpb.2026.109560
Received 26 December 2025; Received in revised form 17 June 2026; Accepted 20 July 2026
Available online 21 July 2026

overcome the limited availability of annotated medical datasets [10]. Moreover, generative models can be used in explainability tasks, providing a reference basis for improving the transparency and interpretability of AI systems [11].

Despite their remarkable potential, GenAI models face significant limitations that challenge their clinical adoption. First, the evaluation of generated outputs remains problematic due to the absence of robust no-reference metrics and inherent model instabilities [18]. Traditional image quality metrics may not adequately capture clinically relevant features, while the stochastic nature of many generative models leads to variability in outputs that can be difficult to assess without ground truth references [19,20].

Another major challenge is distinguishing between genuine medical data and out-of-distribution (OOD) synthetic content. Generative models can generate realistic-looking artifacts or features that appear clinically plausible but do not exist in the original anatomy [21], which is especially concerning when these hallucinations may influence diagnostic decisions [22]. The difficulty in recognizing when models encounter data that differ significantly from their training distribution contributes to this reliability issue. Model instability [23] represents another critical concern, as small perturbations in input data can lead to dramatically different generated outputs. This sensitivity undermines the reproducibility and consistency required for clinical applications, where reliable and predictable behavior is essential for patient safety and diagnostic accuracy.

In the last decade, uncertainty quantification (UQ) has emerged as a promising solution to address these limitations and build trustworthy AI systems for medical imaging [24]. UQ provides a framework for models to communicate their confidence levels and identify regions or scenarios where their predictions may be unreliable [25]. This capability is essential for clinical adoption, as it enables healthcare professionals to make informed decisions about when to trust AI-generated content and when additional verification may be necessary.

While UQ techniques have been extensively studied and adopted in medical image analysis [26], their application has predominantly focused on classification [27] and segmentation tasks [28], with limited exploration in generative modeling scenarios. Classification tasks typically involve discrete outputs with associated confidence scores, where UQ helps assess prediction reliability through model calibration techniques [29]. Similarly, segmentation tasks produce discrete masks with uncertainty estimates that can guide clinician attention to regions of lower confidence [30].

By contrast, the application of UQ to GenAI models in medical imaging has only recently begun to receive broader attention [31]. This emerging body of work suggests that the field is evolving into a distinct and clinically relevant research area. A dedicated review is therefore needed to organize the available evidence, compare methodological approaches, and identify the open challenges that may affect the safe and reliable clinical translation of generative models. This increasing interest is driven by the rising complexity of generative architectures, stronger clinical demands for reliable AI systems, and the recognition that traditional validation approaches may be insufficient for generative tasks [32,33]. Unlike classification or segmentation, generative tasks produce continuous outputs that must be evaluated across complex, high-dimensional image spaces, requiring more specialized approaches to uncertainty estimation and validation [34].

The integration of UQ into generative medical imaging represents a crucial step toward building AI systems that can support clinical workflows while maintaining appropriate levels of confidence and enabling meaningful human oversight.

### 1.1. Research questions and main contribution

The landscape of UQ in generative medical imaging presents several critical unanswered questions that this systematic review aims to address. The primary research question of our study is: "What are the current approaches for quantifying uncertainty in medical image generation, and how do they perform across different imaging modalities and clinical tasks?"

Building upon this primary focus, several secondary questions emerge that are equally important for advancing the field:

- What are the most popular UQ techniques in generative medical imaging, and what factors influence their adoption?
- How does incorporating UQ affect the performance of generative models across various evaluation metrics?
- What are the computational trade-offs between different UQ approaches, and how do they impact clinical feasibility?
- What are the main challenges and opportunities for clinical application?

This systematic review is the first to provide a comprehensive analysis of UQ methods applied to GenAI models in medical imaging. While UQ has been extensively studied in classification and segmentation tasks, its application to generative medical imaging is a new and rapidly evolving field that has yet to be systematically reviewed. We present a systematic taxonomy of UQ approaches designed for generative medical imaging, categorizing them as probabilistic, Monte Carlo (MC) methods, Bayesian inference (BI), and ensemble techniques, or non-probabilistic methods, including interval analysis (IA) and conformal prediction

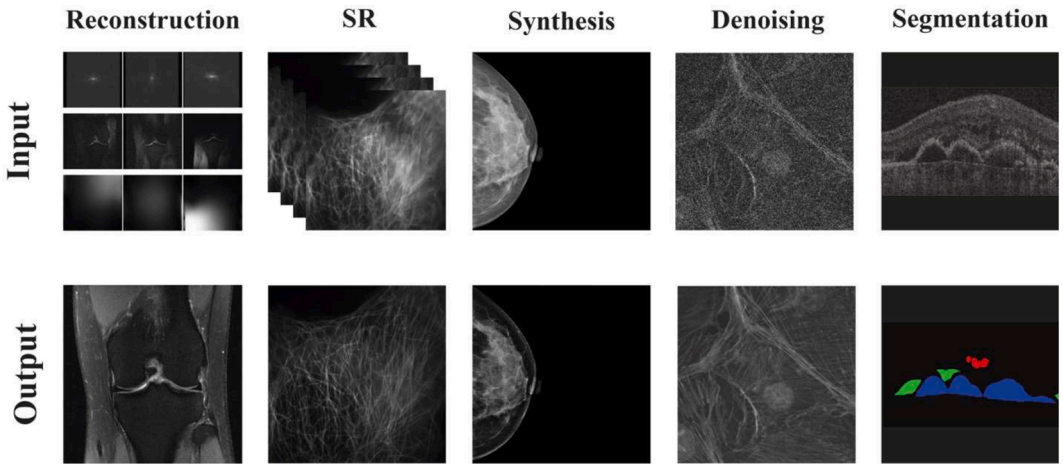


**Fig. 1.** Example of applications of GenAI models in medical imaging. The figure illustrates the diverse range of tasks where GenAI models are currently employed, including image reconstruction [12], super-resolution (SR) imaging [13], cross-domain translation [14,15], denoising [16] and image segmentation [17].

(CP). Our cross-modal analysis examines the implementation of UQ across a broad spectrum of medical imaging modalities and applications, from radiological imaging reconstruction to microscopic analysis and diagnostic assessment.

In this paper, we systematically assess the practical aspects of UQ implementation, such as computational requirements, validation strategies, and calibration techniques. Finally, we identify key challenges and future directions for developing trustworthy GenAI in medical imaging. We connect current technical limitations with emerging opportunities in uncertainty standardization, calibration methodologies, and uncertainty-guided generative frameworks.

## 2. Methods

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [35] were closely followed to select the most relevant articles on UQ methods applied to medical imaging, using generative DL models.

### 2.1. Related reviews

While UQ is highly relevant in medical imaging, its application to GenAI remains relatively unexplored in the literature. To the best of our knowledge, no review articles have been published specifically addressing UQ in generative models for medical imaging. The existing literature includes reviews that address related but distinct topics:

- Osuala et al. [36] provide comprehensive coverage of GANs applied to medical cancer imaging. Although this review does not specifically analyze the UQ techniques used in generative models, it dedicates a section to discussing uncertainty as a potential evaluation metric for GANs. In particular, it highlights how predictive uncertainty can correlate with reconstruction error and serve as an indicator of image realism, model reliability, and adherence to conditioning inputs in generative tasks.
- Bansal et al. [37] and Tan et al. [38] examine generative models applied to individual imaging modalities. The former explores the utilization of GenAI for diabetic retinopathy detection and its potential to enhance AI solutions in healthcare without discussing the possibility of integrating UQ techniques. The latter focuses on GenAI applied to breast cancer diagnosis and treatment, but it does not address the topic of UQ and its role in evaluating model reliability.
- Faghani et al. [39] discuss UQ only briefly within broader discussions on DL in radiological imaging, without explicitly examining its application to generative models.
- Huang et al. [40] concentrate on UQ methodologies across several medical imaging tasks, reconstruction, segmentation, detection, prediction, rather than exploring generative modeling paradigms themselves. Generative approaches are only cited incidentally to illustrate specific applications, without a dedicated discussion of uncertainty modeling within GenAI frameworks.
- Seoni et al. [25] present a comprehensive systematic review of UQ in healthcare applications, covering both machine learning and DL approaches across imaging and signal domains. However, the review does not explicitly address GenAI models or the specific challenges of UQ within generative frameworks.

Taken together, these reviews confirm the relevance of UQ in medical imaging, but they do not provide a focused analysis of how uncertainty is modeled, evaluated, and interpreted in GenAI. This gap supports the need for a dedicated review specifically centered on UQ in GenAI for medical imaging.

### 2.2. Literature search strategy

This review focuses on articles published between January 2018 and December 2025 that investigated and applied UQ methods to generative DL models in medical imaging. The decision to select articles from 2018 onwards is explained by recent advances and trends in the field, ensuring a comprehensive and detailed analysis of the latest developments and capturing recent innovations and methodologies. The search process was conducted in January 2026 across several databases, including Google Scholar, Scopus, PubMed, and the Institute of Electrical and Electronics Engineers (IEEE) databases. The search strategy implemented a Boolean approach, combining keywords related to:

- The generative model based on artificial intelligence, “artificial intelligence”, “deep learning”, “generative model”, “GAN”, “Generative Adversarial Networks”, “Diffusion model”, “VAE”, “Variational Autoencoder”, “Transformer”, “Bayesian Neural Network”, “Normalizing flow”.
- UQ approaches, “uncertainty”, “perturbation”, “uncertainty estimation”, uncertainty quantification”, “stochasticity”, aleatoric uncertainty”, “epistemic uncertainty”.
- Medical imaging modality, “MRI”, “Magnetic Resonance”, “CT”, “Computed Tomography”, Mammography”, “PET”, “Positron Emission Tomography”, “SPECT”, “Single-photon Emission Computed Tomography”, “Digital Pathology”, “Fluorescence”, “OCT”, “Optical Coherence Tomography”, “OCTA”, “Optical Coherence Tomography Angiography”, “Ultrasound”, “Dermatology”, “Dermoscopy”.

The initial search yielded 1,217 papers. After eliminating duplicate papers (n=234) and excluding books, abstracts, conference proceedings, non-English articles, and articles not published in peer review journals, the remaining 721 studies underwent title and abstract assessment. Potentially relevant studies then underwent full-text assessment, including evaluation of both quality and relevance. Study classification, categorization, and initial quality assessment were carried out by one co-author (A.S.). Borderline cases, particularly studies with scores close to the inclusion threshold, were further discussed with a second co-author (S.S.) until consensus was reached. Each study was evaluated across five methodological dimensions, with each dimension scored from 0 to 3 points, for a maximum total score of 15 points.

The assessment evaluated five key methodological dimensions:

1. Study design quality (0-3 points): Appropriateness and rigor of the study design; clarity of the medical imaging task and research objective; adequacy of dataset size and train/validation/test separation; use of internal, external, or multi-center validation; consideration of clinical or technical relevance.
2. Medical imaging data and application assessment (0-3 points): Relevance and quality of the medical imaging data; clarity of imaging modality, anatomical region, disease context, and clinical task; quality of annotations or reference standards; adequacy of preprocessing, harmonization, and data-splitting procedures; consideration of scanner, site, and population variability.
3. Generative model assessment (0-3 points): Clear description and methodological appropriateness of the generative DL model; suitability of the model for the imaging task; quality of model training and validation; comparison with relevant baselines or alternative methods.
4. UQ and statistical validation (0-3 points): Explicit definition of the uncertainty being quantified, including epistemic, aleatoric, or predictive uncertainty; appropriateness of the UQ method; use of calibration, reliability, robustness, or sensitivity analyses; evaluation of whether uncertainty estimates were clinically or methodologically meaningful; transparent handling of statistical assumptions and missing data.
5. Reporting quality and reproducibility (0-3 points): Completeness and transparency of methodological reporting following relevant guidelines where applicable; clarity in the presentation of datasets, model architecture, training details, UQ procedure, and results; discussion

of limitations and potential sources of bias; availability of code, data, pretrained models, or sufficient implementation details to support reproducibility.

A threshold of 8/15 points was selected to ensure methodological quality. Studies with major methodological flaws, unclear UQ procedures, or insufficient reporting were excluded. After screening over 1000 results published over the past years based on defined inclusion criteria and quality assessment, 41 high-impact studies were selected through a rigorous process adhering to PRISMA guidelines (Fig. 2). Detailed quality assessment scores for each included study are reported in the Appendix Tables A2 and A3.

### 2.3. Uncertainty assessment in GenAI models

In GenAI, UQ refers to the analysis and interpretation of variability in the distribution of images that a model associates with a given input. The generation of multiple outputs may depend on the intrinsic probabilistic nature of the generative model [6,41]. However, it can also be achieved by implementing alternative strategies within generative models where uncertainty is not part of their original formulation [42].

Classification and segmentation tasks produce discrete outputs (classes or segmentation masks) with associated confidence scores, where UQ helps assess prediction reliability through model calibration [25]. In contrast, generative tasks produce continuous outputs (synthetic images), where UQ aims to analyze and quantify the variability among multiple plausible outputs that the model can generate from the same input. This analysis provides clinicians with a measure of confidence in the generated images by revealing how consistent or variable these outputs are across different regions or features.

It is common to distinguish between two sources of uncertainty: aleatoric and epistemic. Aleatoric uncertainty is intrinsic to the data. It occurs when data is mislabeled, noisy, or uninformative [39]. On the other hand, epistemic uncertainty arises from limitations in the model's knowledge, resulting from insufficient, biased, or unrepresentative training data, or design errors in the model itself. While epistemic uncertainty can be reduced by improving the quality of training data or the architecture, aleatoric uncertainty is irreducible in most cases [43].

The probabilistic/non-probabilistic taxonomy adopted in this review builds on the framework proposed by Huang et al. [40], but its application to generative medical imaging is not a direct transfer from discriminative models. In classification or segmentation, uncertainty is usually associated with a single output, such as a class probability or a label map. In GenAI models, instead, the output is a distribution of plausible images, and uncertainty may arise from latent-variable sampling, diffusion trajectories, model parameters, conditioning inputs, or adversarial instability. Therefore, aleatoric and epistemic uncertainty may be expressed not only as confidence scores, but also as sample variability, pixel-wise uncertainty maps, prediction intervals, or calibrated risk bounds. In this context, qualitative uncertainty assessment refers to visual or descriptive evaluation of uncertainty patterns, for example whether uncertainty maps localize artifacts, boundaries, or OOD regions, whereas quantitative uncertainty assessment refers to numerical evaluation using calibration, correlation, coverage, or interval-based metrics. Table 1 summarizes the taxonomy and terminology used throughout this review.

## 3. Results

The following section shows the results obtained from the literature

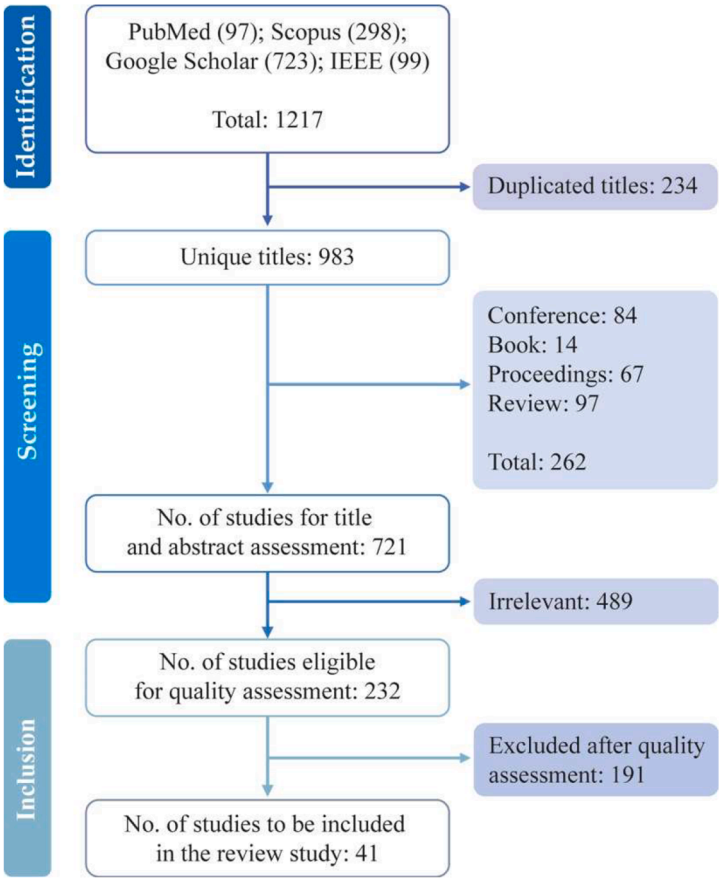


**Fig. 2.** Selection of relevant articles based on PRISMA guidelines. The flowchart shows the systematic review process from 1,217 initially identified papers to 41 high-impact studies selected for final analysis after multiple screening stages.

