Uncertainty Quantification for Machine Learning in Healthcare: A Survey

L. Julián Lechuga López [1,2] LEOPOLDO.LECHUGA@NYU.EDU
Shaza Elsharief [2] SE1525@NYU.EDU
Dhiyaa Al Jorf [2] DA2863@NYU.EDU
Firas Darwish [2] FBD2014@NYU.EDU
Congbo Ma [2] CM7196@NYU.EDU
Farah E. Shamout [1,2] FARAH.SHAMOUT@NYU.EDU
New York University [1], *New York University Abu Dhabi* [2]

Abstract

Uncertainty Quantification (UQ) is pivotal in enhancing the robustness, reliability, and interpretability of Machine Learning (ML) systems for healthcare, optimizing resources and improving patient care. Despite the emergence of ML-based clinical decision support tools, the lack of principled quantification of uncertainty in ML models remains a major challenge. Current reviews have a narrow focus on analyzing the state-of-the-art UQ in specific healthcare domains without systematically evaluating method efficacy across different stages of model development, and despite a growing body of research, its implementation in healthcare applications remains limited. Therefore, in this survey, we provide a comprehensive analysis of current UQ in healthcare, offering an informed framework that highlights how different methods can be integrated into each stage of the ML pipeline including data processing, training and evaluation. We also highlight the most popular methods used in healthcare and novel approaches from other domains that hold potential for future adoption in the medical context. We expect this study will provide a clear overview of the challenges and opportunities of implementing UQ in the ML pipeline for healthcare, guiding researchers and practitioners in selecting suitable techniques to enhance the reliability, safety and trust from patients and clinicians on ML-driven healthcare solutions.

Data and Code Availability This literature review does not rely on any specific dataset, as it synthesizes findings from existing research on UQ in healthcare. No new data was generated, and no code was developed or available for sharing.

Institutional Review Board (IRB) This literature review on UQ in healthcare does not involve human subjects, so IRB approval was not required.

1. Introduction

Machine Learning (ML) is revolutionizing healthcare by enhancing diagnostic performance, personalizing treatment plans, optimizing hospital operations, and accelerating drug discovery, ultimately leading to improved patient outcomes and more efficient and safe medical practices (Alowais et al., 2023). Many systems have been developed to support clinical diagnosis (Browning et al., 2021), from analyzing medical images for anomaly detection (Zou et al., 2023), to providing personalized treatment plans based on patient-specific physiological and genetic characteristics (Durso-Finley et al., 2023b).

However, due to the safety-critical nature of clinical practice, the development of trustworthy and deployable ML in healthcare requires the implementation of robust Uncertainty Quantification (UQ) (Begoli et al., 2019; Gruber et al., 2023). Variations in real-world clinical environments affect the performance of predictive systems and introduce uncertainty at different stages of the ML pipeline: data noise and distribution drift, bias and miscalibration of model parameters, or evaluation of the model in an out-of-distribution scenario, such as deployment in a different hospital (Azizmalayeri et al., 2025). By complementing AI-driven healthcare systems with assessments of the uncertainty in their predictions, methods can help better explain whether errors can be attributed to randomness and noise or whether they

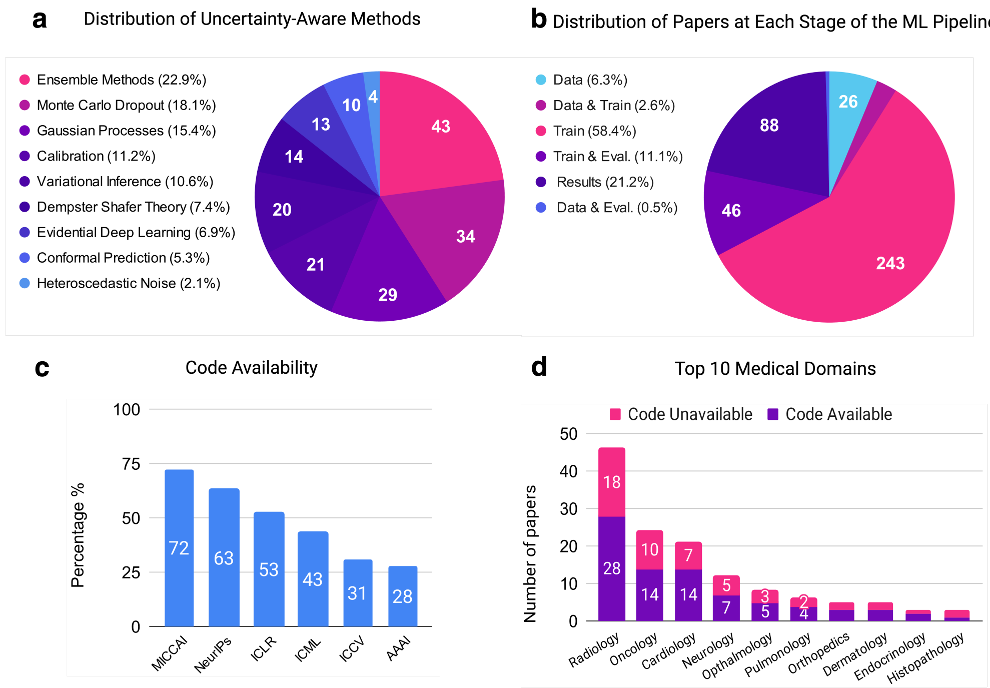

Figure 1: **Overview of distribution and characteristics of reviewed papers. (a)** Prevalence of different uncertainty quantification methods across the surveyed papers. **(b)** Distribution of studies according to the machine learning pipeline stages: data processing, model training, and evaluation. **(c)** Code availability rates across papers published in various conferences and journals. **(d)** Medical domains represented in the reviewed studies, alongside their corresponding code availability.

are due to design choices made during model training and development.

Furthermore, the development of UQ in ML for healthcare can improve confidence in the adoption of these tools by clinicians, patients, and institutions alike (Kurz et al., 2022). Robust and informative UQ diagnostics can help clinicians focus their attention on specific details of patient data, and distinguish between predictions made with high confidence and those with substantial ambiguity, which allows for better risk management (Ren et al., 2023). For patients, understanding the degree of confidence associated with their personalized predictions allows them to weigh the risks and benefits of different interventions, reduces the chances of inappropriate or invasive treatment, and improves their trust in the model (Durso-Finley et al., 2023b). This transparency is vital for healthcare institutions, as it leads to better informed decision-making, optimization of operations, reduction of misdiagnoses and incorrect treatment recommendations, and most importantly, improved patient care (Zou et al., 2023).

Unfortunately, real-world implementation of UQ models in healthcare is still hampered by the limited development of tailored solutions for improving these models (Ovadia et al., 2019; Kompa et al., 2021). Limited underlying theory on how to best adapt predictive uncertainty methods in clinical tasks shows that the use of UQ in clinical applications is not common practice (Begoli et al., 2019; Lambert et al., 2024). In addition, since uncertainty in healthcare applications originates at different stages, it needs to be analyzed from the perspective of the ML pipeline: data processing, model training, and model evaluation. Therefore, there is a crucial need to design and develop Uncertainty Quantification for Machine Learning in Healthcare (UQML4H) to enable the implementation of trustworthy systems that are adapted to robustly mitigate uncertainty during the full ML design lifecycle. Our main goal is to provide a comprehensive overview of State-of-the-Art (SOTA) UQ methods in healthcare, clinical datasets, tasks and domains, and to encourage the development of new UQ methodologies that tackle specific challenges relevant to the nature of each medical domain.

1.1. Motivation

Most existing reviews in UQ cover a wide range of applications and domains, with narrow focus on healthcare. These studies highlighting UQ applications in healthcare primarily focus on a single type of data modality (i.e., medical imaging) (Huang et al., 2024a) or a specific clinical task (Barbano et al., 2022). However, no previous studies pay particular attention to analyzing UQ methods from an ML pipeline point of view. Our work is intended to bridge SOTA research in UQ and tailored clinical applications, with an emphasis on analyzing each phase of model development. Compared to existing work, our survey has four main contributions:

1. We focus on recent SOTA literature on UQ from both medical and nonmedical domains published in the last four years, capturing extensive applications in healthcare (e.g., diagnosis, decision-support systems, etc.) and methodological advances (e.g., theory, algorithms, optimization).
2. We distinguish and analyze UQ methods at each stage of the ML pipeline: data processing (e.g., collection, labeling, alignment), model training (e.g., architecture, tuning, loss design), and evaluation (e.g., inference, metrics, calibration).
3. We present a comprehensive taxonomy of UQ methods categorized by domain, dataset, and task, providing a practical reference for domain-specific applications in healthcare.
4. We connect current methodological advances with practical deployment considerations of UQ in healthcare, identifying key gaps and outlining future research directions and open challenges.

Scope of the Review

To this end, the reviewed papers (overview provided in Figure 1) were selected using the following criteria:

- Recently published peer-reviewed work, (i.e., publication year $\geq$ 2020), and earlier seminal papers in UQ.
- Articles from top-tier AI conferences and medical journals that focus on the application or development of UQ at any stage of the ML pipeline.
- Key information extracted: UQ methods, code availability, clinical datasets and tasks, and specific healthcare domains targeted (e.g., oncology, radiology, cardiology, etc.).

In Section 2, we present our main findings regarding the applications of UQ in healthcare across each stage of the ML pipeline, discussing both current applications and emerging opportunities. The section concludes with a comparative analysis of different UQ techniques and applications in healthcare, highlighting the strengths and weaknesses of seven representative methods. Section 3, presents a thorough discussion of open challenges and future research directions, touching on details of deployment, fairness, regulation, evaluation benchmarks, and a roadmap towards safe UQ implementation in healthcare. Concluding remarks are provided in Section 4.

In Appendix A, we provide a concise overview of key UQ methodologies and identify application domains closely linked to current SOTA developments in machine learning. Furthermore, Appendix B (Table B1) presents detailed information on 94 open-source clinical datasets, grouped by medical domain, currently used in the development and evaluation of UQ methods. Appendix C (Table C1) summarizes the main characteristics of the studies reviewed in this survey, organized by ML pipeline stage and clinical context.

We anticipate that this comprehensive collection of resources will serve as a valuable reference for researchers and practitioners aiming to advance UQ applications in healthcare.

2. Uncertainty Quantification in Healthcare: Applications Across the Machine Learning Pipeline

This section synthesizes key findings on the use of UQ methods in healthcare across the ML pipeline, spanning data preprocessing, model training, and evaluation. We highlight methods applied across diverse clinical tasks, medical domains, and datasets, and point to emerging trends from other fields that may shape future developments. Figure 2 summarizes the main insights, illustrating sources of uncertainty, expected outcomes, and representative UQ techniques at each pipeline stage. Table C1 (Appendix C) provides detailed information on all reviewed studies discussed in this section. This synthesis offers a struc-

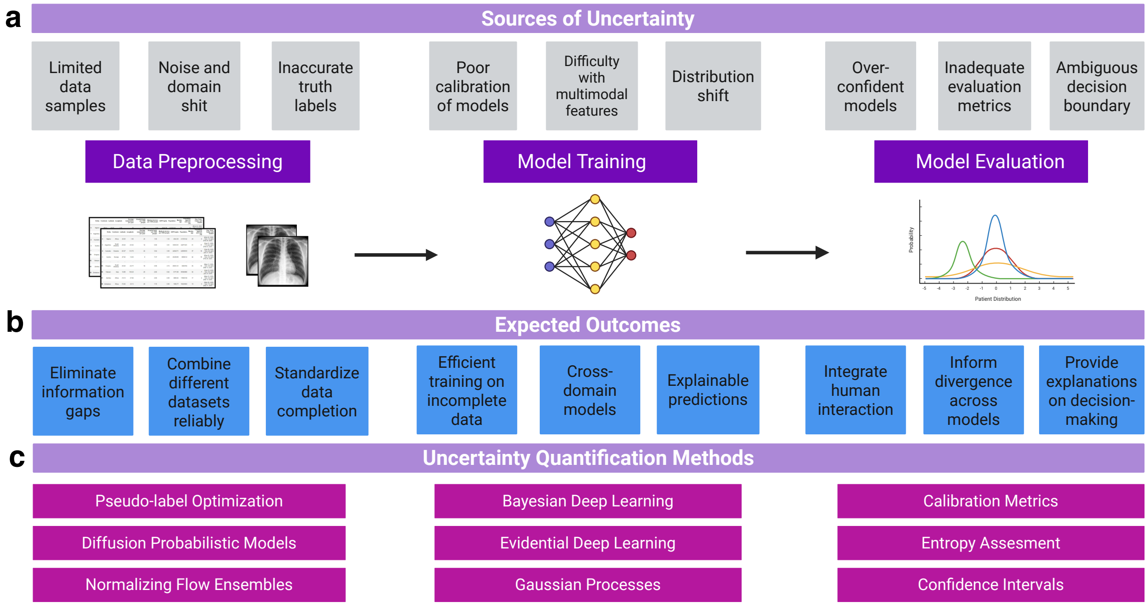

Figure 2: **UQ in the Clinical Machine Learning Pipeline.** **(a)** Key sources of uncertainty identified at each stage of the pipeline. **(b)** Expected outcomes of implementing UQ methods for clinical tasks. **(c)** Relevant UQ techniques applied during data processing, model training, and evaluation.

tured overview of how UQ is currently integrated into healthcare ML workflows.