**Table 1**
Taxonomy and acronyms of UQ concepts used throughout this review. UQ methods are classified according to the probabilistic/non-probabilistic taxonomy adopted in the manuscript, while UQ metrics and outputs reported by the reviewed studies are grouped by category and detailed in Appendix Tables A2 and A3.

| Category | Abbreviation | Definition / Examples |
|---|---|---|
| Probabilistic UQ method | BI | Bayesian Inference |
| | DE | Deep Ensemble |
| | MCD | Monte Carlo Dropout |
| | MCMC | Markov Chain Monte Carlo |
| | MCS | Monte Carlo Sampling |
| | TTE | Test-Time Ensemble |
| Non-probabilistic UQ method | CP | Conformal Prediction |
| | IA | Interval Analysis |
| Uncertainty metrics | CM | Confusion Matrix: Accuracy; Area Under the Curve (AUC); F1-score; Precision; Precision Recall AUC (PR AUC); Recall; Receiver Operating Characteristic AUC (ROC AUC); Sensitivity; Specificity. |
| | Confidence | Confidence: Calibration of confidence; confidence map; confidence score; Expected Calibration Error (ECE); Uncertainty Calibration Error (UCE). |
| | IQ | Image Quality: Contrast-to-Noise Ratio (CNR); Mean Absolute Error (MAE); Mean Absorbed Dose (MAD); Mean Squared Error (MSE); Peak Signal-to-Noise Ratio (PSNR); Root Mean Squared Error (RMSE); Structural Similarity Index Measure (SSIM). |
| | Segm | Segmentation: Dice Similarity Coefficient (DSC); Intersection over Union (IoU). |
| | UQ Map | Uncertainty Quantification Map: Standard Deviation (Std. Dev.); Variance. |
| | UQ Score | Uncertainty Quantification Score: Calibration of uncertainty; conditional uncertainty; conformal and interval guarantees; conformal coverage; conformal set size; coverage; coverage slack; empirical risk; Kullback-Leibler divergence (KL); marginal uncertainty; mean interval size; median interval size; Negative Log-Likelihood (NLL); Pearson correlation; prediction interval size; prediction-satisfies-risk; risk bounds; Spearman correlation; uncertainty estimates. |

review. Methods for quantifying uncertainty are categorized into two main types: probabilistic and non-probabilistic methods.

### 3.1. Probabilistic UQ methods

Probabilistic methods are approaches that represent uncertainty through probability distributions, enabling the quantification of prediction variability through statistical measures. These methods aim to characterize the distribution of plausible outputs. Our literature review revealed three main probabilistic approaches: MC methods, BI, and Ensemble methods (Fig. 3).

#### 3.1.1. Monte Carlo methods

MC methods [44] involve generating random samples by attempting to approximate the probability distributions of the generated data. A larger number of simulated samples allows for a closer approximation to the true distribution, enabling the accurate quantification of uncertainty. Among the reviewed articles, we identified three main MC approaches:

- Monte Carlo sampling (MCS): draws independent random samples to estimate the expected output distribution.
- Monte Carlo dropout (MCD): generates multiple predictions by randomly activating dropout layers during inference, rather than only during training as is typically done.
- Markov Chain Monte Carlo (MCMC): constructs a sequence of dependent samples that gradually converge to the target probability distribution.

For MRI reconstructions, Edupuganti et al. [41] leveraged VAE and created pixel variance maps using MCS. They observed that uncertainty diminished when adversarial losses decreased, and multiple recurrent blocks were employed. Through posterior sampling of a DM, Chu et al. [45] showed that higher acceleration rates for MRI reconstruction maintained a low level of uncertainty. Sloun et al. [46] devised an uncertainty-aware DM that actively reduced uncertainty and maximized diagnostic value in a sequence of experiments during ultrasound (US) image reconstruction. Compared to random sampling, the uncertainty-driven (maximum information sampling) strategy yielded consistently more efficient and more accurate reconstructions. Stevens et al. [47] used diffusion posterior sampling for high-volume-rate 3D cardiac US reconstruction, estimating epistemic uncertainty from the variance of multiple posterior samples, which increased under stronger subsampling and around synthetic OOD anomalies.

MCS was widely used for UQ in Denoising Diffusion Probabilistic Models (DDPM). Dou et al. [48] presented a DDPM for standard plane localization in 3D US, reaching 90% accuracy for detecting anomalies by thresholding the quantified uncertainty computed as the variance of multiple inferences. Dong et al. [7] introduced a DDPM for MRI and estimated uncertainty with standard deviation maps of 50 samples. Zhang et al. [49] introduced a diffusion-based method for MRI quantitative susceptibility mapping reconstruction, using multiple diffusion samples to derive voxel-wise standard-deviation uncertainty maps and assess reconstruction ambiguity. Yu et al. [50] showed that 3D DDPM exhibited lower uncertainty than 2D DDPM for PET image denoising, by using uncertainty maps of 20 samples. Gong et al. [6] proposed different DDPM-based methods for PET image denoising and demonstrated that the addition of MR images increased the correlation between Mean Squared Error (MSE) and the generated predictive uncertainty, reduced uncertainty and improved generative performance. Using a transformer-based model, Mansouri et al. [51] demonstrated that MCS was the best approach for scattering correction in quantitative 90Y Single-photon Emission Computed Tomography (SPECT) imaging, showing an improvement of 8.62 in Peak Signal-to-Noise Ratio (PSNR) and reaching mean absorbed data closer to the reference.

For tumor assessment, Rafael-Palou et al. [52] proposed a VAE that predicted lung nodule growth, future size, and semantic appearance from an initial Computed Tomography (CT) image. Uncertainty-aware models achieved better performance in predicting and segmenting cancer. Xing et al. [53] proposed a MCS DM for 3D MRI and CT image segmentation to generate uncertainty maps and defined an uncertainty-guided loss to improve the segmentation of uncertain pixels. The proposed module achieved the lowest uncertainty value at each inference step, better segmentation performance, and showed more significant changes in uncertainty maps between adjacent inference steps compared to the deep ensemble (DE) and MCD. Liu et al. [54] developed a Bayesian Neural Network (BNN) for structured illumination microscopy, which enhanced the reconstruction of densely labeled structures while enabling the quantification of super-resolution (SR) uncertainty modelling aleatoric uncertainty with a Gaussian distribution and epistemic uncertainty with MCS. High uncertainty values tended to align with super-resolved structures displaying conspicuous errors. Aleatoric uncertainty more accurately reflected the SR error distributions compared with epistemic uncertainty.

The integration of dropout sampling [55] during inference enables approximation of model posterior distributions through neuronal deactivation. This approach proved valuable in MRI synthesis, where Finck et al. [56] proposed the usage of uncertainty maps to enhance clinical utility and trustworthiness of a GAN model, showing they can be used to distinguish true positive lesions from false positive hyperintensities. Similarly, Tolpadi et al. [42] used uncertainty maps for UQ

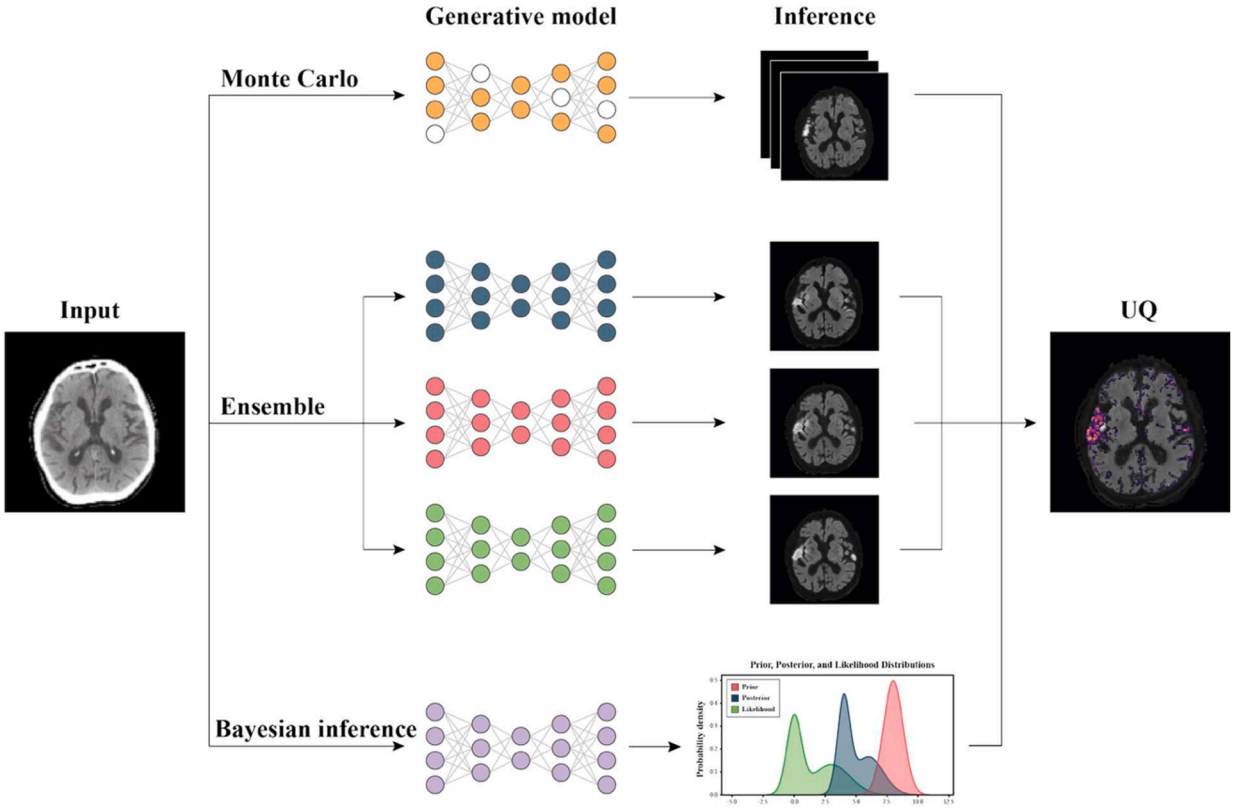


**Fig. 3.** Overview of probabilistic UQ methods in GenAI models. The diagram categorizes the main probabilistic approaches into three branches: MC methods, ensemble methods, and BI techniques. Each approach offers, distinct strategies for capturing and quantifying uncertainty in generative model outputs through statistical and probabilistic frameworks.

of a PatchGAN, developed to synthesize post-contrast wrist MRI from non-contrast scans. Hu et al. [57] used lesion and background scores to quantify the uncertainty of a CT/PET segmentation conditional GAN (cGAN). The adversarial generative model with the combination of CT and PET images reached better performance and lower uncertainty than the U-Net, both in the lesion and in the background. Jebril et al. [58] developed a MCD Bayesian U-Net to segment anomalies in Optical Coherence Tomography Angiography (OCTA) images. The epistemic uncertainty approach achieved higher Dice, Intersection over Union (IoU), and sensitivity scores. Galapon et al. [8] investigated the use of MCD-based uncertainty maps as predictors for verifying proton dose calculations on MRI-derived synthetic CT images in the context of online adaptive proton therapy. The CT images were generated using two DL architectures, a deep convolutional neural network (DCNN) and a cycle-consistent generative adversarial network (cycleGAN) [59], to compare the performance of different model architectures in terms of uncertainty estimation and dosimetric accuracy.

Laves et al. [34] presented a Bayesian model for inverse problems in medical imaging, evaluating it against MCD on four different tasks across a variety of modalities: CT reconstruction, Optical Coherence Tomography (OCT) denoising, dermoscopy inpainting, and MRI SR. For CT reconstruction, the uncertainty for all methods was high at intensity discontinuities, which correlated well with the reconstruction error. For denoising, the model provided well-calibrated predictive uncertainty maps by means of uncertainty calibration error. MCD exhibited less severe overfitting and greater robustness towards eventual hallucinations. For inpainting, the model predicted uncertainty was high in regions corresponding to hair, as desired and low in the region of the chloasma. MCD tended to overly smooth out the inpainted areas. For the SR tasks, both the proposed model and MCD exhibited high uncertainties in edge regions as needed. Tanno et al. [60] accounted for uncertainty in MRI SR through a heteroscedastic noise model. Modelling the combination of both aleatoric and epistemic uncertainty improved overall performance. Modelling aleatoric uncertainty improved robustness to outliers, while modelling epistemic uncertainty defended against overfitting. Uncertainty revealed pathological structures absent from the training data, serving as a metric of reliability.

For medical image reconstruction tasks, Huang et al. [5] introduced the MC Arbitrary Scan Masking (MC-ASM) mechanism for UQ. Rather than randomly deactivating neurons as in traditional MCD, MC-ASM introduces structured randomness by masking redundant scan information. This approach eliminates the need to tune dropout rates and alleviates the performance degradation often encountered when applying dropout to low-level imaging tasks, while still providing reliable estimates of predictive uncertainty.

Recently, the combination of Bayesian DL and MCD was extensively adopted for UQ in fluorescence microscopy for SR tasks. Chang et al. [61] used UQ as an attention mechanism, directing the BNN's focus towards areas of high uncertainty to improve SR reconstruction in structured illumination microscopy. The uncertainty guided MCD method increased the PSNR by 1.7 dB and the Structural Similarity Index Measure (SSIM) by 6%. To increase the reliability of BNN-based reconstruction in fluorescence imaging, Lin et al. [62] provided uncertainty and confidence maps through MCD, allowing biologists to identify potential imaging errors simultaneously and quantitatively without any reference. It achieved a significant uncertainty reduction compared to current methods, overcoming the constraints of unreliability in the practical application of structured illumination microscopy. Qiao et al. [63] proposed an uncertainty-aware Bayesian model, designing an expected calibration error minimization framework that estimates inference confidence. Confidence maps were highly correlated with error maps. Through an iterative fine-tuning process, the uncertainty calibration error was reduced by more than fivefold, from 0.1 to 0.018, while maintaining the output fidelity as measured by PSNR.

MCMC [64,65] constructs a Markov chain to generate dependent samples that approximate a target distribution. Through iterative sampling, where each draw depends only on the previous state, the chain gradually converges to the underlying posterior distribution. It allows sampling from complex and high-dimensional distributions that are otherwise intractable to sample from directly. Chung et al. [66] proposed a denoising method for MRI based on score-based diffusion sampling. Uncertainty in standard deviation maps increased with higher levels of denoising. Luo et al. [67] developed a DM that enabled efficient sampling from learned probability distributions for MRI reconstruction. High undersampling implied high uncertainty in the variance maps. Zach et al. [68] reached a PSNR improvement of 2.01 for MRI reconstruction with a Bayesian energy-based model (EBM). Uncertainty maps provided the clinician with additional information about the inherent uncertainty in the reconstruction, showing high variance in small anatomic structures. Liu et al. [69] aimed to accelerate MRI acquisition by a score-based DM. A Bayesian Convolutional Neural Network (BCNN) was used to retrieve a score to train the DM. The uncertainty-aware proposed method improved MRI reconstruction, yielding an increase in PSNR and SSIM of 2.6% and 2.3%, respectively.

#### 3.1.2. Bayesian inference

BI provides a probabilistic framework for updating knowledge about unknown quantities based on observed data. This approach combines prior information with data likelihood to generate posterior distributions representing uncertainty-aware estimation. BI allows the generative model to capture and express uncertainty. Repeated inference produces a distribution of plausible outputs that is consistent with both the observed data and prior knowledge.

Recent applications in medical imaging demonstrate the versatility of Bayesian methods. Levital et al. [70] proposed a non-parametric Bayesian DL framework for low-dose PET reconstruction, using stochastic gradient Langevin dynamics to generate voxel-wise uncertainty maps that correlated with dose reduction and improved lesion reconstruction compared with MCD. For multimodal registration, Chen et al. [71] presented a Transformer architecture whose uncertainty maps showed strong correlation with actual registration errors, achieving optimal calibration through expected model error estimation. Similarly, Liu et al. [72] employed an uncertainty-aware Transformer for OCT image segmentation. The addition of uncertainty within the attention module improved Dice Similarity Coefficient (DSC) segmentation performance by 3.8%.

Approximate BI in the form of variational dropout was also employed in [60], to provide an estimation of epistemic uncertainty in the application of SR methods to MRI-based neuroimaging. BI was jointly combined with a heteroscedastic noise model for aleatoric UQ, to quantify predictive uncertainty over the system output. The results on various upsampling methods and datasets demonstrated how the joint modelling of uncertainty improves reconstruction accuracy. Modelling aleatoric uncertainty improved robustness to outliers, while modelling epistemic uncertainty defended against overfitting. Uncertainty revealed pathological structures absent from the training data, serving as a metric of reliability.

#### 3.1.3. Ensemble methods

Ensemble methods leverage multiple trained architectures or variants of the same architecture to enable UQ through analysis of their collective outputs. This approach generates deterministic results from individual models while capturing uncertainty through disagreement between ensemble members.

In fluorescence microscopy, Weigert et al. [73] trained ensembles of neural networks to detect image areas that cannot reliably be restored. Their method quantified uncertainty by measuring prediction disagreement between ensemble members, highlighting problematic areas in the images. Shin et al. [74] proposed an ensemble DM for MRI reconstruction. The ensemble approach led to better performance, increasing PSNR and SSIM by 2.53 and 1.4%, respectively. Quintero et al. [75] applied ensemble methods to a Unet- and a CycleGAN-based model used for generating synthetic CT images of the abdomen for radiotherapy. This study demonstrated a correlation between the uncertainty maps and the image quality and dose calculation accuracy. Through the calculated uncertainty maps, it was possible to represent the variations derived by the randomness inherited by the models during the training process. An improvement in accuracy was also proved as more predictions per input were included in the standard deviation calculations.