2.1. Data Preprocessing

The data preprocessing stage is fundamental for ML modeling, addressing challenges like noise, imbalance, and incompleteness that increase uncertainty. UQ methods at this stage focus on enhancing the quality and reliability of inputs for subsequent stages. Techniques include correcting label noise in imbalanced datasets, UQ pseudo-labeling, and modeling uncertainty with normalizing flows.

2.1.1. Applications

Although methods for data preprocessing in healthcare are limited, and the boundary between preprocessing and early training stages may be ambiguous, several studies have proposed UQ methods to enhance data reliability before training. For instance, Angelopoulos et al. (2022) made use of a distribution-free UQ method for image-to-image regression for MRI and microscopy imaging, providing pixel-wise uncertainty intervals to preprocess the input data with formal statistical guarantees addressing uncertainty. Similarly, Li et al. (2022a) improved the quality of data annotations using label probability distributions for tumor cellularity assessment in breast cancer histological images. Das et al. (2024) proposed AnoMed, a semi-supervised confidence guided pseudo-label optimizer, to capture anatomical structures and underlying representations in chest X-rays. In multichannel brain MRI, Tanno et al. (2019) introduced a method to decompose predictive uncertainty and quantify the effects of intrinsic and parameter uncertainty of data. Working on echocardiography images, Gu et al. (2024a) presented Re-Training for Uncertainty (RT4U), a data-centric method to introduce uncertainty to weakly informative inputs in the data. Using diffusion probabilistic modeling, Oh et al. (2024) used data augmentation and synthesis to address domain shift issues, while Khader et al. (2023); Adib et al. (2023); Iuliano et al. (2024) focused on generating high-quality synthetic data and assessed the association between original and synthetic data for 3D medical images, MRI and malaria images, and ECG data.

Beyond specific applications, the effectiveness of UQ methods at the data processing stage is also shaped by the type of data modality and its resolution, which introduce distinct sources of uncertainty and influence methodological suitability. For instance, uncertainty in MRI may arise from acquisition parameters or reconstruction artifacts (Edupuganti et al., 2021; Zhao et al., 2024b), X-ray imaging can suffer from low contrast and dose variability (Cai et al., 2022; Gong et al., 2023a), and EHR data is prone to missing values and documentation inconsistencies across institutions (Horii and Chikahara, 2023; Deng et al., 2025). In time-series data, higher sampling rates can capture finer dynamics but risk overfitting to noise; lower rates reduce noise sensitivity but may miss short-term patterns (Puri et al., 2022; Folgado et al., 2023).

To address this, UQ methods should be tailored to modality-specific characteristics: Bayesian Neural Networks (BNNs) are well-suited for pixel-level uncertainty in imaging, while deep ensembles are more effective for structured EHR data (Caldeira and Nord, 2020; Peng et al., 2020). Multi-scale UQ approaches and ensemble techniques can also help strike a balance between sensitivity and robustness (Vranken et al., 2021; Hamedani-KarAzmoudehFar et al., 2023). Data resolution also plays a pivotal role, as high resolution imaging scans offer greater detail but introduce noise and computational overhead, whereas low-resolution data may omit clinically relevant features, affecting the reliability of UQ estimates (Nehme et al., 2023; Fu et al., 2025). Finally, in multimodal settings, such as those involving imaging and EHR, heterogeneous uncertainty sources must be jointly modeled with techniques like hierarchical Bayesian models. Uncertainty-weighted fusion can also account for these modality-specific variances, enhancing the overall reliability of clinical predictions (Fu et al., 2025).

2.1.2. Opportunities

Label Handling. This focuses on addressing noisy and imbalanced datasets through adaptive labeling techniques. Relevant examples include uncertainty quantification pseudo-labeling strategies for semi-supervised learning that filter unreliable pseudo-labels to enhance model robustness (Rizve et al., 2021; Yan et al., 2022) and Uncertainty Correction of Labels (UCL) proposed by Huang et al. (2022a) to identify and correct mislabeled data.

Noise Reduction and Augmentation. These methods attenuate data variance and calibrate noise levels, preventing spurious uncertainty signals. Techniques employing aleatoric uncertainty estimation and UQ augmentation can help refine domain adaptation by aligning source and target distributions (Yan et al., 2022; Zhang et al., 2024b).

Input Transformations. Input transformation methods are often used to improve uncertainty estimates. For example, normalizing flow ensembles and the use of confidence score calibration during preprocessing enables accurate representation of both aleatoric and epistemic uncertainties (Berry and Meger, 2023; Yang et al., 2024). These studies demonstrate the meaningful integration of UQ into the data preprocessing stage to mitigate the impact of poor quality data, enhance reliability, improve annotation quality, and mitigate uncertainties before training begins.

2.2. Model Training

This stage is crucial for integrating UQ, as it impacts generalization and reliability of predictions. UQ methods at this stage focus on capturing both epistemic and aleatoric uncertainty to enhance model robustness, reduce overfitting, and improve predictive performance under uncertainty.

2.2.1. Applications

The training stage has seen the most development of UQ methods, with various approaches used to improve reliability and robustness in medical domains.

Evidential Deep Learning (EDL) has been widely applied across different tasks for its ability to provide meaningful uncertainty estimates while maintaining high performance. For instance, Ren et al. (2023) used EDL for joint image classification and segmentation in ophthalmology and Yang et al. (2023a) employed it to improve segmentation reliability in general surgery. In oncology, Dong et al. (2024) adopted EDL for prostate cancer grading, while Jeong et al. (2024) incorporated pixel-wise uncertainty into diffusion models for adversarial colonoscopy image generation.

Bayesian methods have also emerged as prominent UQ techniques at the training stage. Zhao et al. (2022) applied Bayesian techniques for efficient uncertainty estimation in cardiology segmentation tasks, while Bayesian variational inference was leveraged by Adams and Elhabian (2023) for super shape prediction in cardiology. Additionally, sparse Bayesian networks were proposed by Abboud et al. (2024), to efficiently quantify uncertainty in skin cancer classification and chest radiology segmentation.

Monte Carlo-based approaches and ensemble methods are also commonly used for UQ during training. Monte Carlo (MC) Dropout was utilized by Aljuhani et al. (2022) for histological image classification in oncology. Ensemble methods were explored by Kazemi Esfeh et al. (2022) on video frames to address ejection fraction regression in cardiology. Similarly, Wu et al. (2022) combined Bayesian and ensemble techniques to predict remaining surgery duration in ophthalmology using surgical video frames, producing a prediction and uncertainty estimation in the same inference run. In another vision application, Zhao et al. (2024a) combined Bayesian methods with ensembles for uncertainty-guided segmentation tasks in cardiology.

Beyond these core methods, several innovative techniques have been introduced to address domain- or task-specific challenges. For example, uncertainty attention modules were employed by Xie et al. (2022c) to handle ambiguous boundary segmentation in cardiology and fetal ultrasound. Uncertainty-weighted class activation maps were leveraged by Fu et al. (2025) for weakly-supervised segmentation in neonatal medicine. Epistemic and aleatoric uncertainties were explicitly modeled by Xiang et al. (2022a), who employed these techniques for supervised and unsupervised learning in segmentation tasks within pancreatology and cardiology. Larrazabal et al. (2023) proposed regularization techniques, such as the use of maximum entropy calibration to improve segmentation reliability in radiology and cardiology. Moreover, contrastive learning and latent space comparisons were developed by Judge et al. (2022a), enabling robust uncertainty estimation across multiple imaging datasets.

Entropy-based methods are also frequently employed for UQ. Sharma et al. (2024) introduced entropy-driven self-distillation learning to enhance classification performance in ophthalmology, oncology, and skin cancer applications. Gaussian probability distributions were leveraged by Judge et al. (2023a), addressing challenges in cardiology and chest X-ray image segmentation, while Li et al. (2023a) proposed a Dirichlet distribution classifier for curriculum learning in skin cancer and COVID-19 image classification tasks.

Other notable methods include the use of autoencoders by Lennartz and Schultz (2023a), who introduced a segmentation distortion measure to improve uncertainty estimation under domain shifts in neurology imaging. Generative adversarial networks (GANs) have also been adapted by Upadhyay et al. (2021), who introduced Uncertainty-Guided Progressive GANs (UP-GAN) for image-to-image translation tasks in medical imaging. Hung et al. (2024) employed cross-slice attention and evidential critical loss for prostate cancer detection using MRI scans. Finally, Browning et al. (2021) used deep reinforcement learning to estimate uncertainty for pathology landmark detection in orthopedics. Zou et al. (2022) utilized subjective logic theory to achieve trusted brain tumor segmentation in neurology-oncology.

2.2.2. Opportunities

Architecture-Level Approaches. Recent advancements focus on novel architectures to address specific challenges for UQ. Rudner et al. (2022) introduced tractable function-space variational inference for BNNs, resulting in improved computational efficiency while maintaining robust uncertainty estimates. Mao et al. (2021b) proposed UAS-Net, an architecture that adapts sampling based on uncertainty estimation to help handle noisy data and refine depth estimation accuracy. More commonly, existing architectures are modified to adapt UQ methods to specific applications. For example, Zhang et al. (2022c) integrated transformers with uncertainty modeling using attention mechanisms that adapt based on uncertainty levels in the data. Similarly, Zhu et al. (2022) designed a depth completion network that incorporates uncertainty into the architectural layers, and Ma (2024) integrated UQ into the GAN architecture with data augmentation techniques to improve model stability and performance. MC dropout has also found widespread use across a variety of applications due to its simplicity and effectiveness. For example, Tölle et al. (2024) introduced a federated uncertainty-weighted averaging method, combining Bayesian methods with MC dropout to address diverse label distributions in federated learning systems. Similarly, an uncertainty-driven dropout was introduced by Feng et al. (2021) to enhance the robustness of Graph Neural Networks (GNN) through stochastic techniques.

Loss Modification. Modifying loss functions enables the explicit incorporation of uncertainty into ML models, enhancing their robustness and reliability. For instance, Warburg et al. (2020) introduced a Bayesian triplet loss to generate stochastic embeddings for image retrieval tasks. Similarly, Do et al. (2021) proposed loss modifications for semi-supervised learning that adapt to uncertain labels, effectively reducing overfitting to noisy data. In the context of GNNs, Feng et al. (2021) developed adaptive loss functions that respond dynamically to adversarial perturbations based on uncertainty. Caldeira and Nord (2020) investigated how different loss functions, such as Mean Squared Error (MSE) and log-likelihood, influence uncertainty estimation and model robustness in Bayesian frameworks.

These approaches highlight the growing focus on embedding UQ directly into model design, presenting potential avenues of research that can significantly improve the design and development of robust, tailored UQ methods for model training in healthcare applications.

2.3. Model Evaluation

The evaluation stage is key for assessing how UQ translates into real-world insights, focusing on model confidence, calibration, and robustness. Uncertainty decomposition techniques can also help distinguish between epistemic and aleatoric uncertainty, offering deeper insights into model behavior and guiding clinical improvements at the inference stage.

2.3.1. Applications

Recent advances have focused on improving predictive uncertainty estimation, Out-Of-Distribution (OOD) detection, and interpretability. For instance, Hu et al. (2021a) decomposed prediction error into random and systematic components, proposing a two-step method that estimates target labels and error magnitude, evaluating the method on an MRI reconstruction task. Teichmann et al. (2024) introduced a statistical method for OOD detection and for improving the precision of contouring target structures and organs-at-risk, showing that epistemic uncertainty estimation is highly effective for radiotherapy workflows. Kushibar et al. (2022) introduced an image-level uncertainty metric to improve uncertainty estimation in segmentation tasks compared to the commonly used pixel-wise metrics such as entropy and variance, validating their method on oncology and cardiology applications. To help in medical image understanding, Chen et al. (2024a) proposed an efficient conformal prediction method along with an uncertainty explanation method to identify the most influential training samples, offering a more interpretable uncertainty estimate for organs and blood imaging datasets.

In orthopedic imaging applications, Skärström et al. (2024) aligned model uncertainty estimates with intra-reader variability, demonstrating reliability comparable to human annotators, and providing calibrated uncertainty maps to enhance interpretability in vertebral fracture assessment. Yang et al. (2022b) considered the information present in annotations introducing a multi-confidence mask, to predict regions with varying uncertainty levels in lung nodule segmentation, suggesting that regions causing segmentation uncertainty are not random but are related to disagreements in radiologist annotations. Similarly, Konuk et al. (2024) argued that current uncertainty evaluation metrics fall short in clinical contexts, and proposed an evaluation framework to inform joint human-AI systems. To tackle overconfident predictions, Popordanoska et al. (2021) investigated the relationship between calibrated predictions and volume estimation in medical image segmentation, validating their findings on glioma and ischemic stroke lesion volume estimation. For example, to address the lack of UQ methods that are adapted to precision medicine, Durso-Finley et al. (2023b) used Bayesian deep learning to assess model uncertainty in MRI scans for multiple sclerosis, correlating predictive uncertainty with treatment options to enhance clinical decision-making. Finally, in an effort to improve the evaluation of UQ methods on real-world applications, Band et al. (2022) built an open-source benchmark for diabetic retinopathy detection tasks. Their benchmark uses a set of task-specific reliability and performance metrics to evaluate Bayesian methods on safety-critical scenarios, reflecting the complexities of real-world clinical data.