Huang et al. [76] proposed AQuA, a DE-based framework for uncertainty-aware hallucination detection in virtual tissue staining and digital pathology (DP), using repeated virtual staining-autofluorescence cycles and ensemble voting to identify unreliable generative outputs without requiring histochemical stained ground truth at inference. Their method achieved 99.8% accuracy on kidney virtual-staining images and 97.8% accuracy on lung virtual-staining images, demonstrating that ensemble-based UQ can flag realistic hallucinations and poor-quality outputs that may otherwise mislead human experts.

### 3.2. Non-probabilistic UQ methods

Non-probabilistic methods address uncertainty without depending on explicit probabilistic models and are typically employed when precise probabilities or distributions are unavailable or difficult to ascertain. Rather than making strong assumptions about the underlying distribution, these approaches represent uncertainty using alternative mathematical frameworks, such as intervals, to characterize and analyze the range of possible outcomes (Fig. 4).

#### 3.2.1. Interval analysis

IA quantifies uncertainty by defining input values within bounded ranges that contain the true value. This approach propagates these intervals through models to determine output bounds, providing conservative variability estimates without requiring probabilistic distribution assumptions.

Recent applications demonstrate the value of interval-based UQ in medical imaging. For retinal disease classification, Peng et al. [77] employed the encoder of a trained Transformer-based generative model, developed for OCT images, to classify multiple retinal diseases. The proposed foundation model predicted high uncertainty scores for more than 85% images of non-target-category diseases or with low quality to prompt manual checks and prevent misdiagnosis.

In microscopy, Ye et al. [78] proposed a generative method that simultaneously denoises and predicts pixel-wise uncertainty. Their method improved algorithm trustworthiness by providing statistical guarantees and identifying model hallucinations through high uncertainty regions. The system implemented adaptive scanning of uncertain areas, reducing light exposure and acquisition time by 1.25-17 × compared to conventional approaches. Furthermore, they demonstrated uncertainty reduction as denoised images converged to ground truth features, confirming the reliability of their uncertainty estimates.

#### 3.2.2. Conformal prediction

A promising methodology for UQ in GenAI is CP, a relatively new framework for deriving statistically rigorous uncertainty sets for the predictions, making no assumption about the used black-box model nor the distribution of the data. The derived uncertainty regions contain the true unknown outcome with a specified error tolerance. Up to now, CP was successfully used in medical imaging for classification tasks, also in presence of OOD data [79].

In the field of GenAI for medical imaging, a novel approach adapted from CP was proposed by Dey et al. [80], focusing on SR models applied for enhancing the visual quality of CT images. This is based on the use of residual image prediction and enables the estimation of bounds for the generated SR images. The uncertainty in the model prediction is

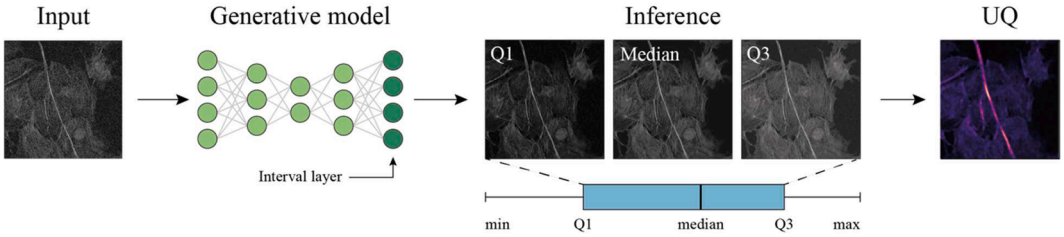


**Fig. 4.** Overview of non-probabilistic UQ methods. The figure illustrates alternative approaches to uncertainty estimation that do not rely on explicit probabilistic models, such as IA. These methods represent uncertainty through bounded ranges and mathematical frameworks that provide conservative estimates without requiring distributional assumptions.

evaluated by calculating the L1 distance of the high-resolution and SR images from the upper and lower bounds. In the absence of a reference high-resolution image the uncertainty is computed as the difference between the boundary images. The bounds are predicted by means of a score function, which acts as an attention map to the residual image and is approximated using the CP, introducing a small correction factor evaluated at different attention levels. The so-calculated bounds provide a probable range of images that a specific model may generate.

Wang et al. [81] explicitly addressed this intersection by integrating CP with a DDPM for MRI SR, proposing E-DM as an uncertainty-controlled evaluation framework in which pixel-wise upper and lower bounds are generated through conformalized quantile regression and risk-controlling prediction sets. This study is particularly relevant because it demonstrates how CP can be embedded into DM-based image generation to provide user-defined reliability guarantees while preserving the quality of the generated medical images. Dey et al. [82] further expanded the use of CP in diffusion-based generative medical imaging through PathGen, a crossmodal DM that synthesizes transcriptomic features from digital histopathology. In this setting, conformal analysis is applied to the outputs enabled by the diffusion-generated modality, providing conformal sets for tumour grading and conformal risk bounds for survival estimation, thereby linking diffusion-based generation with patient-level reliability assessment and fairness analysis across demographic groups.

This type of methodology seems promising also for generative neural style transfer models, used to transfer the style of medical images with features strongly dependent on the equipment post-processing software, like in mammography [83].

## 4. Discussion

### 4.1. Main findings

A temporal analysis of the literature reveals a growing interest in using UQ techniques with GenAI for medical imaging (Fig. 5). Until 2020, only a few isolated studies investigated this combination, reflecting the early development stage of both generative techniques and UQ methods in the medical domain. The period 2021-2024 saw a gradual but consistent increase, with more pronounced growth in the 2022-2024 period. However, 2025 represents a clear inflection point, with the number of publications increasing from 5 to 23. This sharp rise likely reflects not only the rapid expansion of GenAI in medical imaging, but also a broader recognition that image realism alone is insufficient for clinical translation without some form of reliability assessment. In this sense, the increased attention to UQ appears to mirror a shift in the field from simple image generation toward more clinically oriented frameworks that also consider confidence, controllability, and safety. More generally, this temporal evolution is consistent with the maturation of UQ in medical imaging, where techniques initially developed for classification and segmentation are now being extended to the more complex setting of generative models.

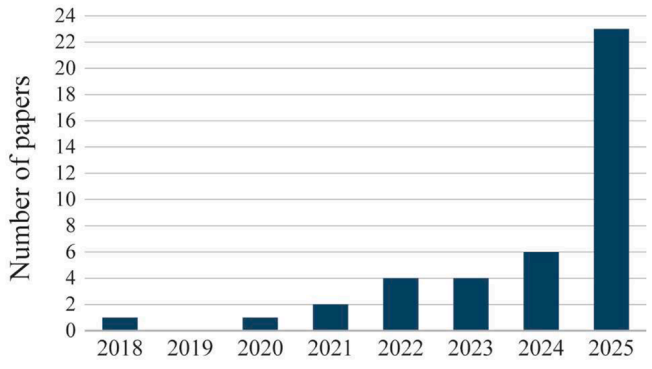


**Fig. 5.** Temporal distribution of published studies on UQ in GenAI for medical imaging (2018-2025).

The analysis of distribution by imaging modality and computational task (Fig. 6) reveals a clear predominance of radiological imaging. MRI is the most widely studied modality (n=19), followed by microscopy (n=10), including DP and fluorescence imaging, CT (n=9), optical techniques and nuclear imaging that contribute with the same number of studies (n=4). This distribution is likely influenced by several factors. First, MRI and CT are among the most widely used modalities in the development of AI methods because they are supported by relatively large public datasets, standardized acquisition protocols, and clinically important applications in reconstruction, denoising, and synthesis. Second, these modalities are particularly compatible with generative modeling because they often involve inverse problems, sparse acquisition, or cross-domain translation tasks, where uncertainty information can provide practical value. Several studies have demonstrated cross-modal applications, with Laves et al. [34] investigating UQ across CT, MRI, OCT, dermatological imaging, and Huang et al. [5] across CT, MRI, PET, while others investigated dual modalities such as CT/MRI [8,53], and PET/MRI [6].

In terms of computational tasks, reconstruction is the most common application (n=17), particularly in radiology, where image quality is critical for diagnosis accuracy. This predominance is not surprising, since reconstruction tasks often require the model to recover missing or degraded information under ill-posed conditions, making uncertainty estimates especially relevant for identifying unreliable regions or sampling ambiguity. Denoising (n=7) and Segmentation (n=7), span a variety of imaging modalities. SR is very common in the microscopic domain (n=5), where improved spatial resolution can provide essential diagnostic information for cellular and tissue investigation. In this domain, Weigert et al. [73] demonstrate the applicability of UQ across numerous fluorescence microscopy tasks such as denoising, reconstruction, and SR.

The analysis of UQ techniques and generative architectures in Fig. 7 reveals significant methodological heterogeneity across imaging modalities, suggesting that no single UQ-framework architecture pair has yet emerged as dominant across all applications. DMs are the most common architecture (n=16), followed by BNNs (n=10), reflecting the

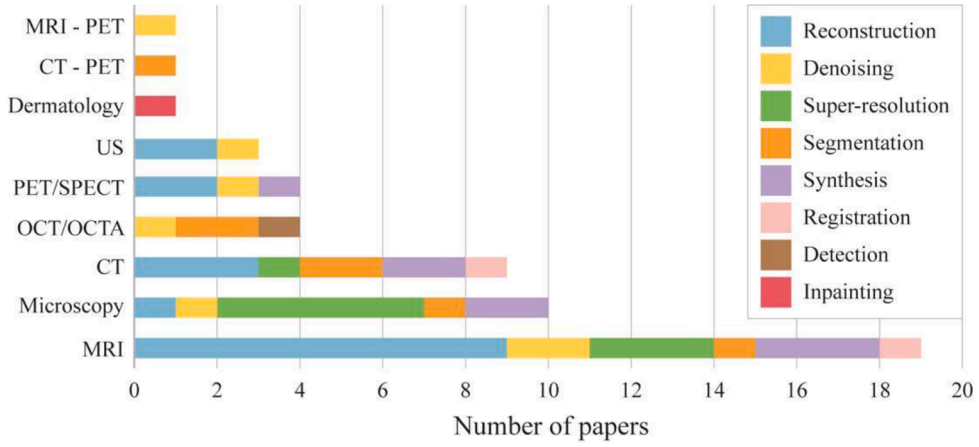


**Fig. 6.** Distribution of studies by imaging modality and task. The matrix visualization shows the relationships between different medical imaging modalities (MRI, CT, PET, SPECT, US, OCT, Fluorescence Microscopy, Dermatology) and GenAI tasks (reconstruction, denoising, synthesis, segmentation, SR, registration, detection, inpainting). The total number of instances (n=52) exceeds the number of papers (n=41), as several studies address multiple modalities or implement different tasks within the same imaging domain.

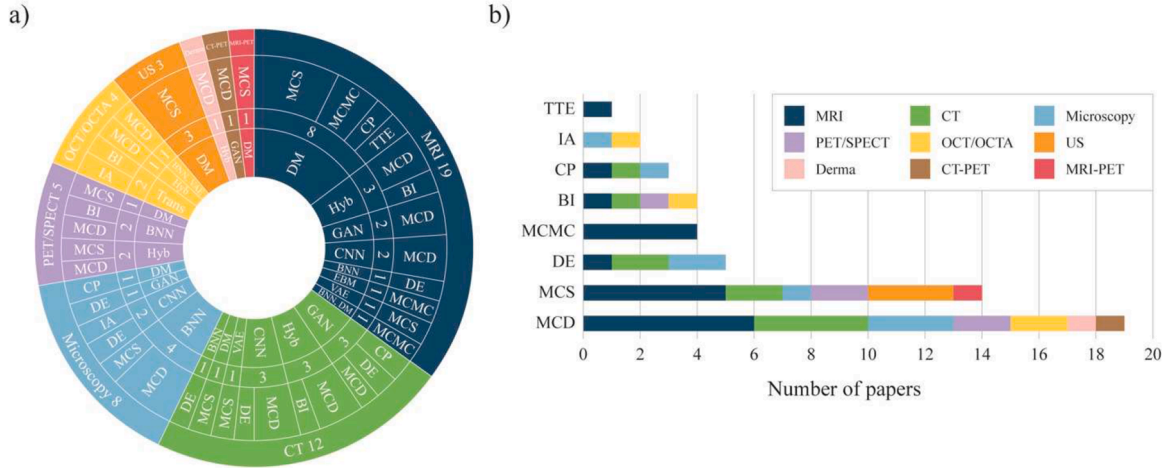


**Fig. 7.** Analysis of UQ methods across imaging modalities and generative models. (a) Sunburst diagram showing the hierarchical relationship between imaging modalities, UQ techniques, and generative model architectures used in the reviewed studies. (b) Distribution of UQ techniques across different medical imaging modalities. The total number of instances (n=54) exceeds the number of papers (n=41) as several studies address multiple modalities or generative models in the same work.

increasing popularity of these approaches due to their inherent ability to model uncertainty. Convolutional Neural Networks (CNNs) and GANs account for the same number of studies (n=7). A significant number of studies (n=9) employ hybrid architectures that combine convolutional network principles with DMs, suggesting an emerging trend toward integrating the representational efficiency of CNN-based feature extraction with the probabilistic generation and uncertainty-aware behavior of diffusion frameworks.

MC methods dominate the UQ landscape, with MCD and MCS accounting for 19 and 14 studies, respectively. This predominance likely reflects their practical appeal, as these approaches are comparatively easy to implement, can often be integrated into existing architectures with limited modification, and are flexible across reconstruction, segmentation, and image synthesis tasks. By contrast, MCMC methods are reported in only 4 studies, primarily for applications requiring more rigorous Bayesian uncertainty estimation. Overall, the distribution shown in Fig. 7 suggests that the field currently favors UQ approaches that are both methodologically practical and computationally feasible, particularly in applications where uncertainty estimation must be incorporated into already demanding generative pipelines.

### 4.2. UQ outputs in GenAI models

The study of validation procedures demonstrates that the majority of research employs a quantitative approach to uncertainty evaluation (n=10 studies) or a combination of qualitative and quantitative methods (n=19 studies) (Fig. 8a). This trend reflects growing recognition within the research community that UQ requires numerical validation, while also acknowledging the importance of qualitative assessment for clinical interpretability (e.g., uncertainty maps). The comparatively limited number of studies that depend entirely on qualitative validation (n=14) indicates that the discipline has progressed from purely subjective evaluations to more objective, quantitative approaches to uncertainty assessment.

In terms of uncertainty output formats, UQ maps and confidence measures are the most common among the research examined (Fig. 8b). UQ maps are graphical representations of uncertainty that enable physicians to identify parts of the synthetic images where the model has higher or lower confidence, which is very useful for clinical decision-making. Confidence metrics provide a more comprehensive assessment of model certainty, giving the end user an overall reliability score

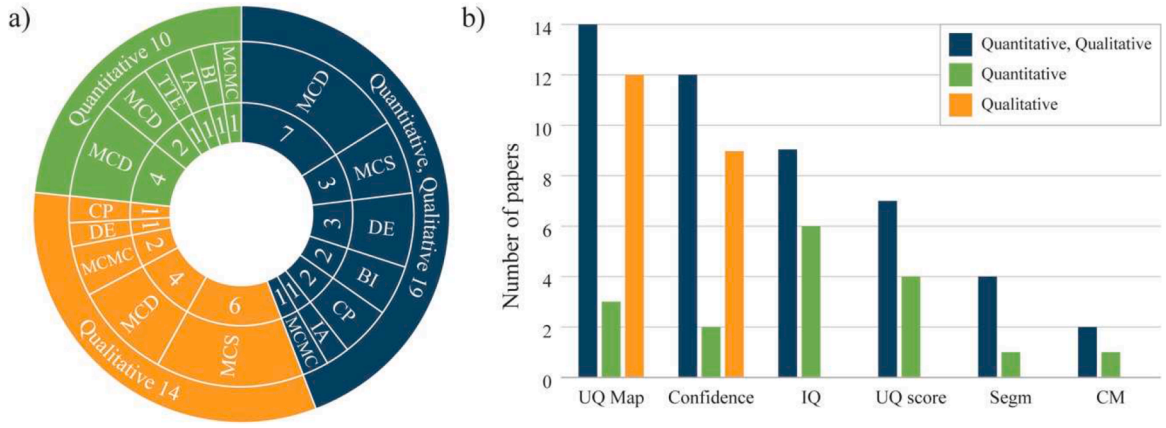


**Fig. 8.** Validation approaches and outputs in uncertainty-aware GenAI models. (a) Sunburst diagram illustrating the types of validation methods employed (quantitative, qualitative, or both) for UQ in generative models. (b) Distribution of output types provided by uncertainty-aware generative models across the reviewed studies.

for the generated data. Other output types, including segmentation masks and image quality metrics, represent specialized applications where UQ is tailored to specific clinical needs.