In EHR-based applications, Horii and Chikahara (2023) estimated heterogeneous treatment effects using a Bayesian Gaussian-process-based partially linear model, enabling fine-grained UQ in observational data. Deng et al. (2025) integrated variational dropout and deep ensembles to enhance both calibration and counterfactual decision-making. To capture temporal dynamics, Hess et al. (2024) modeled patient trajectories through a Bayesian neural controlled differential equation framework, quantifying both model and outcome uncertainty. Additionally, Huang et al. (2024b) employed Gaussian random fuzzy numbers within an evidential regression model to simultaneously estimate epistemic and aleatoric uncertainties for time-to-event prediction.

2.3.2. Opportunities

Post-hoc Calibration. Methods such as test-time augmentation are used to improve model calibration by generating diverse inputs for model evaluation to improve generalization (Hekler et al., 2023). Dirichlet-based models also help adjust the model's output by recalibrating probabilities (Shen et al., 2023; Kopetzki et al., 2021). Additionally, Li et al. (2022b) proposed a response-scaling method of the input to improve the numerical stability of UQ methods, and enhancing the overall reliability of the predictions.

Bayesian, Ensemble and Probabilistic Methods. Monte Carlo methods have been used extensively across different domains to provide uncertainty estimates by simulating dropout during inference (Zheng et al., 2021; Bethell et al., 2024; Oberdiek et al., 2022; Wagh et al., 2022). Moreover, Yao et al. (2020) used Bayesian stacking during inference to construct a weighted average of posterior distributions. Additionally, Dai et al. (2023) designed loss functions that integrate uncertainty consistency with Bayesian ensemble methods, enabling robust pseudo-labeling and improved performance in settings with limited supervision.

2.4. Comparative Evaluation of UQ Methods in Clinical Applications

In this section, we compare and highlight seven widely used UQ techniques, discussing strengths, limitations, and suitability for healthcare applications based on key criteria such as predictive performance, calibration, scalability, robustness, and computational cost.

BNNs model both epistemic and aleatoric uncertainty but are computationally intensive and difficult to scale (Antorán et al., 2021; Morales-Álvarez et al., 2021). They have been applied to medical image classification, disease progression modeling, and decision support (Adams and Elhabian, 2023; Zhao et al., 2024b). GPs provide well-calibrated estimates but do not scale well to large datasets. In healthcare, they are used for disease progression prediction, time-series forecasting, and biomarker discovery (Wang and Rockova, 2020; Peluso et al., 2024). Ensemble methods quantify uncertainty via inter-model variability, improving robustness in image segmentation, classification, and anomaly detection, though they lack a principled Bayesian formulation and are computationally costly (Dusenberry et al., 2020; Vranken et al., 2021). Evidential Deep Learning captures both types of uncertainty in a single pass and is used in autonomous diagnostics and decision support, but it may yield overconfident predictions without proper regularization (Shi et al., 2024; Hung et al., 2024). Conformal Prediction offers distribution-free confidence intervals based on past errors and is applied in risk assessment and predictive modeling, although its guarantees rely on the assumption that past data distributions hold (Dutta et al., 2023; Stutts et al., 2023). Bayesian Deep Ensembles improve calibration and epistemic uncertainty estimation over standard ensembles, but their high computational cost limits real-time use (Wilson and Izmailov, 2020). Monte Carlo Dropout approximates Bayesian inference with lower overhead and is widely used in imaging and predictive modeling, though its estimates depend on the choice of dropout rate and may not fully capture uncertainty (Abdar et al., 2021; Bethell et al., 2024). While many of these methods demonstrate strong empirical performance, their practical adoption in healthcare requires careful consideration of trade-offs between computational cost, scalability, interpretability, and robustness to ensure reliable clinical deployment.

Our systematic analysis reveals that strategically integrating UQ into ML pipelines in healthcare, whether during data processing, model training, or evaluation, has tremendous potential to enhance clinical workflow efficiency and ensures that uncertainty is addressed at the right stage.

3. Open Challenges and Future Research

Expanding Clinical Dataset Diversity

Despite a growing use of UQ methods in healthcare, most applications remain concentrated on medical imaging, particularly MRI (Bernard et al., 2018), CT (Heller et al., 2021), and X-rays (Nguyen et al., 2020), limiting generalizability across other healthcare domains. While relevant imaging applications such as skin cancer detection (Ren et al., 2024a) and brain tumor segmentation (Fuchs et al., 2021) are well-studied, research into other modalities, such as sensor data and ECG, remains limited. Decentralized learning approaches, including federated, swarm, and split learning, can offer privacy-preserving solutions for efficient datasharing in healthcare AI (Antunes et al., 2022). Key directions include developing standardized UQ frameworks for decentralized settings, improving calibration across heterogeneous non-iid datasets, and designing lightweight UQ methods to mitigate communication and computational overhead (Antunes et al., 2022; Nguyen et al., 2022). While federated learning applications exist in healthcare (Nguyen et al., 2022), further research is needed to fully extend UQ into decentralized frameworks.

Future work should prioritize diversifying datasets and clinical tasks to broaden UQ applicability (Loftus et al., 2022) and explore multimodal UQ methods (Dutta et al., 2023; Jung et al., 2024) for more robust and realistic clinical models.

Analyzing Sources of Uncertainty

Most research focuses on quantifying predictive uncertainty (Lakshminarayanan et al., 2017), but few studies address understanding its origins, an essential step for trustworthy AI systems. Efforts should focus on identifying whether uncertainty stems from data noise (Alizadehsani et al., 2024), model specification issues (Do et al., 2021), or training data limitations (Huang et al., 2022a; MacDonald et al., 2023). Understanding these sources can guide better clinical decision support and model design.

Building a Unified Framework for UQ Across the ML Pipeline

Our analysis shows that UQ is typically applied in isolation at different ML stages, particularly during training (Abboud et al., 2024; Aljuhani et al., 2022), with little integration across preprocessing (Angelopoulos et al., 2022) and evaluation (Hu et al., 2021b). This fragmentation leads to uncertainty propagation and compounded errors (Valdenegro-Toro et al., 2024). A unified pipeline-oriented approach would enable systematic uncertainty management, categorizing and evaluating methods at each ML stage, identifying gaps where specific uncertainties remain unaddressed (Gruber et al., 2023; Jürgens et al., 2024) and guiding the development of holistic solutions.

Developing Tailored UQ Methods for Healthcare

Current UQ studies often focus on empirical gains without advancing theoretical foundations or understanding limitations in clinical contexts (Durso-Finley et al., 2023b; Teichmann et al., 2024). A balanced approach addressing both theory and application is critical. Further refinement of popular methods such as deep ensembles (Abdollahi et al., 2021; Gu et al., 2021), MC dropout (Bethell et al., 2024), and BNNs (Herzog et al., 2020) is needed, alongside development of novel adaptations for underexplored healthcare domains.

Enhancing Interpretability in UQ for Healthcare

Interpretability of UQ in clinical settings remains underdeveloped. Although uncertainty and noise are often intertwined, distinguishing true uncertainty (i.e., aleatoric, epistemic) from noise due to data variations (e.g., incorrect measurements, missing labels) or model training (e.g., parameter selection) is crucial for actionable insights (Xiang et al., 2022b; Zhang et al., 2023a). Yet, no consensus exists on clinically meaningful UQ metrics, and many approaches lack clinician input. Research still heavily focuses on training-stage uncertainty, with limited exploration at inference and deployment, where clinical decisions occur (Angelopoulos et al., 2022; Kushibar et al., 2022; Konuk et al., 2024; Leibig et al., 2022). Stronger interdisciplinary collaboration between AI researchers and clinicians is needed, to ensure UQ methods deliver clinically actionable information.

Mitigating Fairness and Bias Challenges

Bias in AI-driven healthcare arises from multiple sources, including data (e.g., underrepresentation of certain populations), algorithmic (e.g., modeling choices amplifying disparities), and selection biases (e.g., systematic exclusions in data collection) (Tripepi et al., 2010; Gianfrancesco et al., 2018; Chen et al., 2024b). Evaluating uncertainty across demographic subgroups can reveal discrepancies in model confidence, identifying populations for which predictions are systemically unreliable (Bozkurt et al., 2020), therefore adjusting decision thresholds to ensure consistent predictive performance across diverse patient cohorts and clinical settings (Ojha et al., 2025). While efforts to mitigate bias are gaining traction, current UQ methods often lack systematic evaluations across diverse demographic groups. Future research should prioritize stratified uncertainty analyses to ensure models maintain consistent reliability across age, sex, ethnicity, and disease subpopulations.

Ensuring Safety and Risk Management in Clinical Applications

The risk profile of healthcare applications dictates the required level of UQ rigor, reliability and interpretability, since decisions can have life-altering consequences for patients (Huang et al., 2024a). In wellness applications (e.g., fitness trackers, general health monitoring), UQ can improve transparency as errors have lower stakes despite the potential of misleading information for the user (Zou et al., 2023). In safety-critical domains such as radiology or surgery, UQ must provide high reliability and clinical guarantees (Khalighi et al., 2024). Regulatory frameworks should differentiate between wellness tools and safety-critical AI, enforcing stronger UQ integration where patient safety is more critical. Adaptive methods such as conformal prediction and deep ensembles (Zhou et al., 2024; Thompson et al., 2025) can support real-time clinical decision-making, ensuring uncertainty aligns with dynamic clinical contexts. In addition, thresholds for uncertainty alarms should be adapted based on clinical application and risk level, with regulatory standards working towards enforcing safeguards to ensure AI reliability in safety-critical settings.

Establishing a Standardized Framework for Evaluation

A structured framework to evaluate UQ methods in healthcare is essential, given the diversity of clinical tasks, datasets, and models (Barandas et al., 2024b; Seoni et al., 2023a; Lanini et al., 2024). To address challenges such as inconsistent evaluation metrics (Seoni et al., 2023a), limited generalizability (Barandas et al., 2024b), and lack of OOD benchmarking (Barandas et al., 2024b), we propose a framework comprising four core components. The **first component** focuses on standardized performance-based metrics for uncertainty-aware predictions. Although many studies assess UQ indirectly via task performance (e.g., classification, segmentation) (Barandas et al., 2024b; Tabarisaadi et al., 2024), evaluation protocols should consistently report traditional metrics such as accuracy, precision-recall, F1-score, and AUC. The **second component** emphasizes human-machine interaction metrics, particularly selective prediction. UQ enables selective deferral of uncertain cases to human experts, improving clinical decision-making (Barandas et al., 2024b). Comparative evaluations should assess deferral strategies and their impact, especially in high-risk scenarios. The **third component** addresses calibration analysis to ensure the clinical reliability of uncertainty estimates. Well-calibrated predictions are critical to prevent misleading outputs (Barandas et al., 2024b; Lanini et al., 2024; Chen et al., 2024c). Metrics such as Expected Calibration Error (ECE) and Brier scores should be systematically reported across clinical tasks, yet are often overlooked (Chen et al., 2024c; Xia et al., 2023). The **fourth component** involves OOD detection to assess model robustness under distributional shifts. Given the frequent domain shifts in healthcare applications, evaluation should explicitly test models' ability to distinguish in-distribution from OOD samples (Barandas et al., 2024b; Xia et al., 2023). Methods such as controlled data perturbations (Xia et al., 2023) can facilitate systematic OOD benchmarking on clinical datasets. While domain-specific adaptations may be necessary, adopting a standardized evaluation framework would substantially improve reproducibility, robustness, and trustworthiness of UQ methods, particularly for long-term clinical deployment.

Integrating UQ into Regulatory Frameworks

Current AI healthcare regulations emphasize transparency, reliability, risk management, and patient safety principles, although they rarely explicitly mention UQ (Schmidt et al., 2024). The FDA, Health Canada, and the UK's MHRA have issued "*Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles*," emphasizing interpretability, reliability, and adaptability, which are core principles underlying the objectives of UQ (Food et al., 2024). Additionally, the World Health Organization (WHO) has highlighted the importance of regulatory oversight in AI-driven health applications, emphasizing transparency and risk mitigation as central elements (Tsaneva-Atanasova et al., 2025). In particular, UQ can optimize workflows by enabling uncertainty-adapted triage and optimization risk stratification, ensuring ambiguous cases receive additional scrutiny before critical decisions are made. Experts have also argued that AI regulations should explicitly require uncertainty-aware metrics to ensure the safe deployment of AI models, alongside task-specific continuous monitoring protocols (Chua et al., 2023). In the future, holistic regulatory frameworks should include benchmarks for calibration, OOD detection, and selective deferral.

Addressing Deployment Challenges

While UQ in healthcare AI has been extensively studied from a theoretical perspective, its real-world clinical deployment remains limited. Most current research emphasizes potential benefits and challenges but offers limited practical implementation strategies.