The task-specific analysis of UQ implementation (Fig. 9) demonstrates no significant differences between UQ techniques and generative tasks, with all UQ approaches being fairly well-represented across tasks. This distribution indicates that the choice of the UQ approach is more reliant on the application's unique needs, computational restrictions, and desired output qualities than on the generating task itself. However, when the relationship between created outputs and UQ approaches is examined, a clear pattern is observed: MCS and MCD are primarily utilized to build UQ maps and confidence measurements. This choice is most likely due to their ease of implementation, which makes them ideal for providing pixel-level uncertainty estimates that can be successfully presented as spatial maps. The versatility of these MC-based approaches allows for flexible uncertainty representation, accommodating both local (pixel-level) and global (image-level) UQ needs across different medical imaging applications.

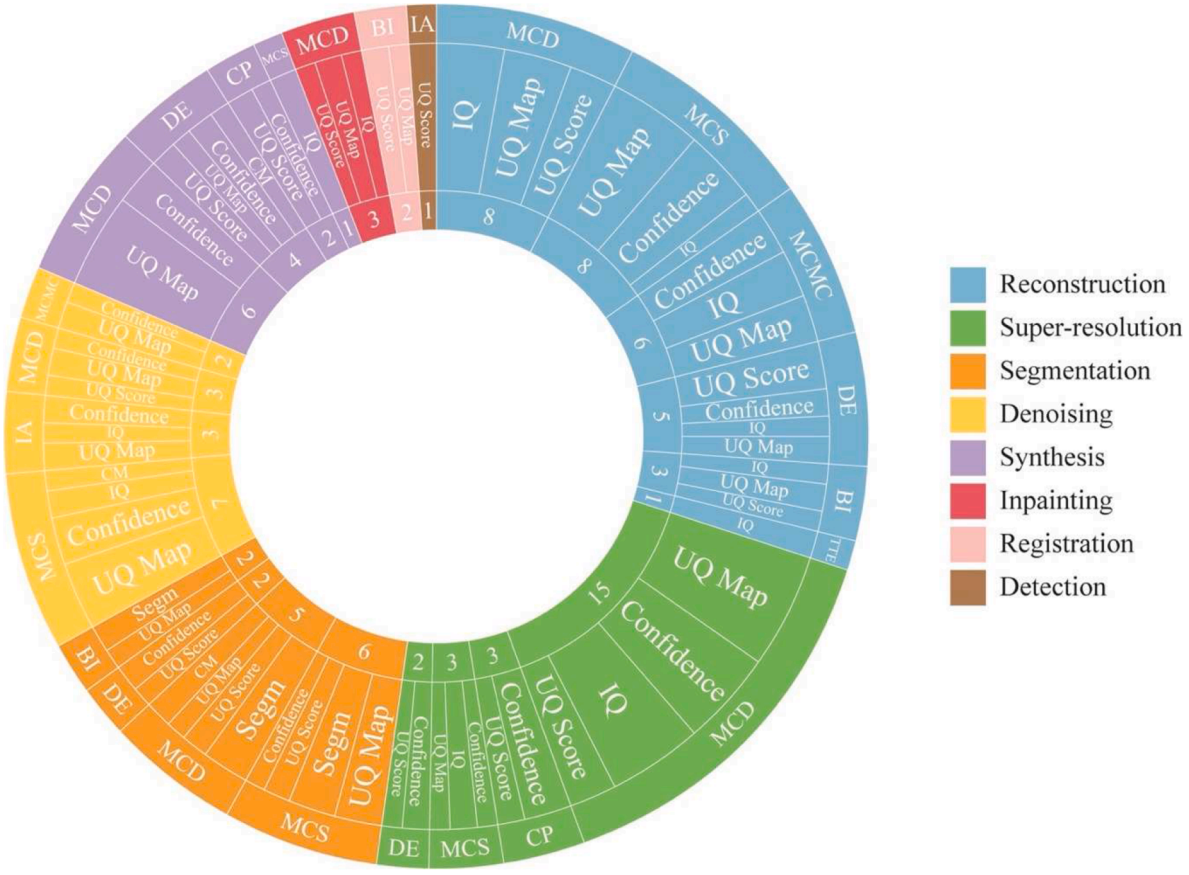


**Fig. 9.** Task-specific UQ implementation and outputs. Sunburst diagram showing the relationship between GenAI tasks, types of uncertainty outputs produced, and UQ techniques employed.

### 4.3. Impact of UQ on GenAI output

Across the reviewed studies, UQ was often associated with improved performance, more informative reliability assessment, or both, but these effects were not universal and depended strongly on how uncertainty was incorporated into the model [5,34,53]. In general, UQ does not improve performance simply by estimating uncertainty. Rather, gains were more likely when uncertainty estimates were actively used to guide training or inference, for example through uncertainty-aware loss functions [53], confidence-aware attention mechanisms [61,72], adaptive sampling [46,78], or the identification of unreliable regions requiring refinement or closer inspection [34,76]. By contrast, post-hoc UQ strategies may primarily improve calibration and interpretability without necessarily increasing task performance [42,63].

In MRI reconstruction, several studies reported gains associated with UQ-aware model design or inference. Liu et al. [69] demonstrated PSNR improvements of 2.6% and SSIM gains of 2.3% through Bayesian UQ, with particular effectiveness at high acceleration rates (R=14). Similarly, Shin et al. [74] improved MRI reconstruction through an uncertainty-guided ensemble of DM inferences, increasing PSNR from 34.01 dB to 36.54 dB (+2.53 dB) and SSIM from 0.930 to 0.944 (+1.4%). Tanno et al. [60] found a 12.3% decrease in Root Mean Squared Error (RMSE) in diffusion MRI enhancement using combined heteroscedastic and variational uncertainty models. In these cases, the benefit appears to arise not only from quantifying uncertainty, but from using it to regularize the model, improve robustness to outliers, or guide the reconstruction process toward more reliable solutions.

In several studies, UQ was also associated with improved segmentation accuracy and reliability, particularly when uncertainty estimates were directly integrated into the segmentation pipeline. Xing et al. [53] found that the Dice score of their DM improved gradually from 89.53% (baseline) to 90.36% (+0.83%) with MC-Diff uncertainty, and then to 91.87% (+2.34%) with complete uncertainty components. In OCT imaging, Liu et al. [72] improved Dice by 3.8%, 3.4%, and 1.8% across separate studies by integrating semantic uncertainty, which was especially beneficial in border regions and semantic ambiguity. Jebril et al. [58] found that employing VQ-VAE with uncertainty estimation improved F1 scores from 0.75 to 0.92 (+0.17) for OCTA anomaly detection. Here, the likely mechanism is that uncertainty helps the model focus on ambiguous boundaries, sparse lesions, or difficult regions where deterministic predictions are more error-prone.

In the context of DL-based generation of synthetic CT images for adaptive proton therapy, Galapon et al. [8] reported significant correlations between uncertainty maps and several quantitative metrics, including Hounsfield units (HU) difference, range error, water equivalent thickness (WET) error, and dose difference, supporting the potential of MCD-based uncertainty estimation as a robust quality assurance tool for synthetic CT images. For the DCNN model, uncertainty correlated most strongly with HU difference with $r = 0.92 \pm 0.03$, followed by dose difference with $r = 0.65 \pm 0.09$, WET error with $r = 0.64 \pm 0.06$, and range error with $r = 0.66 \pm 0.09$. In this case, the main contribution of UQ was not necessarily to improve the generated image itself, but to better characterize when the generated output could be considered reliable for downstream clinical use.

Some studies demonstrated the versatility of UQ across multiple tasks and modalities. For example, Laves et al. [34] demonstrated consistent improvements across inverse challenges, obtaining PSNR increases of 2.2 dB in CT reconstruction, 0.88 dB in MRI SR, and 0.04 dB in OCT denoising. Their method also avoided overfitting, minimized hallucinatory artifacts, and increased cross-task generalization. Huang et al. [5] showed improvements in different reconstruction tasks, including fast MRI, sparse-view CT, and low-dose PET, mitigating the typical performance drop caused by the commonly used dropout. Their method achieved the best reconstruction fidelity, and the GAN version showed the best perceptual quality.

Beyond quantitative performance improvements, UQ integration has provided several qualitative benefits. Tanno et al. [60] showed improved generalization to OOD data and the capacity to detect pathological features unseen during the training process. Jebril et al. [58] demonstrated more consistent anomaly localization and increased robustness across scanners and fields of view. Laves et al. [34] achieved better calibration of uncertainty estimates, making the models more reliable for clinical applications. At the same time, not all UQ implementations improved raw task performance [42,63]. In particular, post-hoc methods such as MC dropout may mainly provide better-calibrated uncertainty maps or confidence estimates, while their effect on predictive accuracy depends on the architecture, the number of stochastic samples, and the task-specific role of dropout regularization (for the complete set of study-level metrics, see Tables A2 and A3).

### 4.4. Advantages of UQ methods and computational considerations

The choice of the UQ method has a significant impact on both the reliability of uncertainty estimates and the computational feasibility of clinical deployment. MC methods represent the most widely adopted approach due to their model-agnostic nature and straightforward implementation. MCD and MCS require multiple forward passes through the network, with computational overhead directly proportional to the number of samples. Reviewed studies typically employ 100-200 samples for adequate uncertainty estimation [34,56,60], which translates to a corresponding 100-200x increase in inference time. This overhead is acceptable for single-slice generative tasks such as 2D MRI reconstruction [60], CT slice denoising [34] or CT image generation from MRI [8], but becomes prohibitive for computationally intensive applications like 3D volumetric reconstruction or gigapixel whole-slide imaging in microscopy [76]. The primary advantage of MC methods lies in their flexibility: dropout layers can be added to existing architectures without requiring a fundamental redesign. However, current limitations include the lack of systematic guidelines for dropout placement and the absence of universal quality metrics for uncertainty with studies showing significant variation in implementation strategies [42,56].

MCMC methods represent a more sophisticated MC approach that constructs dependent sample sequences to approximate complex posterior distributions [67]. While enabling sampling from high-dimensional distributions that would be intractable for direct methods [67,68], MCMC imposes the highest computational burden, often requiring hundreds to thousands of iterations for convergence. Despite this complexity, MCMC provides the most rigorous uncertainty estimates as demonstrated in MRI reconstruction applications [68], where high computational resources can be leveraged to achieve superior reconstruction quality.

BNNs provide the most theoretically principled approach by modeling weight distributions rather than point estimates, offering uncertainty estimates that closely approximate true epistemic uncertainty [54,62]. Computational requirements differ considerably across implementations. Variational approaches need only a single forward pass [34], whereas sampling-based methods require multiple passes, mirroring the computational pattern of MC approaches [54]. Training requires specialized optimization techniques that can be computationally expensive and may exhibit slower convergence [62]. Despite their potential benefits, BNNs have seen limited adoption due to the architectural modifications they require.

Ensemble methods use the diversity of different models to provide robust uncertainty estimates, and interestingly, they can outperform MC approaches with fewer models. When working with 3-5 models in the ensemble, the inference time scales linearly, which typically results in faster uncertainty estimation than alternative approaches. However, the efficiency of the ensemble depends on the performance of the individual models: poorer models in the ensemble impair uncertainty quality and may introduce systematic biases [75]. In practice, this means that developers must not only select comparable models, but also potentially build weighting algorithms to account for inevitable performance

variances within the ensemble.

Finally, non-probabilistic methods such as IA provide computational efficiency at the cost of less informative uncertainty representations. Requiring only a single forward pass, these approaches are suitable for resource-constrained environments [77]. However, their deterministic nature limits the ability to capture the full uncertainty spectrum, providing binary or bounded estimates that may be less clinically informative than rich distributional information from probabilistic approaches [80].

### 4.5. Challenges

UQ in generative medical imaging still faces several challenges that limit the reliability and clinical adoption of AI-generated medical content.

One important issue is the current imbalance between probabilistic and non-probabilistic UQ strategies, with the latter still only sparsely represented. Most non-probabilistic studies in our corpus were published in 2025, which suggests increasing interest in these methods, but also indicates that they are still recent and require further methodological development, broader validation, and more systematic comparison across tasks. These approaches are particularly relevant because they can transform sample variability into statistically interpretable prediction sets. This is especially valuable for DMs, whose stochastic sampling may generate outputs that appear plausible but may still be clinically unsafe. For example, Teneggi et al. [84] introduced a CP approach for score-based DMs, providing calibrated intervals and risk-controlled bounds in image-to-image regression, with additional validation on abdominal CT denoising. Ekmekci and Cetin et al. [85] similarly highlighted the need for scalable UQ frameworks by combining generative posterior sampling, BNNs, and conformal calibration to jointly address aleatoric and epistemic uncertainty; however, their CP experiments were demonstrated on MNIST inpainting rather than on medical imaging data.

A second challenge concerns the close relationship between UQ methodology and the underlying generative architecture. This is particularly evident in the recent transition from adversarial to diffusion-based models. As documented by Azad et al. [86], diffusion frameworks have rapidly become the dominant generative paradigm, and this shift is also reflected in our corpus, where DMs are the most frequent backbone (n=16), compared with GAN-based studies (n=7). This architectural transition has important implications for UQ. MCD and MCS were the most frequently used UQ techniques in our review (n=19, n=14, respectively), but they emerged before the diffusion era as model-agnostic, post-hoc estimators based on dropout insertion or repeated stochastic inference. Accordingly, MCD remains concentrated in deterministic, adversarial, and convolutional settings, including GAN [56], PatchGAN [42], cGAN [57], and cycleGAN [8]. By contrast, DMs offer a more native route to uncertainty estimation through variability across posterior samples or denoising trajectories, without requiring architectural modification [45,46,66,67], including diffusion posterior sampling for high-volume-rate 3D cardiac US [47]. In our corpus, none of the diffusion-based studies used MCD; uncertainty was instead estimated through MCS, MCMC, or conformal and ensemble-based approaches. This does not mean that MCD has become obsolete, but rather that the preferred UQ strategy appears to shift with the dominant generative backbone. A key limitation is that only a few studies provide architecture-controlled, head-to-head comparisons between diffusion-native uncertainty estimation and post-hoc approaches such as dropout-based methods on the same task and dataset. Without such comparisons, it remains difficult to determine whether apparent advantages reflect the UQ method itself or the architecture in which it is applied.

Another challenge is the lack of standardization in uncertainty outputs and reporting practices [70]. Current studies use heterogeneous uncertainty representations, ranging from pixel-wise variance maps [7, 50] to global confidence scores [63], with little agreement on terminology, scaling, or clinical interpretation. This makes it difficult to compare results across studies and limits the practical usability of uncertainty information. A related challenge concerns the interpretation of uncertainty in relation to hallucinated or weakly supported generated content. In most cases, UQ methods do not provide a formal proof that a generated structure is hallucinated; rather, they identify regions where the output is poorly constrained by the input data, unstable across stochastic realizations, or weakly supported by the learned posterior distribution. In Bayesian and sampling-based settings, elevated posterior predictive variance, broad uncertainty intervals, or strong disagreement across MC samples or ensemble members may indicate that a generated region is less robust and therefore at higher risk of reflecting hallucinated content [67,76]. Similarly, in diffusion-based models, variability across repeated conditional generations or posterior samples can reveal anatomically or texturally unstable regions that warrant closer inspection [47,49]. Conformal approaches provide a complementary mechanism by enabling coverage-aware prediction sets or rejection strategies for low-confidence outputs, even when hallucinations cannot be directly localized [80,81]. However, the relationship between uncertainty and hallucination is not one-to-one: true but rare findings may also receive high uncertainty, while some confidently generated artifacts may remain undetected. For this reason, uncertainty estimates should be interpreted primarily as indicators of risk for unreliable content rather than as definitive markers of hallucination.

Calibration also remains insufficiently developed for generative medical imaging. Most existing calibration tools were originally designed for classification, whereas generative tasks involve continuous outputs distributed over high-dimensional image spaces. As a result, standard tools such as reliability diagrams or expected calibration error may not adequately capture spatially structured uncertainty or the quality of confidence estimates in generated images. More suitable calibration strategies for continuous and spatially dependent outputs are still needed.

### 4.6. Future directions

Several research directions may help address the current limitations of UQ in generative medical imaging. A first priority is the development of more standardized frameworks for uncertainty representation, visualization, and reporting. Because uncertainty outputs are inherently dependent on the imaging task, the scale at which uncertainty is expressed, and the computational constraints of the application, a fully uniform standard may not be realistic. However, a minimum reporting framework could still substantially improve comparability across studies. At a minimum, future work should specify the level of uncertainty representation (e.g., voxel-, region-, or patient-level), the type of uncertainty being modeled (aleatoric, epistemic, or combined), the format used to express uncertainty (e.g., uncertainty maps, predictive intervals, entropy-based scores, etc.), and the metrics adopted to evaluate reliability, calibration, and the relationship between uncertainty and model error. Reporting conventions may also need to remain partly modality- and task-specific, since clinically meaningful uncertainty outputs may differ across reconstruction, segmentation, and image synthesis settings. Consistent terminology and reporting practices across imaging modalities, tasks, and model families would make it easier to compare studies and to interpret uncertainty outputs in clinically meaningful ways [76]. Recent framework-oriented work outside medical imaging suggests a move from isolated post-hoc uncertainty maps toward more generalizable UQ layers and uncertainty-guided generation strategies [87,88]. Although these approaches have not yet been systematically translated into medical imaging, they indicate a possible direction for more unified and model-agnostic UQ design.

A second priority is systematic, architecture-aware benchmarking of UQ estimators. At present, methodological choice is often entangled with the generative backbone itself. Future studies should therefore

compare diffusion-native posterior sampling, dropout-based approximations, Bayesian formulations, and ensemble methods on common datasets and tasks, under matched sampling budgets and with standardized reporting of calibration performance and computational cost. Such comparisons would help clarify when the iterative structure of DMs provides a real advantage over post-hoc estimators, and when lower-cost approaches such as MCD remain competitive and easier to deploy, especially in computationally demanding settings such as 3D volumetric reconstruction or gigapixel whole-slide imaging. As DMs continue to expand in medical imaging [86], uncertainty estimation may increasingly be embedded within the denoising process itself rather than added after generation.

Another promising direction is the development of uncertainty-aware evaluation metrics. Current studies still rely heavily on conventional image-quality measures such as PSNR or SSIM, which do not account for the confidence associated with different image regions. Future evaluation frameworks could combine fidelity and uncertainty, for example by weighting errors according to local confidence or by explicitly assessing whether uncertainty maps correctly identify unreliable outputs.

Uncertainty-guided generation is another area with strong potential for clinical use. Rather than producing a single result, future systems could use uncertainty estimates to iteratively refine outputs, trigger additional processing when confidence is low, or adapt computational effort according to case difficulty. In this way, uncertainty could become an active component of the generation process rather than only a retrospective descriptor of model confidence.

Finally, physics-informed UQ is a promising direction for applications in which image formation is governed by known acquisition principles. Although physics-based generative models already incorporate imaging constraints, they may still fail under OOD conditions, artifacts, or unexpected pathological patterns. Integrating UQ into these models could help quantify not only uncertainty in the prediction itself, but also uncertainty in the validity of the underlying physical assumptions. This may be especially relevant in settings where reliable generation depends both on adherence to imaging physics and on the ability to detect when those constraints no longer hold.

## 5. Conclusions

The capacity of GenAI to produce highly realistic medical images could be transformative for healthcare, but current models lack mechanisms to communicate the reliability of their outputs or alert clinicians when generated content may be unreliable. Our systematic review, analyzing 41 studies, offers the first comprehensive look at UQ in medical imaging AI. Incorporating UQ was frequently associated with improved performance or with more informative reliability assessment, including reported gains in reconstruction quality and segmentation accuracy. However, these findings were reported across heterogeneous tasks, modalities, architectures, and evaluation protocols. MC methods currently lead the field, offering flexibility and compatibility across different models, though their computational demands remain a challenge for real-time and large-scale applications. The integration of UQ within GenAI marks an important step toward more trustworthy systems in healthcare. Based on our findings, this research field needs to focus on three key areas: developing standardized uncertainty measures, creating task-specific calibration methods, and building generation frameworks that can refine their outputs based on confidence levels.

## Ethical approval

This systematic review did not require ethical approval as it analyzed only previously published studies.

## Data availability

All data generated or analyzed during this study are included in this published article. The complete dataset used for the analysis is available from the corresponding author upon reasonable request.

## Ethics in publishing statement

All authors agree that:

This research presents an accurate account of the work performed, all data presented are accurate and methodologies detailed enough to permit others to replicate the work.
This manuscript represents entirely original works and or if work and/or words of others have been used, that this has been appropriately cited or quoted and permission has been obtained where necessary.
This material has not been published in whole or in part elsewhere.
The manuscript is not currently being considered for publication in another journal.
That generative AI and AI-assisted technologies have not been utilized in the writing process or if used, disclosed in the manuscript the use of AI and AI-assisted technologies and a statement will appear in the published work.
That generative AI and AI-assisted technologies have not been used to create or alter images unless specifically used as part of the research design where such use must be described in a reproducible manner in the methods section.
All authors have been personally and actively involved in substantive work leading to the manuscript and will hold themselves jointly and individually responsible for its content.

## CRediT authorship contribution statement

**Alen Shahini:** Methodology, Formal analysis, Writing – original draft. **Silvia Seoni:** Data curation, Formal analysis, Writing – review & editing. **Alessandra Manzin:** Methodology, Formal analysis, Writing – review & editing. **Martina Oria:** Formal analysis, Writing – review & editing. **Anjan Gudigar:** Visualization, Writing – review & editing. **U Raghavendra:** Investigation, Writing – review & editing. **Abdulkadir Sengur:** Investigation, Formal analysis, Writing – review & editing. **Rajendra Acharya:** Formal analysis, Writing – review & editing. **Massimo Salvi:** Conceptualization, Formal analysis, Supervision, Writing – original draft.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

Alessandra Manzin and Martina Oria acknowledge the 22HLT05 MAIBAI project, which has received funding from the European Partnership on Metrology, co-financed from the European Union's Horizon Europe Research and Innovation Programme and by the Participating States.

## Appendix

### *Study quality assessment framework*

Each paper was assessed using a standardized quality appraisal framework based on predefined criteria. Five key dimensions were evaluated, as summarized in Table A1. Studies with a total score below 8 points were excluded to ensure the methodological rigor and overall quality of the review. This assessment framework enabled the inclusion of only methodologically sound and clinically relevant studies, thereby maintaining consistent standards across all articles analyzed.

**Table A1**
Quality assessment criteria for studies on UQ of GenAI in medical imaging.

| Criterion | Excellent (3 points) | Good (2 points) | Basic (1 point) | Insufficient (0 points) |
|---|---|---|---|---|
| **Study design quality** | Clear research objective and imaging task; rigorous design with training, validation, and independent test sets; adequate dataset size; external or multicenter validation; clear clinical or technical relevance. | Clear objective and appropriate design; adequate data split; reasonable sample size; internal validation; partial discussion of relevance. | Basic design; limited sample size or validation; relevance only briefly discussed. | Objective or task unclear; data split not adequately described; possible data leakage; insufficient methodological detail. |
| **Medical imaging data and application assessment** | Imaging modality, anatomy, disease context, and task clearly described; reliable annotations or reference standards; preprocessing well reported; dataset variability considered. | Imaging data and task adequately described; annotations mostly clear; preprocessing reported; some discussion of heterogeneity. | General description of imaging data; limited detail on annotations or preprocessing; little discussion of variability. | Imaging application unclear; modality, dataset, annotations, or preprocessing insufficiently reported. |
| **Generative model assessment** | Model architecture clearly described and appropriate for the task; training procedure well reported; performance assessed with suitable metrics; relevant baselines included; limitations discussed. | Model described with sufficient detail; training and evaluation mostly clear; at least one comparison method included; limitations partly discussed. | Model only briefly described; limited training or evaluation detail; weak or no comparison with baselines. | Model poorly described or not appropriate for the task; insufficient implementation detail; no meaningful evaluation. |
| **Uncertainty quantification and statistical validation** | Type of uncertainty clearly defined; UQ method well justified; uncertainty validated through calibration or robustness analyses; results clearly interpreted. | UQ method clearly described and generally appropriate; some validation performed; limited but adequate statistical analysis. | UQ method only briefly described; uncertainty type unclear; limited validation. | No clear UQ method; uncertainty not meaningfully validated; insufficient information to assess the approach. |
| **Reporting quality and reproducibility** | Methods and results reported clearly and in full; limitations discussed; code, data, pretrained models, or enough implementation detail provided. | Most methodological details reported; results clearly presented; limitations discussed; study mostly reproducible. | Basic reporting of methods and results; several implementation details missing; limited discussion of limitations. | Incomplete methods or unclear results; major details missing; insufficient information for replication. |

### *Included articles*

(Table A2,Table A3)

**Table A2**
Studies with probabilistic UQ methods applied to generative models.

| Author | Study quality and aim | Generative Task | Uncertainty type and UQ method | Architecture of the GenAI model | Metrics | Results | Findings |
|---|---|---|---|---|---|---|---|
| Weigert et al. [73] | Score 12/15. To show how content-aware image restoration based on deep learning extends the range of biological phenomena observable by microscopy. | SR, Segmentation, Reconstruction of fluorescence images | Aleatoric: Laplace distribution Epistemic: DE | CNN | Confidence, UCE | Qualitative and quantitative. Ablation: No. 90% pixel-wise confidence intervals were estimated from Laplace distributions; epistemic disagreement was quantified with 5-network ensembles using $D \in [0,1]$. Calibration was assessed with expected calibration error, but no numerical UQ values were reported. | Training ensembles of networks was useful for detecting problematic image areas that cannot reliably be restored. |
| Hu et al. [57] | Score 9/15. To propose an automatic, coarse-to-fine approach for Extranodal natural killer/T cell lymphoma segmentation using adversarial networks. | Segmentation of CT and PET images | Epistemic: MCD | GAN | DSC, UQ score | Quantitative. Ablation: No. Lesion uncertainty score: proposed 0.064 vs U-Net 25.572. Background uncertainty score: proposed 0.9777 vs U-Net 0.2915. CT+PET yielded better | A well-designed NN reduced lesion uncertainty, minimizing prediction errors related to lesion volume and location. |

(*continued on next page*)

**Table A2** (*continued*)

| Author | Study quality and aim | Generative Task | Uncertainty type and UQ method | Architecture of the GenAI model | Metrics | Results | Findings |
|---|---|---|---|---|---|---|---|
| | | | | | | segmentation than single-modality inputs and fewer outliers than U-Net-based methods. | |
| Edupuganti et al. [41] | Score 8/15. To quantify the uncertainty in MRI reconstruction with DL models. | Reconstruction of MR images | Epistemic: MCS | VAE | Variance, Uncertainty Map | Qualitative. Ablation: No. 1,000 MC reconstructions produced pixel-wise variance maps. Adversarial loss increased uncertainty, while recurrent blocks reduced it. | Increased adversarial loss led to larger uncertainty. Using multiple recurrent blocks (cascaded VAE and data consistency layers), decreased uncertainty. |
| Tanno et al. [60] | Score 12/15. To propose to account for intrinsic uncertainty through a heteroscedastic noise model and for parameter uncertainty through approximate BI and integrate the two to quantify predictive uncertainty over the output image. | SR of MR images | Aleatoric: Gaussian distribution Epistemic: MCD | CNN | RMSE, Variance, Uncertainty map | Qualitative and quantitative. Ablation: Yes. MCD reduced RMSE versus 3D-ESPCN (DTI: 6.212→6.116; MAP-MRI: 2.285→1.951). Predictive uncertainty correlated with RMSE and detected >90% of unreliable voxels. Uncertainty maps highlighted pathological structures outside the training distribution. Ablation showed that larger training sets reduced epistemic, but not aleatoric, uncertainty. | Combining aleatoric and epistemic uncertainty improved overall performance. Aleatoric uncertainty increased robustness to outliers, while epistemic uncertainty reduced overfitting. Uncertainty also highlighted pathological structures not seen during training. |
| Chen et al. [71] | Score 13/15. To present TransMorph, a hybrid Transformer-ConvNet model for volumetric medical image registration, and Bayesian variants of TransMorph. | Registration of CT and MR images | Epistemic: BI | Hybrid: Transformer, BNN | UCE, Uncertainty map | Quantitative. Ablation: No. Calibration using predictive variance: miscalibrated uncertainty. Calibration using expected model error: near-perfect calibration and uncertainty maps better aligned with registration errors. | Bayesian variant produced a well-calibrated registration uncertainty estimate. |
| Finck et al. [56] | Score 10/15. To investigate the generalizability of a refined GAN for synthesizing high-contrast double inversion recovery images and propose the use of uncertainty maps to further enhance its clinical utility and trustworthiness. | Synthesis of MR images | Epistemic: MCD | GAN: GAN | Variance, Uncertainty Map | Qualitative. Ablation: No. Uncertainty maps from 100 test-time dropout runs showed high confidence for most multiple sclerosis lesions, while artificial hyperintensities had high uncertainty, helping distinguish true-positive lesions from synthesis artifacts. | Uncertainty maps allowed to discern TP lesions from FP hyperintensities. The study showed the potential benefit of uncertainty maps to help gain trust in GANs and make informed medical decisions. |
| Laves et al. [34] | Score 11/15. To present an unsupervised Bayesian approach to inverse problems in medical imaging using mean-field variational inference with a fully tempered posterior. | Reconstruction of CT images SR of MR images Denoising of OCT images Inpainting of dermatological images | Aleatoric: Gaussian distribution Epistemic: MCD | Hybrid: CNN, BNN VAE, EBM | PSNR, SSIM, UCE, Uncertainty Map | Quantitative. Ablation: Yes. Across CT reconstruction, MRI SR, OCT denoising, and dermoscopy inpainting, the proposed tempered Bayesian model outperformed non-UQ baselines on image-quality metrics and produced better calibrated UCE than SGLD/MCD in the reported examples (CT PSNR 35.05 vs DIP 34.22 | For CT reconstruction, although all models showed a high reconstruction error in the ribs, only the proposed model associated it with high uncertainty values. |

(*continued on next page*)

**Table A2** (*continued*)

| Author | Study quality and aim | Generative Task | Uncertainty type and UQ method | Architecture of the GenAI model | Metrics | Results | Findings |
|---|---|---|---|---|---|---|---|
| | | | | | | and SR PSNR 29.91 vs 29.10). | |
| Rafael-Palou et al. [52] | Score 11/15. To propose a deep hierarchical generative and probabilistic network that, given an initial image of the nodule, predicts whether it will grow, quantifies its future size, and provides its expected semantic appearance at a future time. | Segmentation of CT images | Epistemic: MCS | VAE | Mean and Std. Dev, Uncertainty Map, DSC | Qualitative and quantitative. Ablation: Yes. Dice: Pix2Pix 71%; U-Net 77%; BNN with dropout 78%; proposed 78%. Predictions were accompanied by uncertainty scores for growth probability, nodule size, and segmentation variability. | The performances are accompanied by an uncertainty score. Uncertainty-aware models achieved better performance in predicting and segmenting cancer. |
| Chung et al. [66] | Score 9/15. To propose a new denoising method based on score-based reverse diffusion sampling. | Denoising of MR images | Aleatoric: Gaussian distribution Epistemic: MCMC | DM | Std. Dev., Uncertainty Map | Qualitative. Ablation: No. UQ used 5 posterior samples, with variance increasing as α (denoising) increased. | UQ is used to better understand the denoising process. It increased with higher levels of denoising. |
| Luo et al. [67] | Score 9/15. To introduce a framework that enables efficient sampling from learned probability distributions for MRI reconstruction. | Reconstruction of MR images | Aleatoric: Gaussian distribution Epistemic: MCMC | DM | Variance, Uncertainty Map | Qualitative. Ablation: No. Variance maps indicated greater reconstruction uncertainty under higher MRI undersampling, suggesting less certain solutions and potential hallucination risk. | High undersampling implied a high uncertainty about the solution, which may lead to hallucinations. |
| Tolpadi et al. [42] | Score 9/15. To synthesize post-contrast wrist MRI from non-contrast scans. | Synthesis of MR images | Epistemic: MCD | GAN: PatchGAN | Variance, Uncertainty Map | Qualitative. Ablation: No. Uncertainty maps from 100 latent perturbations showed PatchGAN was more confident in synovial joints than UNet. | Uncertainty analyses are vital to address the criticism of DL algorithms being “black boxes”. |
| Zach et al. [68] | Score 12/15. To develop a generative model for MRI reconstruction that improves generalizability, interpretability, and UQ. | Reconstruction of MR images | Epistemic: MCMC | EBM | Variance, PSNR, Uncertainty Map | Qualitative and quantitative. Ablation: Yes. PSNR: U-Net 34.16; DM: 34.65; proposed 36.17. Improvement: +2.01 PSNR over U-Net. Posterior variance was high around small anatomical structures. | Uncertainty maps provided the clinician with additional information about the inherent uncertainty in the reconstruction. |
| Galapon et al. [8] | Score 12/15. To evaluate the predictive value of uncertainty maps generated with MCD for verifying proton dose calculations on DL-based synthetic CTs derived from MRIs in online adaptive proton therapy. | Synthesis of synthetic CT from MRI | Epistemic: MCD | CNN | Uncertainty Map | Qualitative and quantitative. Ablation: Yes. The DCNN produced the strongest uncertainty-error relationships, with uncertainty-HU 0.92, uncertainty-range 0.66, uncertainty-dose 0.065, and dosimetric agreement 99.76% ± 0.43%. Uncertainty maps corresponded to HU, range, and dose-error patterns. | Correlations between uncertainty maps and quantitative metrics (HU, range, and dose errors) support the use of MCD-based uncertainty estimation as a robust quality assurance approach for DL-based synthetic CT accuracy assessment. |
| Gong et al. [6] | Score 12/15. To propose and evaluate different DDPM-based methods for PET image denoising. | Denoising of PET and MR images | Epistemic: MCS | DM | SSIM, PSNR, MSE, Uncertainty Map | Quantitative. Ablation: No. MSE-predictive uncertainty correlation: DDPM-MR 60.48%; DDPM-PET 94.70%; DDPM-PETMR 92.09%; DDPM-MR-PETCons 92.46%. DDPM with PET information | The addition of MR images improved performance both globally, regionally and reduced the uncertainty values. Adding PET data-consistency constraint further reduced the uncertainty. |