Several key barriers must be addressed to translate UQ advances into clinical practice. First, real-time clinical decision support requires UQ methods that are computationally efficient and scalable, capabilities that many existing techniques lack (Verma et al., 2021). Second, healthcare data quality and accessibility issues, such as noise, incompleteness, and privacy restrictions, complicate the development of reliable uncertainty estimates (Zhang et al., 2022a). Decentralized learning approaches offer promising solutions by enabling robust training without sharing raw data across institutions (Yuan, 2024). Third, the lack of open science practices, including limited code and model sharing (Figure 1c), hinders transparency, reproducibility, and comparative evaluation. Finally, the absence of standardized UQ evaluation frameworks in healthcare leads to inconsistencies across studies, complicating clinical translation. Addressing these barriers requires closer collaboration between ML researchers and clinicians to align UQ with real-world clinical needs and operational constraints.

Focused efforts on computational efficiency, data robustness, open benchmarking, and standardized evaluation can significantly advance the integration of UQ into clinical AI workflows, ultimately improving trust, reliability, and patient outcomes. Addressing these challenges highlights the need for a structured roadmap to guide future research and facilitate the practical deployment of uncertainty-aware AI systems in clinical settings.

Defining a Roadmap for Future Research

Building on the identified challenges and regulatory considerations, the shift from theoretical UQ research development to real-world clinical deployment requires strategic advancements in evaluation, interpretability, communication, and integration into decision-support systems. We outline four actionable research directions to facilitate this transition:

1. **Standardizing UQ Evaluation Metrics:** The lack of consistent evaluation metrics across clinical tasks hinders comparability. Reporting guidelines should mandate standardized assessment of uncertainty calibration, coverage error, out-of-distribution OOD detection, ensuring that uncertainty is evaluated alongside performance metrics in a structured manner.
2. **Contextualizing UQ with Task-Specific Safety Thresholds:** UQ must be aligned with the safety-critical nature of specific clinical applications. Regulatory bodies and clinical experts should define impact-sensitive uncertainty thresholds to ensure that AI models meet appropriate patient safety standards.
3. **Enhancing Interpretability and Clinical Trust in UQ:** Uncertainty estimates must be both clinically meaningful and interpretable. Developing intuitive visualizations and fostering close collaboration with clinicians on UQ interpretation can bridge the gap between AI model outputs and real-world clinical decision-making.
4. **Integrating UQ into Assistive AI and Decision Support:** AI systems should leverage UQ to highlight high-uncertainty cases, allowing clinicians to exercise greater caution where needed. Future efforts should prioritize interpretable, impact-sensitive UQ methods developed collaboratively with clinicians and patients to ensure practical utility.

4. Conclusion

In this survey, we reviewed and synthesized recent advancements in UQ methods for healthcare, providing a comprehensive analysis of their application across the ML pipeline. We discussed popular methodologies, key clinical domains, relevant medical datasets, and outlined current challenges alongside promising future research directions. We emphasize the need for an integrated and systematic approach to incorporate UQ across all stages of model development for clinical applications. We encourage researchers to evaluate proposed algorithms across diverse medical datasets, integrate UQ techniques throughout the ML pipeline, and conduct detailed analyses of uncertainty sources to more effectively mitigate them. Our recommendations aim to bridge existing research gaps and guide future work in UQML4H, ultimately supporting the development of trustworthy, reliable, and clinically meaningful ML systems.

Acknowledgements

This work was supported by the NYUAD Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award CG010.

References

Zeinab Abboud, Herve Lombaert, and Samuel Kadoury. Sparse bayesian networks: Efficient uncertainty quantification in medical image analysis. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 675–684, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72117-5.

Moloud Abdar, Maryam Samami, Sajjad Dehghani Mahmoodabad, Thang Doan, Bogdan Mazoure, Reza Hashemifesharaki, Li Liu, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. Uncertainty quantification in skin cancer classification using three-way decision-based bayesian deep learning. *Computers in Biology and Medicine*, 135:104418, 2021. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2021.104418. URL https://www.sciencedirect.com/science/article/pii/S0010482521002122.

Moloud Abdar, Mohammad Amin Fahami, Leonardo Rundo, Petia Radeva, Alejandro F. Frangi, U. Rajendra Acharya, Abbas Khosravi, Hak-Keung Lam, Alexander Jung, and Saeid Nahavandi. Hercules: Deep hierarchical attentive multilevel fusion model with uncertainty quantification for medical image classification. *IEEE Transactions on Industrial Informatics*, 19(1):274–285, 2023. doi: 10.1109/TII.2022.3168887.

Jafar Abdollahi, Babak Nouri-Moghaddam, and Mehdi Ghazanfari. Deep neural network based ensemble learning algorithms for the healthcare system (diagnosis of chronic diseases). *arXiv preprint arXiv:2103.08182*, 2021.

Adam Abeshouse, Clement Adebamowo, Sally N Adebamowo, Rehan Akbani, Teniola Akeredolu, Adrian Ally, Matthew L Anderson, Pavana Anur, Elizabeth L Appelbaum, Joshua Armenia, et al. Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell*, 171(4):950–965, 2017.

Zaid Abulawi, Rui Hu, Prasanna Balaprakash, and Yang Liu. Bayesian optimized deep ensemble for uncertainty quantification of deep neural networks: A system safety case study on sodium fast reactor thermal stratification modeling. *arXiv preprint arXiv:2412.08776*, 2024. URL https://arxiv.org/abs/2412.08776.

Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in brief*, 30:105474, 2020.

Jadie Adams and Shireen Y Elhabian. Fully bayesian vib-deepssm. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 346–356. Springer, 2023.

Edmonmd Adib, Amanda S Fernandez, Fatemeh Afghah, and John J Prevost. Synthetic ecg signal generation using probabilistic diffusion models. *IEEe Access*, 11:75818–75828, 2023.

Ben Adlam, Jaehoon Lee, Lechao Xiao, Jeffrey Pennington, and Jasper Snoek. Exploring the uncertainty properties of neural networks' implicit priors in the infinite-width limit. *ArXiv*, abs/2010.07355, 2020. URL https://api.semanticscholar.org/CorpusID:222378172.

Sabbir Ahmed, Mohammad Abu Yousuf, Muhammad Mostafa Monowar, Abdul Hamid, and Madini O. Alassafi. Taking all the factors we need: A multimodal depression classification with uncertainty approximation. *IEEE Access*, 11:99847–99861, 2023. doi: 10.1109/ACCESS.2023.3315243.

Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, et al. Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. *Physiological measurement*, 41(12):124003, 2020.