(*continued on next page*)

**Table A2** (*continued*)

| Author | Study quality and aim | Generative Task | Uncertainty type and UQ method | Architecture of the GenAI model | Metrics | Results | Findings |
|---|---|---|---|---|---|---|---|
| | | | | | | outperformed GAN baselines. | |
| Jebril et al. [58] | Score 13/15. To propose a DL-based anomaly detection model to identify anomalies in OCTA. | Segmentation of OCTA images | Epistemic: MCD | BNN, VAE | DSC, IoU, Sensitivity, Specificity, Uncertainty Map | Qualitative and quantitative. Ablation: Yes. DSC/IoU/Sensitivity/ Specificity: Base 52/37/ 80/67%; BNN 61/46/ 80/80%; VAE 60/45/57/91%. Improvements: DSC +9%; IoU +9%; specificity +13%. VAE and BNN localized anomalies in OCTA images. | Improvement DSC: +9% IoU: +9% Specificity: +13% Epistemic uncertainty approach successfully segmented anomalous regions. |
| Liu et al. [72] | Score 11/15. To present a novel approach for the simultaneous segmentation of multi-class macular edema. | Segmentation of OCT images | Epistemic: BI | Transformer | DSC, Uncertainty Map | Qualitative and quantitative. Ablation: Yes. DSC: U-Net 73.5/79.1/ 73.0%; +UQ 77.3/82.5/ 74.8%. Improvement: +3.8/ +3.4/+1.8% DSC. Epistemic uncertainty was higher at lesion/ background junctions and lower in lesion cores. | The addition of uncertainty guided attention module improved segmentation performance. |
| Van Sloun et al. [46] | Score 10/15. To equip US systems with a mechanism to actively reduce uncertainty and maximize diagnostic value across a sequence of experiments. | Reconstruction of US images | Aleatoric: Gaussian distribution Epistemic: MCS | DM | Variance, MAE | Quantitative. Ablation: No. Acquisition: 12.5% scanlines per frame. Maximum-information sampling: 4 lines outperformed random sampling with 6 lines, equivalent to about 50% more useful measurements. Reconstruction accuracy improved versus random sampling. | Uncertainty-aware sampling enabled more efficient and more accurate US reconstruction. |
| Chang et al. [61] | Score 11/15. To introduce a two-step strategy that not only quantifies the uncertainty of DL models but also enhances SR reconstruction. | SR of fluorescence images | Aleatoric: Laplace distribution Epistemic: MCD | BNN | Confidence, PSNR, SSIM, Uncertainty Map | Qualitative and quantitative. Ablation: No. PSNR: +1.7 dB on average. SSIM: +6%. Uncertainty maps guided attention toward high-uncertainty regions during SIM reconstruction. | UQ served as an attention mechanism, directing the NN's focus towards areas of high uncertainty to improve the reconstruction. |
| Chu et al. [45] | Score 9/15. To present a novel posterior sampler for DMs using implicit neural representation, named DiffINR. | Reconstruction of MR images | Epistemic: MCS | DM | Std. Dev., Uncertainty Map | Qualitative. Ablation: No. Standard-deviation maps indicated that DiffINR maintained lower uncertainty at higher acceleration rates than compared reconstruction settings. | With higher acceleration rates the proposed model maintained lower uncertainty. |
| Dong et al. [7] | Score 8/15. To introduce a Flow-based Truncated DDM for SR MR Spectroscopic Imaging. | Denoising of MR images | Epistemic: MCS | DM | Std. Dev., Uncertainty Map | Qualitative. Ablation: No. Standard-deviation maps computed from 50 diffusion samples. Uncertainty was higher near brain boundaries and was used as an error-estimation map. | The uncertainty maps qualitatively highlighted less reliable regions, especially near the brain periphery where lipid contamination and field inhomogeneity are expected. |
| Dou et al. [48] | Score 11/15. To present a probabilistic method based on the conditional | Denoising of US images | Epistemic: MCS | DM | Confidence, Accuracy, AUC | Quantitative. Ablation: No. Accuracy: 90.67%. | UQ could be used in abnormality detection and in development of |

(*continued on next page*)

**Table A2** (*continued*)

| Author | Study quality and aim | Generative Task | Uncertainty type and UQ method | Architecture of the GenAI model | Metrics | Results | Findings |
|---|---|---|---|---|---|---|---|
| | DDM for standard plane localization in 3D US. | | | | | AUC: 0.98. Uncertainty: positive cases 11.82 ± 40.03 vs negative cases 0.256 ± 0.176. Decision threshold: maximum validation-set uncertainty. | reliable devices for clinical practice. |
| Ekmekci and Cetin et al. [85] | Score 13/15. To propose a scalable framework that quantifies both aleatoric and epistemic uncertainty in imaging inverse problems by extending generative posterior sampling with BNNs with latent variables and DE. | Reconstruction of CT and MR images | Aleatoric: generative posterior sampling Epistemic: DE | BNN | Uncertainty Map, NLL, SSIM, MSE | Qualitative and quantitative. Ablation: Yes. DE improved UQ/ reconstruction over single posterior samplers: average DPS CT NLL/ SSIM/MSE = 12.0958/ 0.9211/~0.0002; average UQ-VAE MRI = 272.9950/0.5667/ ~0.0031. DE achieved lower NLL and higher SSIM. Epistemic uncertainty decreased with larger CT training sets and highlighted abnormal/ OOD-like artifacts in CT and MRI. | The framework produced uncertainty estimates with expected epistemic behavior, including reducibility with more training data and sensitivity to abnormal test measurements. It improved predictive uncertainty quality over individual generative posterior sampling models. |
| Huang et al. [5] | Score 13/15. To introduce MambaMIR, an Arbitrary-Masked Mamba-based model with wavelet decomposition for joint medical image reconstruction and uncertainty estimation. | Reconstruction of CT, MR and PET images | Epistemic: MCS | Hybrid | Uncertainty Map, PSNR, SSIM | Qualitative and quantitative. Ablation: Yes. MambaMIR achieved strong fidelity (PSNR 32.51/SSIM 0.768; abdomen SVCT PSNR 45.72/SSIM 0.983) whereas MambaMIR-GAN achieved the best perceptual quality across reconstruction tasks. Uncertainty maps provided MC-based reliability information without the usual dropout-related performance drop. | The MC based proposed model provided uncertainty maps as an additional tool for clinicians, while mitigating the typical performance drop caused by the commonly used dropout. |
| Huang et al. [76] | Score 14/15. To present an autonomous quality and hallucination assessment method for virtual tissue staining and DP. | Synthesis of DP images | Epistemic: DE | GAN: cycleGAN | Confidence score, KL divergence, t-statistic, ROC AUC, PR AUC, sensitivity, specificity, accuracy | Qualitative and quantitative. Ablation: Yes. Accuracy: kidney virtual staining 99.8%; lung virtual staining 97.8%. Ablation: repeated VS-AF cycles and ensemble voting improved hallucination detection versus single-pass/ single-classifier baselines. Qualitative: UQ flagged hallucinated or poor-quality virtual staining outputs. | The UQ strategy helps identify unreliable generative outputs without ground-truth images at inference. It improves trustworthiness of virtual staining by flagging hallucinations and model failures that may look realistic to human experts. |
| Levital et al. [70] | Score 12/15. To propose NPB-LDPET, a DL-based non-parametric Bayesian framework for low-dose PET reconstruction and uncertainty assessment. | Reconstruction of PET images | Aleatoric: BI Epistemic: MCD | BNN | UCE, Std. Dev., Uncertainty Map, Pearson correlation, SSIM, PSNR, NRMSE, MAE | Qualitative and quantitative. Ablation: Yes. Samples: 50. UCE-Dose reduction factor correlation: BNN r2=0.9174; MCD r2=0.6144. Bayes framework vs MCD: MAE -21%; local lesion contrast | UQ using a non-parametric Bayesian DL framewor, produced voxel-wise uncertainty maps and demonstrated strong correlation between uncertainty and dose reduction, outperforming MCD in both reconstruction |

(*continued on next page*)

**Table A2** (*continued*)

| Author | Study quality and aim | Generative Task | Uncertainty type and UQ method | Architecture of the GenAI model | Metrics | Results | Findings |
|---|---|---|---|---|---|---|---|
| | | | | | | +8.3%. Uncertainty maps aligned with reconstruction errors, lesions, and low-SNR regions. | accuracy and uncertainty reliability. |
| Lin et al. [62] | Score 12/15. To present PG-SIM, a probabilistic SIM reconstruction method based on BNNs and incorporating graph representation learning to model optical prior knowledge. | SR of fluorescence images | Aleatoric: Laplace distribution Epistemic: MCD | BNN | Confidence, Uncertainty Map | Qualitative and quantitative. Ablation: No. Proposed method produced pixel-wise uncertainty and confidence maps. Removing the graph-based optical-prior module reduced reliability. It achieved higher confidence than probabilistic baselines and reached PSNR 32.60 ±0.56, SSIM 0.9146 ±0.0213 on high-SNR microtubules. | The proposed model provided uncertainty and confidence maps corresponding to the imaging results, allowing biologists to identify potential imaging errors simultaneously and quantitatively without any reference. |
| Liu et al. [69] | Score 11/15. To accelerate 3D multi-contrast cardiac MR acquisition by a novel method based on score-based DMs with self-supervised learning. | Reconstruction of MR images | Aleatoric: Gaussian distribution Epistemic: MCMC | BNN, DM | PSNR, SSIM | Quantitative. Ablation: Yes. PSNR/SSIM: base 27.1/91.0%; BCNN 28.3/92.2%; proposed BCNN+DM 29.7/93.3%. Improvement over base: PSNR +2.6; SSIM +2.3%. | BCNN is used to retrieve a score to train the DM. The proposed uncertainty-aware model improved reconstruction performance. |
| Liu et al. [54] | Score 13/15. To develop Bayesian DL for structured illumination microscopy, which enhances the reconstruction of densely labeled structures while enabling the quantification of SR uncertainty. | SR of fluorescence images | Aleatoric: Gaussian distribution Epistemic: MCS | BNN | Confidence, PSNR, Uncertainty Map | Qualitative and quantitative. Ablation: No. BayesDL-SIM (5 samples) achieved ~4 × lower calibration error than CARE and >3 × lower error with decoupled training. Model mismatch increased epistemic uncertainty by ~400%, bead-density mismatch by 3–4 ×, and live-cell mismatch by 1.5–2.5 × . It also reconstructed faint and intricate structures more accurately than SIM baselines, with aleatoric and epistemic uncertainties correlating with true SR errors; aleatoric uncertainty better matched error distributions. | The proposed model quantified both aleatoric and epistemic uncertainties, ensuring more reliable and transparent SR imaging. |
| Mansouri et al. [51] | Score 10/15. To develop DL models for CT-free attenuation and MC-based scatter correction in quantitative 90Y SPECT imaging for improved dose calculation. | Synthesis of SPECT images | Aleatoric: MCS | Hybrid: UNet, Transformer | SSIM, PSNR, MAE, Mean Absorbed Dose | Quantitative. Ablation: No. MC correction: SSIM +0.35%; PSNR +8.62; MAE -0.11. MAD closer to reference data. | The most accurate approach for modelling attenuation and scatter was MC simulation. |
| Qiao et al. [63] | Score 14/15. To develop Bayesian DPA-TISR and design an expected calibration error minimization framework that reliably infers inference confidence. | SR of fluorescence images | Aleatoric: Laplace distribution Epistemic: MCD | BNN | Confidence, PSNR, UCE, Uncertainty Map | Qualitative and quantitative. Ablation: No. UCE: 0.100 to 0.018, greater than 5-fold reduction. PSNR: 33.48 to 33.42 after calibration. Confidence maps were | After optimization, the UCE of Bayesian DPA-TISR can be substantially reduced, resulting in well-calibrated confidence quantification capability. |

(*continued on next page*)

**Table A2** (*continued*)

| Author | Study quality and aim | Generative Task | Uncertainty type and UQ method | Architecture of the GenAI model | Metrics | Results | Findings |
|---|---|---|---|---|---|---|---|
| | | | | | | highly consistent with error maps. | |
| Quintero et al. [75] | Score 13/15. To evaluate the impact of model uncertainty on predicted HUs and dose calculation on synthetic CT for abdominal patients. | Synthesis of CT images | Aleatoric: ensemble methods for cycle-GAN and CNN<br>Epistemic: MCD for BNN | CNN, BNN, GAN: cycle-GAN | Uncertainty map | Qualitative and quantitative.<br>Ablation: Yes.<br>UQ setups: DE for CNN/cycleGAN; MCD for BNN with 100 iterations.<br>Whole-image uncertainty vs gamma index: Spearman=0.84, $p<0.001$.<br>PTV-only uncertainty vs gamma index: Spearman=0.01, p=0.8.<br>High uncertainty appeared near body contours, vertebrae, density transitions, motion, and misregistration. | By combining DE and MCD, the study shows that uncertainty maps provide clinically relevant information beyond MAE, RMSE, and SSIM, particularly for identifying image regions where synthetic CT errors may affect adaptive abdominal radiotherapy dose evaluation. |
| Shin et al. [74] | Score 8/15. To propose ensemble and adaptive low frequency mixing on the DM. | Reconstruction of MR images | Epistemic: TTE | DM | PSNR, SSIM | Quantitative.<br>Ablation: Yes.<br>PSNR/SSIM: no ensemble 34.01/93.0%; ensemble 36.54/94.4%.<br>Improvement: PSNR +2.53; SSIM +1.4%. | The study proposed an ensemble approach using multiple inferences of a DM. The study did not mention UQ, but the core concepts and methodology of the proposed ensemble model aligned with UQ principles. Ensemble approach led to better performance. |
| Stevens et al. [47] | Score 9/15. To introduce a novel approach to 3D US reconstruction from a reduced set of elevation planes by employing DMs to achieve increased spatial and temporal resolution. | Reconstruction of US images | Epistemic: MCS | DM | Variance, Uncertainty Map | Qualitative.<br>Ablation: No.<br>Posterior-sampling uncertainty maps increased under stronger subsampling and around synthetic OOD anomalies. | The proposed framework not only improves 3D US reconstruction quality but also provides interpretable uncertainty estimates that support robustness analysis and could guide future uncertainty-aware imaging workflows. |
| Xing et al. [53] | Score 13/15. To develop Diff-UNet, a DM for general 3D medical image segmentation, with MC-based uncertainty maps, an uncertainty-guided loss, and progressive uncertainty-driven refinement across diffusion steps. | Segmentation of CT and MR images | Epistemic: MCS | DM | UCE, DSC, IoU, Uncertainty map | Qualitative and quantitative.<br>Ablation: Yes.<br>HRCT DSC: DE 76.20%; MC dropout 76.34%; MC-Diff 76.90%.<br>HRCT UCE: DE 0.540; MC dropout 0.522; MC-Diff 0.515.<br>Ablation: MC-Diff average DSC 90.36% (+0.83% vs base); PURE average DSC 91.87%.<br>Uncertainty maps changed more clearly between diffusion steps. | Diff-UNet quantitatively and qualitatively outperformed SOTA 3D medical segmentation methods. Improvements MC-Diff led to an average DSC of 90.36%, representing an improvement of 0.83% over base model. The model PURE module achieved better performance, with an average DSC of 91.87%. |
| Yu et al. [50] | Score 10/15. To propose and validate a 3D DDPM for whole-body PET image denoising. | Denoising of PET images | Epistemic: MCS | DM | Uncertainty Map | Qualitative.<br>Ablation: No.<br>PSNR/SSIM/CNR values are only shown in plots.<br>Uncertainty maps from 20 DDPM inference runs showed lower uncertainty for 3D DDPM than 2D DDPM. | 3D DDPM exhibited lower uncertainty than 2D DDPM. Inference process was repeated 20 times with different random seeds. The proposed model exhibited more reliable quantitative characteristics. |
| Zhang et al. [49] | Score 10/15. To propose a DL-based method for quantitative susceptibility mapping reconstruction that is | Reconstruction of MR images | Epistemic: MCS | DM | Std. Dev., Uncertainty Map | Qualitative.<br>Ablation: No.<br>Voxel-wise uncertainty was estimated as the standard deviation across | The proposed diffusion framework naturally enables uncertainty maps, making reconstruction |

(*continued on next page*)

**Table A2** (*continued*)

| Author | Study quality and aim | Generative Task | Uncertainty type and UQ method | Architecture of the GenAI model | Metrics | Results | Findings |
|---|---|---|---|---|---|---|---|
| | robust to data perturbations. | | | | | multiple diffusion reconstructions. Diffusion-QSM achieved the best NRMSE: 50.30 ±3.94. | ambiguity observable while primarily targeting improved robustness and generalization in QSM inversion. |

AF: Autofluorescence; AUC: Area Under Curve; BI: Bayesian Inference; BNN: Bayesian Neural Network; CNN: Convolutional Neural Network; CNR: Contrast-to-Noise Ratio; CT: Computer Tomography; DCNN: Deep Convolutional Neural Network; DDM: Denoising Diffusion Model; DDPM: Denoising Diffusion Probabilistic Model; DE: Deep Ensemble; DIP: Deep Image Prior; DL: Deep Learning; DM: Diffusion Model; DP: Digital Pathology; DSC: Dice Similarity Coefficient; DTI: Diffusion Tensor Imaging; EBM: Energy-based Model; GAN: Generative Adversarial Network; HU: Hounsfield Unit; IoU: Intersection over Union; KL: Kullback-Leibler; MAE: Mean Absolute Error; MAP: Mean Apparent Propagator; MC: Monte Carlo; MCD: Monte Carlo Dropout; MCS: Monte Carlo Sampling; MR: Magnetic Resonance; MSE: Mean Squared Error; NLL: Negative Log-Likelihood; NN: Neural Network; OCT: Optical Coherence Tomography; OOD: Out-of-Distribution; PET: Positron Emission Tomography; PR: Precision-Recall; PSNR: Peak Signal-to-Noise Ratio; RMSE: Root Mean Squared Error; QSM: Quantitative Susceptibility Mapping; ROC: Receiver Operating Characteristic; SGLD: Stochastic Gradient Langevin Dynamics; SIM: Structured Illumination Microscopy; SNR: Signal-to-Noise Ratio; SOTA: State of the Art; SPECT: Single-Photon Emission Computed Tomography; SR: Super-Resolution; SSIM: Structural Similarity Index Measure; TTE: Test-Time Ensemble; UCE: Uncertainty Calibration Error; UQ: Uncertainty Quantification; US: Ultrasound; VAE: Variational Autoencoder; VS: Virtual Staining.

**Table A3**
Studies with non-probabilistic UQ methods applied to generative models.

| Author | Study quality and aim | Generative Task | Uncertainty type and UQ method | Architecture of the GenAI model | Metrics | Results | Findings |
|---|---|---|---|---|---|---|---|
| Wang et al. [81] | Score 11/15. To evaluate and improve the trustworthiness of DM outputs by adding CP-based uncertainty intervals to SR3-DDPM for MRI SR. | SR of MR images | Aleatoric and epistemic: CP | DM | Prediction satisfy risk, Empirical risk, Prediction Interval size, Mean interval size, Median interval size, SR-error, HR-error, Calibration | Qualitative and quantitative. Ablation: No. Risk setting: alpha=0.1, delta=0.1. Average SR-Error: 0.0488; average HR-Error: 0.0497, both <0.1 threshold. Test HR-Errors: 0.0591, 0.0543, 0.0498, 0.0418, 0.0435. Calibration: initial lambda=[0.9, 1.3]; risk fluctuated around 0.1 after fine-tuning. Pixel-wise lower/upper confidence intervals were produced. | DM provides statistically risk-controlled, uncertainty-aware MRI SR. It can align pixel-level risk with user-defined confidence requirements and improve trustworthiness without reported degradation of image quality. |
| Dey et al. [80] | Score 12/15. To propose a novel approach to predict a range of SR images that any generative SR model may yield for a given LR image using residual image prediction. | SR of CT images | Aleatoric: CP | GAN | Uncertainty estimates, Coverage, Variance | Qualitative. Ablation: No. Residual-image bounds quantified coverage and uncertainty. For 4 × lung CT, MoMSGAN reached ground-truth/generated-image coverage 0.458/0.901 with uncertainty 0.056; uniform attention improved coverage to 0.592/0.951 with uncertainty 0.075. The proposed boundary-prediction pipeline achieved the best coverage with comparatively lower uncertainty across attention levels compared with similar models. | The study provides a pipeline to identify SR models for medical imaging leading to optimal coverage and reduced uncertainty. |
| Dey et al. [82] | Score 14/15. To show that PathGen, a diffusion-based crossmodal generative model, can synthesize gene expression from digital histopathology for accurate, reliable, and interpretable prediction of cancer grade and survival risk. | Synthesis of gene-expression profiles from DP images | Epistemic: CP | DM | UQ score, Marginal uncertainty, Conditional uncertainty, Conformal coverage, Conformal set Size, Risk bounds, Coverage slack, Significance tests | Qualitative and quantitative. Ablation: No. CP for grade sets and survival risk bounds showed that PathGen-synthesized transcriptomic features had uncertainty and coverage closely matching real transcriptomic data, | Uncertainty analysis showed that synthesized transcriptomic features provide clinically meaningful reliability estimates comparable to real transcriptomic data, while enabling fairness-aware evaluation of prediction confidence across patient groups. |

(*continued on next page*)

**Table A3** (*continued*)

| Author | Study quality and aim | Generative Task | Uncertainty type and UQ method | Architecture of the GenAI model | Metrics | Results | Findings |
|---|---|---|---|---|---|---|---|
| | | | | | | with no significant cohort or subgroup differences. Synthesized transcriptomics improved prediction over WSI alone and matched real transcriptomics, with similar uncertainty (e.g., grade: 0.413 vs 0.407; survival: 0.438 vs 0.446 in TCGA-GBMLGG). | |
| Peng et al. [77] | Score 10/15. To develop a foundation model with uncertainty estimation to detect 16 retinal conditions on OCT. | Detection of OCT images | Epistemic: IA | Transformer | UQ Score | Quantitative. Ablation: No. Uncertainty thresholding improved F1 from 95.74% to 97.44% internally and 77.11% to 91.47% externally. High uncertainty flagged 89.22%/85.44%/89.03% of non-target or low-quality OCT images. | The encoder of a pretrained Transformer-based generative model, developed for OCT reconstruction, was used to classify multiple retinal diseases on OCT images. The study evaluated the uncertainty of the encoder used as classifier. |
| Ye et al. [78] | Score 12/15. To propose a method for simultaneous denoising and pixel-wise uncertainty prediction in scanning microscopy, using uncertainty-guided rescanning of only high-uncertainty regions to reduce acquisition time and light dose. | Denoising of fluorescence images | Aleatoric: IA | CNN | Confidence, MSE, SSIM, Uncertainty Map | Qualitative and quantitative. Ablation: No. Adaptive acquisition savings: 1.25-17x lower potential light dose/time than non-adaptive acquisition. Hallucinated structures had high uncertainty; uncertainty decreased after multi-image denoising as predictions converged to ground truth. | DL methods for microscopy could be designed to be trustworthy by building in UQ to provide error bars for each prediction. |

CNN: Convolutional Neural Network; CP: Conformal Prediction; DM: Diffusion Model; DP: Digital Pathology; GAN: Generative Adversarial Network; IA: Interval Analysis; MC: Monte Carlo; MCD: Monte Carlo Dropout; MSE: Mean Squared Error; OCT: Optical Coherence Tomography; SR: Super-Resolution; SSIM: Structural Similarity Index Measure; UQ: Uncertainty Quantification.

## References


[1] M.H. Akpinar, A. Sengur, M. Salvi, S. Seoni, O. Faust, H. Mir, F. Molinari, U. R. Acharya, Synthetic data generation via Generative adversarial networks in healthcare: a systematic review of image- and signal-based studies, IEEE Open. J. Eng. Med. Biol. 6 (2025) 183–192, https://doi.org/10.1109/OJEMB.2024.3508472.

[2] J. Ehrhardt, M. Wilms, Autoencoders and variational autoencoders in medical image analysis, in: Biomedical Image Synthesis and Simulation, Elsevier, 2022, pp. 129–162, https://doi.org/10.1016/B978-0-12-824349-7.00015-3.

[3] A. Kazerouni, E.K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, D. Merhof, Diffusion models in medical imaging: a comprehensive survey, Med. Image Anal. 88 (2023) 102846, https://doi.org/10.1016/j.media.2023.102846.

[4] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, D. Shen, Transformers in medical image analysis, Intell. Med. 3 (2023) 59–78, https://doi.org/10.1016/j.imed.2022.07.002.

[5] J. Huang, L. Yang, F. Wang, Y. Wu, Y. Nan, W. Wu, C. Wang, K. Shi, A.I. Aviles-Rivero, C.-B. Schönlieb, D. Zhang, G. Yang, Enhancing global sensitivity and uncertainty quantification in medical image reconstruction with Monte Carlo arbitrary-masked mamba, Med. Image Anal. 99 (2025) 103334, https://doi.org/10.1016/j.media.2024.103334.

[6] K. Gong, K. Johnson, G. El Fakhri, Q. Li, T. Pan, PET image denoising based on denoising diffusion probabilistic model, Eur. J. Nucl. Med. Mol. Imaging 51 (2024) 358–368, https://doi.org/10.1007/s00259-023-06417-8.

[7] S. Dong, Z. Cai, G. Hangel, W. Bogner, G. Widhalm, Y. Huang, Q. Liang, C. You, C. Kumaragamage, R.K. Fulbright, A. Mahajan, A. Karbasi, J.A. Onofrey, R.A. De Graaf, J.S. Duncan, A flow-based truncated denoising diffusion model for super-resolution magnetic resonance spectroscopic imaging, Med. Image Anal. 99 (2025) 103358, https://doi.org/10.1016/j.media.2024.103358.

[8] A.V. Galapon, A. Thummerer, J.A. Langendijk, D. Wagenaar, S. Both, Feasibility of Monte Carlo dropout-based uncertainty maps to evaluate deep learning-based synthetic CTs for adaptive proton therapy, Med. Phys. 51 (2024) 2499–2509, https://doi.org/10.1002/mp.16838.

[9] A. Shahini, A. Gambella, F. Molinari, M. Salvi, Semantic-driven synthesis of histological images with controllable cellular distributions, Comput. Methods Progr. Biomed. 261 (2025) 108621, https://doi.org/10.1016/j.cmpb.2025.108621.

[10] R. Montoya-del-Angel, K. Sam-Millan, J.C. Vilanova, R. Martí, MAM-E: mammographic synthetic image generation with diffusion models, Sensors 24 (2024) 2076, https://doi.org/10.3390/s24072076.

[11] B. Khosravi, S. Purkayastha, B.J. Erickson, H.M. Trivedi, J.W. Gichoya, Exploring the potential of generative artificial intelligence in medical image synthesis: opportunities, challenges, and future directions, Lancet Digit. Heal. 7 (2025) 100890, https://doi.org/10.1016/j.landig.2025.100890.

[12] J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M.J. Muckley, A. Defazio, R. Stern, P. Johnson, M. Bruno, M. Parente, K.J. Geras, J. Katsnelson, H. Chandarana, Z. Zhang, M. Drozdzal, A. Romero, M. Rabbat, P. Vincent, N. Yakubova, J. Pinkerton, D. Wang, E. Owens, C.L. Zitnick, M.P. Recht, D.K. Sodickson, Y.W. Lui, fastMRI: an open dataset and benchmarks for accelerated MRI, (2018). https://doi.org/10.48550/ARXIV.1811.08839.

[13] C. Qiao, D. Li, BioSR: a biological image dataset for super-resolution microscopy, (2024) 12479559462 Bytes. https://doi.org/10.6084/M9.FIGSHARE.13264793.V9.

[14] M. Oria, R. Ferrero, C. Andreis, M. Vicentini, R. Van Engen, C. Roozemond, P. Lamberti, S. Remogna, A. Manzin, Image-based reconstruction of anthropomorphic breast phantoms for synthetic mammogram generation, Comput. Biol. Med. 198 (2025) 111121, https://doi.org/10.1016/j.compbiomed.2025.111121.

[15] M. Oria, R. Ferrero, A. Manzin, Neural style transfer based generation and quality assessment of synthetic mammograms, Manuscr. Under Rev., (n.d.).

[16] Y.Z. Varun Mannam, Fluorescence microscopy denoising (FMD) dataset, (2020). https://doi.org/10.7274/R0-ED2R-4052.

[17] H. Bogunovic, F. Venhuizen, S. Klimscha, S. Apostolopoulos, A. Bab-Hadiashar, U. Bagci, M.F. Beg, L. Bekalo, Q. Chen, C. Ciller, K. Gopinath, A.K. Gostar, K. Jeon, Z. Ji, S.H. Kang, D.D. Koozekanani, D. Lu, D. Morley, K.K. Parhi, H.S. Park, A. Rashno, M. Sarunic, S. Shaikh, J. Sivaswamy, R. Tennakoon, S. Yadav, S. De Zanet, S.M. Waldstein, B.S. Gerendas, C. Klaver, C.I. Sanchez, U. Schmidt-Erfurth,

RETOUCH: the retinal OCT fluid detection and segmentation benchmark and challenge, IEEE Trans. Med. Imaging 38 (2019) 1858–1874, https://doi.org/10.1109/TMI.2019.2901398.
[18] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, 2018, pp. 586–595, https://doi.org/10.1109/CVPR.2018.00068.
[19] M. Lucic, K. Kurach, M. Michalski, S. Gelly, O. Bousquet, Are GANs created equal? A large-scale study, (2017). https://doi.org/10.48550/ARXIV.1711.10337.
[20] A. Borji, Pros and cons of GAN evaluation measures, Comput. Vis. Image Underst. 179 (2019) 41–65, https://doi.org/10.1016/j.cviu.2018.10.009.
[21] Y. Ning, S. Teixayavong, Y. Shang, J. Savulescu, V. Nagaraj, D. Miao, M. Mertens, D.S.W. Ting, J.C.L. Ong, M. Liu, J. Cao, M. Dunn, R. Vaughan, M.E.H. Ong, J.J.-Y. Sung, E.J. Topol, N. Liu, Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist, Lancet Digit. Heal. 6 (2024) e848–e856, https://doi.org/10.1016/S2589-7500(24)00143-2.
[22] F. Chen, Y. Li, Y. Chen, Z. Bian, L. Duo, Q. Zhou, L. Zhang, ADVANCED working group, strategies for the analysis and elimination of hallucinations in artificial intelligence generated medical knowledge, J. Evid. Based. Med. 18 (2025) e70075, https://doi.org/10.1111/jebm.70075.
[23] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training GANs, (2016). https://doi.org/10.48550/ARXIV.1606.03498.
[24] L.M. Kimpton, L.M. Paun, M.J. Colebank, V. Volodina, Challenges and opportunities in uncertainty quantification for healthcare and biological systems, Phil. Trans. R. Soc. A 383 (2025) 20240232, https://doi.org/10.1098/rsta.2024.0232.
[25] S. Seoni, V. Jahmunah, M. Salvi, P.D. Barua, F. Molinari, U.R. Acharya, Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013–2023), Comput. Biol. Med. 165 (2023) 107441, https://doi.org/10.1016/j.compbiomed.2023.107441.
[26] S. Seoni, R.M. Scardigno, B. Cotrufo, M. Salvi, A. Brunetti, A. Guerriero, G. Rotunno, D. Buongiorno, A. Vallan, F. Molinari, K. Meiburger, Exploring uncertainty quantification for photoacoustic image reconstruction and quantitative oxygenation mapping, in: A.A. Oraevsky, L.V. Wang (Eds.), Photons Plus Ultrasound: Imaging and Sensing 2025, SPIE, San Francisco, United States, 2025, p. 128, https://doi.org/10.1117/12.3043687.
[27] M. Salvi, S. Seoni, A. Campagner, A. Gertych, U.R. Acharya, F. Molinari, F. Cabitza, Explainability and uncertainty: two sides of the same coin for enhancing the interpretability of deep learning models in healthcare, Int. J. Med. Inform. 197 (2025) 105846, https://doi.org/10.1016/j.ijmedinf.2025.105846.
[28] M. Salvi, A. Mogetta, U. Raghavendra, A. Gudigar, U.R. Acharya, F. Molinari, A dynamic uncertainty-aware ensemble model: application to lung cancer segmentation in digital pathology, Appl. Soft. Comput. 165 (2024) 112081, https://doi.org/10.1016/j.asoc.2024.112081.
[29] F. Cabitza, C. Fregosi, L. Vicente, Too sure for trust. The paradoxical effect of calibrated confidence in case of uncalibrated trust in hybrid decision making, in: R. Guidotti, U. Schmid, L. Longo (Eds.), Explainable Artificial Intelligence, Springer Nature Switzerland, Cham, 2026, pp. 233–254, https://doi.org/10.1007/978-3-032-08317-3_11.
[30] F.C. Maruccio, W. Eppinga, M.-H. Laves, R.F. Navarro, M. Salvi, F. Molinari, P. Papaconstadopoulos, Clinical assessment of deep learning-based uncertainty maps in lung cancer segmentation, Phys. Med. Biol. 69 (2024) 035007, https://doi.org/10.1088/1361-6560/ad1a26.
[31] S. Sai, A. Gaur, R. Sai, V. Chamola, M. Guizani, J.J.P.C. Rodrigues, Generative AI for transformative healthcare: a comprehensive study of emerging models, applications, case studies, and limitations, IEEE Access. 12 (2024) 31078–31106, https://doi.org/10.1109/ACCESS.2024.3367715.
[32] S. Islam, M.D.T. Aziz, H.R. Nabil, J.R. Jim, M.F. Mridha, M.D.M. Kabir, N. Asai, J. Shin, Generative adversarial networks (GANs) in medical imaging: advancements, applications, and challenges, IEEE Access. 12 (2024) 35728–35753, https://doi.org/10.1109/ACCESS.2024.3370848.
[33] N.L. Bragazzi, S. Garbarino, Toward Clinical Generative AI: conceptual framework, JMIR AI 3 (2024) e55957, https://doi.org/10.2196/55957.
[34] M.-H. Laves, M. Tölle, A. Schlaefer, S. Engelhardt, Posterior temperature optimized Bayesian models for inverse problems in medical imaging, Med. Image Anal. 78 (2022) 102382, https://doi.org/10.1016/j.media.2022.102382.
[35] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, D. Moher, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, BMJ (2021) n71, https://doi.org/10.1136/bmj.n71.
[36] R. Osuala, K. Kushibar, L. Garrucho, A. Linardos, Z. Szafranowska, S. Klein, B. Glocker, O. Diaz, K. Lekadir, Data synthesis and adversarial networks: a review and meta-analysis in cancer imaging, Med. Image Anal. 84 (2023) 102704, https://doi.org/10.1016/j.media.2022.102704.
[37] V. Bansal, A. Jain, N.K. Walia, Diabetic retinopathy detection through generative AI techniques: a review, Results Opt. 16 (2024) 100700, https://doi.org/10.1016/j.rio.2024.100700.
[38] X.J. Tan, W.L. Cheor, E.M. Cheng, C.C. Lim, K.S.A. Rahman, Generative artificial intelligence (GAI) in breast cancer diagnosis and treatment: a systematic review, CMC 84 (2025) 2015–2060, https://doi.org/10.32604/cmc.2025.063407.
[39] S. Faghani, M. Moassefi, P. Rouzrokh, B. Khosravi, F.I. Baffour, M.D. Ringler, B. J. Erickson, Quantifying uncertainty in deep learning of radiologic images, Radiology 308 (2023) e222217, https://doi.org/10.1148/radiol.222217.
[40] L. Huang, S. Ruan, Y. Xing, M. Feng, A review of uncertainty quantification in medical image analysis: probabilistic and non-probabilistic methods, Med. Image Anal. 97 (2024) 103223, https://doi.org/10.1016/j.media.2024.103223.
[41] V. Edupuganti, M. Mardani, S. Vasanawala, J. Pauly, Uncertainty quantification in deep MRI reconstruction, IEEE Trans. Med. Imaging 40 (2021) 239–250, https://doi.org/10.1109/TMI.2020.3025065.
[42] A.A. Tolpadi, J. Luitjens, F.G. Gassert, X. Li, T.M. Link, S. Majumdar, V. Pedoia, Synthetic inflammation imaging with PatchGAN Deep Learning networks, Bioengineering 10 (2023) 516, https://doi.org/10.3390/bioengineering10050516.
[43] J. Gawlikowski, C.R.N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, X.X. Zhu, A survey of uncertainty in Deep Neural networks, (2021). https://doi.org/10.48550/ARXIV.2107.03342.
[44] C.Z. Mooney, Monte Carlo simulation, Nachdr., Sage, Thousand Oaks, Calif, 2003.
[45] J. Chu, C. Du, X. Lin, X. Zhang, L. Wang, Y. Zhang, H. Wei, Highly accelerated MRI via implicit neural representation guided posterior sampling of diffusion models, Med. Image Anal. 100 (2025) 103398, https://doi.org/10.1016/j.media.2024.103398.
[46] R.J.G. Van Sloun, Active inference and deep generative modeling for cognitive ultrasound, IEEE Trans. Ultrason. Ferroelect. Freq. Contr. 71 (2024) 1478–1490, https://doi.org/10.1109/TUFFC.2024.3466290.
[47] T.S.W. Stevens, O. Nolan, O. Somphone, J.-L. Robert, R.J.G. Van Sloun, High volume rate 3-D ultrasound reconstruction with diffusion models, IEEE Trans. Med. Imaging 45 (2026) 2053–2064, https://doi.org/10.1109/TMI.2025.3645849.
[48] H. Dou, Y. Huang, Y. Huang, X. Yang, C. Zhen, Y. Zhang, Y. Xiong, W. Huang, D. Ni, Standard plane localization using denoising diffusion model with multi-scale guidance, Comput. Methods Progr. Biomed. 261 (2025) 108619, https://doi.org/10.1016/j.cmpb.2025.108619.
[49] M. Zhang, C. Liu, Y. Zhang, H. Wei, Diffusion-QSM: diffusion model with time-travel and resampling refinement for quantitative susceptibility mapping, IEEE Trans. Biomed. Eng. 73 (2026) 1401–1413, https://doi.org/10.1109/TBME.2025.3605587.
[50] B. Yu, S. Ozdemir, Y. Dong, W. Shao, T. Pan, K. Shi, K. Gong, Robust whole-body PET image denoising using 3D diffusion models: evaluation across various scanners, tracers, and dose levels, Eur. J. Nucl. Med. Mol. Imaging 52 (2025) 2549–2562, https://doi.org/10.1007/s00259-025-07122-4.
[51] Z. Mansouri, Y. Salimi, N.B. Wolf, I. Mainta, H. Zaidi, CT-free attenuation and Monte-Carlo based scatter correction-guided quantitative 90Y-SPECT imaging for improved dose calculation using deep learning, Eur. J. Nucl. Med. Mol. Imaging 52 (2025) 3484–3499, https://doi.org/10.1007/s00259-025-07191-5.
[52] X. Rafael-Palou, A. Aubanell, M. Ceresa, V. Ribas, G. Piella, M.A.G. Ballester, Prediction of lung nodule progression with an uncertainty-aware hierarchical probabilistic network, Diagnostics 12 (2022) 2639, https://doi.org/10.3390/diagnostics12112639.
[53] Z. Xing, L. Wan, H. Fu, G. Yang, Y. Yang, L. Yu, B. Lei, L. Zhu, Diff-UNet: a diffusion embedded network for robust 3D medical image segmentation, Med. Image Anal. 105 (2025) 103654, https://doi.org/10.1016/j.media.2025.103654.
[54] T. Liu, J. Liu, D. Li, S. Tan, Bayesian deep-learning structured illumination microscopy enables reliable super-resolution imaging with uncertainty quantification, Nat. Commun. 16 (2025) 5027, https://doi.org/10.1038/s41467-025-60093-w.
[55] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: representing model uncertainty in Deep learning, in: Proceedings of the 33rd International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059, in: https://proceedings.mlr.press/v48/gal16.html. accessed November 3, 2025.
[56] T. Finck, H. Li, S. Schlaeger, L. Grundl, N. Sollmann, B. Bender, E. Bürkle, C. Zimmer, J. Kirschke, B. Menze, M. Mühlau, B. Wiestler, Uncertainty-aware and lesion-specific image synthesis in multiple sclerosis magnetic resonance imaging: a multicentric validation study, Front. Neurosci. 16 (2022) 889808, https://doi.org/10.3389/fnins.2022.889808.
[57] X. Hu, R. Guo, J. Chen, H. Li, D. Waldmannstetter, Y. Zhao, B. Li, K. Shi, B. Menze, Coarse-to-fine adversarial networks and zone-based uncertainty analysis for NK/T-cell lymphoma segmentation in CT/PET images, IEEE J Biomed Heal. Inf. 24 (2020) 2599–2608, https://doi.org/10.1109/JBHI.2020.2972694.
[58] H. Jebril, M. Esengönül, H. Bogunović, Anomaly detection in optical coherence tomography angiography (OCTA) with a vector-quantized variational auto-encoder (VQ-VAE), Bioengineering 11 (2024) 682, https://doi.org/10.3390/bioengineering11070682.
[59] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, (2017). https://doi.org/10.48550/ARXIV.1703.10593.
[60] R. Tanno, D.E. Worrall, E. Kaden, A. Ghosh, F. Grussu, A. Bizzi, S.N. Sotiropoulos, A. Criminisi, D.C. Alexander, Uncertainty modelling in deep learning for safer neuroimage enhancement: demonstration in diffusion MRI, Neuroimage 225 (2021) 117366, https://doi.org/10.1016/j.neuroimage.2020.117366.
[61] X. Chang, X. Zhu, Y. Feng, Z. Chen, L. Bian, Structured illumination microscopy with uncertainty-guided deep learning, IEEE Trans. Comput. Imaging 11 (2025) 389–398, https://doi.org/10.1109/TCI.2025.3550715.
[62] K. Lin, J. Dai, H. Chen, Y. Jin, Toward reliable fluorescence imaging: optical prior-guided probabilistic reconstruction for structured illumination microscopy, APL. Photonics. 10 (2025) 066108, https://doi.org/10.1063/5.0270116.
[63] C. Qiao, S. Liu, Y. Wang, W. Xu, X. Geng, T. Jiang, J. Zhang, Q. Meng, H. Qiao, D. Li, Q. Dai, A neural network for long-term super-resolution imaging of live cells with reliable confidence quantification, Nat. Biotechnol. 44 (2026) 110–119, https://doi.org/10.1038/s41587-025-02553-8.

[64] W.R. Gilks, Markov Chain Monte Carlo in Practice, CRC Press LLC, Boca Raton, 1995.
[65] S. Brooks, Markov chain Monte Carlo method and its application, J. R. Stat. Soc. D 47 (1998) 69–100, https://doi.org/10.1111/1467-9884.00117.
[66] H. Chung, E.S. Lee, J.C. Ye, MR image denoising and super-resolution using regularized reverse diffusion, IEEE Trans. Med. Imaging 42 (2023) 922–934, https://doi.org/10.1109/TMI.2022.3220681.
[67] G. Luo, M. Blumenthal, M. Heide, M. Uecker, Bayesian MRI reconstruction with joint uncertainty estimation using diffusion models, Magn. Reson. Med. 90 (2023) 295–311, https://doi.org/10.1002/mrm.29624.
[68] M. Zach, F. Knoll, T. Pock, Stable deep MRI reconstruction using generative priors, IEEE Trans. Med. Imaging 42 (2023) 3817–3832, https://doi.org/10.1109/TMI.2023.3311345.
[69] Y. Liu, Z.-X. Cui, S. Qin, C. Liu, H. Zheng, H. Wang, Y. Zhou, D. Liang, Y. Zhu, Score-based diffusion models with self-supervised learning for accelerated 3D multi-contrast cardiac MR imaging, IEEE Trans. Med. Imaging 44 (2025) 2436–2448, https://doi.org/10.1109/TMI.2025.3534206.
[70] M.F. Levital, S. Khawaled, J.A. Kennedy, M. Freiman, Non-parametric bayesian deep learning approach for whole-body low-dose PET reconstruction and uncertainty assessment, Med. Biol. Eng. Comput. 63 (2025) 1715–1730, https://doi.org/10.1007/s11517-025-03296-z.
[71] J. Chen, E.C. Frey, Y. He, W.P. Segars, Y. Li, Y. Du, TransMorph: transformer for unsupervised medical image registration, Med. Image Anal. 82 (2022) 102615, https://doi.org/10.1016/j.media.2022.102615.
[72] H. Liu, W. Gao, L. Yang, D. Wu, D. Zhao, K. Chen, J. Liu, Y. Ye, R.X. Xu, M. Sun, Semantic uncertainty guided Cross-transformer for enhanced macular edema segmentation in OCT images, Comput. Biol. Med. 174 (2024) 108458, https://doi.org/10.1016/j.compbiomed.2024.108458.
[73] M. Weigert, U. Schmidt, T. Boothe, A. Müller, A. Dibrov, A. Jain, B. Wilhelm, D. Schmidt, C. Broaddus, S. Culley, M. Rocha-Martins, F. Segovia-Miranda, C. Norden, R. Henriques, M. Zerial, M. Solimena, J. Rink, P. Tomancak, L. Royer, F. Jug, E.W. Myers, Content-aware image restoration: pushing the limits of fluorescence microscopy, Nat. Methods 15 (2018) 1090–1097, https://doi.org/10.1038/s41592-018-0216-7.
[74] Y. Shin, G. Son, D. Hwang, T. Eo, Ensemble and low-frequency mixing with diffusion models for accelerated MRI reconstruction, Med. Image Anal. 101 (2025) 103477, https://doi.org/10.1016/j.media.2025.103477.
[75] P. Quintero, L. Cerviño, H. Zhang, W. Harris, Evaluation of model uncertainty in AI-based synthetic CT generation from CBCT for abdominal adaptive radiotherapy, Med. Phys. 52 (2025) 4054–4067, https://doi.org/10.1002/mp.17721.
[76] L. Huang, Y. Li, N. Pillar, T. Keidar Haran, W.D. Wallace, A. Ozcan, A robust and scalable framework for hallucination detection in virtual tissue staining and digital pathology, Nat. Biomed. Eng. 9 (2025) 2196–2214, https://doi.org/10.1038/s41551-025-01421-9.
[77] Y. Peng, A. Lin, M. Wang, T. Lin, L. Liu, J. Wu, K. Zou, T. Shi, L. Feng, Z. Liang, T. Li, D. Liang, S. Yu, D. Sun, J. Luo, L. Gao, X. Chen, B. Huang, C. Zheng, C. Jin, D. Zheng, D. Huang, D. Li, G. Zhang, H. Wu, H. Xia, H. Lin, H. Liang, J. Yi, J. Huang, J. Liu, M. Chen, Q. Zeng, T. Li, W. Chen, X. Huang, X. Chen, X. Ke, X. Liao, Y. Wang, Y. Huang, Y. Cheng, Y. Zhang, Y. Xiong, Y. Huang, Z. Wu, Z. Huang, C.-Y. Cheng, H. Fu, H. Chen, Enhancing AI reliability: a foundation model with uncertainty estimation for optical coherence tomography-based retinal disease diagnosis, Cell Rep. Med. 6 (2025) 101876, https://doi.org/10.1016/j.xcrm.2024.101876.
[78] C.T. Ye, J. Han, K. Liu, A. Angelopoulos, L. Griffith, K. Monakhova, S. You, Learned, uncertainty-driven adaptive acquisition for photon-efficient scanning microscopy, Opt. Express. 33 (2025) 12269, https://doi.org/10.1364/OE.542640.
[79] J. Fayyad, S. Alijani, H. Najjaran, Empirical validation of conformal prediction for trustworthy skin lesions classification, Comput. Methods Progr. Biomed. 253 (2024) 108231, https://doi.org/10.1016/j.cmpb.2024.108231.
[80] S. Dey, P. Basuchowdhuri, D. Mitra, R. Augustine, S.K. Saha, T. Chakraborti, Uncertainty estimation using boundary prediction for medical image super-resolution, Comput. Vis. Image Underst. 256 (2025) 104349, https://doi.org/10.1016/j.cviu.2025.104349.
[81] C. Wang, Y. Gu, P. Liu, S. Zhao, A.S. Kevin Zhou, E-DM: evaluating diffusion model by conformal prediction, in: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), IEEE, Athens, Greece, 2024, pp. 1–5, https://doi.org/10.1109/ISBI56570.2024.10635216.
[82] S. Dey, C.R.S. Banerji, P. Basuchowdhuri, S.K. Saha, D. Parashar, T. Chakraborti, Generating crossmodal gene expression from cancer histopathology improves multimodal AI predictions, Nat. Commun. 17 (2025) 259, https://doi.org/10.1038/s41467-025-66961-9.
[83] Z. Li, Z. Cui, S. Wang, Y. Qi, X. Ouyang, Q. Chen, Y. Yang, Z. Xue, D. Shen, J.-Z. Cheng, Domain generalization for mammography detection via multi-style and Multi-view contrastive learning, in: M. De Bruijne, P.C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham, 2021, pp. 98–108, https://doi.org/10.1007/978-3-030-87234-2_10.
[84] J. Teneggi, M. Tivnan, J.W. Stayman, J. Sulam, How to trust your diffusion model: a convex optimization approach to conformal risk control, (2023). https://doi.org/10.48550/ARXIV.2302.03791.
[85] C. Ekmekci, M. Cetin, Conformalized Generative Bayesian Imaging: an uncertainty quantification framework for computational imaging, (2025). https://doi.org/10.48550/ARXIV.2504.07696.
[86] M. Azad, N.M. Fahad, M.A.K. Raiaan, T.R. Anik, M.F.K. Khan, H.M.K. Toyé, G. Muhammad, A systematic review of diffusion models for medical image-based diagnosis: methods, taxonomies, clinical integration, explainability, and future directions, Diagnostics 16 (2026) 211, https://doi.org/10.3390/diagnostics16020211.
[87] S. Chattopadhyay, B. Kennedy, S. Munikoti, S. Sarkar, K. Pazdernik, Directional Concentration Uncertainty: A representational approach to uncertainty quantification for generative models, (2026). https://doi.org/10.48550/ARXIV.2602.13264.
[88] J. Han, L.L. Beyer, S. Karaman, Flow matching with uncertainty quantification and guidance, (2026). https://doi.org/10.48550/ARXIV.2602.10326.