Daniel C Alexander, Darko Zikic, Aurobrata Ghosh, Ryutaro Tanno, Viktor Wottschel, Jiaying Zhang, Enrico Kaden, Tim B Dyrby, Stamatios N Sotiropoulos, Hui Zhang, et al. Image quality transfer and applications in diffusion mri. *NeuroImage*, 152:283–298, 2017.

Roohallah Alizadehsani, Mohamad Roshanzamir, Sadiq Hussain, Abbas Khosravi, Afsaneh Koohestani, Mohammad Hossein Zangooei, Moloud Abdar, Adham Beykikhoshk, Afshin Shoeibi, Assef Zare, Maryam Panahiazar, Saeid Nahavandi, Dipti Srinivasan, Amir F. Atiya, and U. Rajendra Acharya. Handling of uncertainty in medical data using machine learning and probability theory techniques: a review of 30 years (1991–2020). 339(3):1077–1118, 2024. doi: 10.1007/s10479-021-04006-2. URL https://link.springer.com/10.1007/s10479-021-04006-2.

Asmaa Aljuhani, Ishya Casukhela, Jany Chan, David Liebner, and Raghu Machiraju. Uncertainty aware sampling framework of weak-label learning for histology image classification. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 366–376, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16434-7.

Shuroug A Alowais, Sahar S Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I Alshaya, Sumaya N Almohareb, Atheer Aldairem, Mohammed Alrashed, Khalid Bin Saleh, Hisham A Badreldin, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*, 23(1):689, 2023.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. URL https://arxiv.org/abs/2107.07511.

Anastasios N Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pages 717–730. PMLR, 2022.

Javier Antorán, Adrian Weller, Umang Bhatt, Tameem Adel, and José Miguel Hernández-Lobato. GETTING a CLUE: A METHOD FOR EXPLAINING UNCERTAINTY ESTIMATES. 2021.

Rodolfo Stoffel Antunes, Cristiano André da Costa, Arne Küderle, Imrana Abdullahi Yari, and Björn Eskofier. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–23, 2022.

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

Gundeep Arora, Srujana Merugu, Anoop S V K K Saladi, and Rajeev Rastogi. Leveraging uncertainty estimates to improve classifier performance. 2024. URL https://www.amazon.science/publications/leveraging-uncertainty-estimates-to-improve-classifier-performance.

Awais Ashfaq, Markus Lingman, Murat Sensoy, and Sławomir Nowaczyk. Deed: Deep evidential doctor. *Artificial Intelligence*, 325:104019, 2023.

Mohammad Azizmalayeri, Ameen Abu-Hanna, and Giovanni Cinà. Unmasking the chameleons: A benchmark for out-of-distribution detection in medical tabular data. *International Journal of Medical Informatics*, 195:105762, 2025. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2024.105762. URL https://www.sciencedirect.com/science/article/pii/S1386505624004258.

Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13117–13126, 2021. doi: 10.1109/ICCV48922.2021.01289.

Chenjia Bai, Lingxiao Wang, Jianye Hao, Zhuoran Yang, Bin Zhao, Zhen Wang, and Xuelong Li. Pessimistic value iteration for multi-task data sharing in offline reinforcement learning. *Artificial Intelligence*, 326:104048, 2024. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2023.104048. URL https://www.sciencedirect.com/science/article/pii/S0004370223001947.

Neil Band, Tim GJ Rudner, Qixuan Feng, Angelos Filos, Zachary Nado, Michael W Dusenberry, Ghassen Jerfel, Dustin Tran, and Yarin Gal. Benchmarking bayesian deep learning on diabetic retinopathy detection tasks. *arXiv preprint arXiv:2211.12717*, 2022.

Marília Barandas, Lorenzo Famiglini, Andrea Campagner, Duarte Folgado, Raquel Simão, Federico Cabitza, and Hugo Gamboa. Evaluation of uncertainty quantification methods in multi-label classification: A case study with automatic diagnosis of electrocardiogram. *Information Fusion*, 101: 101978, 2024a. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101978. URL https://www.sciencedirect.com/science/article/pii/S1566253523002944.

Marília Barandas, Lorenzo Famiglini, Andrea Campagner, Duarte Folgado, Raquel Simão, Federico Cabitza, and Hugo Gamboa. Evaluation of uncertainty quantification methods in multi-label classification: A case study with automatic diagnosis of electrocardiogram. *Information Fusion*, 101: 101978, 2024b. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101978. URL https://www.sciencedirect.com/science/article/pii/S1566253523002944.

Riccardo Barbano, Simon Arridge, Bangti Jin, and Ryutaro Tanno. Chapter 26 - uncertainty quantification in medical image synthesis. In Ninon Burgos and David Svoboda, editors, *Biomedical Image Synthesis and Simulation*, The MICCAI Society book Series, pages 601–641. Academic Press, 2022. ISBN 978-0-12-824349-7. doi: https://doi.org/10.1016/B978-0-12-824349-7.00033-5. URL https://www.sciencedirect.com/science/article/pii/B9780128243497000335.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11): 2514–2525, 2018.

Lucas Berry and David Meger. Normalizing flow ensembles for rich aleatoric and epistemic uncertainty modeling. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i6.25834. URL https://doi.org/10.1609/aaai.v37i6.25834.

Daniel Bethell, Simos Gerasimou, and Radu Calinescu. Robust uncertainty quantification using conformalised monte carlo prediction. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i19.30084. URL https://doi.org/10.1609/aaai.v38i19.30084.

Grigor Bezirganyan. Data and decision fusion with uncertainty quantification for ml-based healthcare decision systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5169–5172, 2023.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023a.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023b.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1613–1622. JMLR.org, 2015.

Ralf Bousseljot, Dieter Kreiseler, and Allard Schnabel. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. 1995. doi: 10.1515/bmte.1995.40.s1.317.

Selen Bozkurt, Eli M Cahan, Martin G Seneviratne, Ran Sun, Juan A Lossio-Ventura, John PA Ioannidis, and Tina Hernandez-Boussard. Reporting of demographic data and representativeness in machine learning models using electronic health records. *Journal of the American Medical Informatics Association*, 27(12):1878–1884, 2020.

Edward De Brouwer, Javier González Hernández, and Stephanie Hyland. Predicting the impact of treatments over time with uncertainty aware neural differential equations, 2022. URL https://arxiv.org/abs/2202.11987.

Katherine E. Brown, Farzana Ahamed Bhuiyan, and Douglas A. Talbert. Uncertainty quantification in multimodal ensembles of deep learners. In *The Florida AI Research Society*, 2020. URL https://api.semanticscholar.org/CorpusID:219319991.

James Browning, Micha Kornreich, Aubrey Chow, Jayashri Pawar, Li Zhang, Richard Herzog, and Benjamin L Odry. Uncertainty aware deep reinforcement learning for anatomical landmark detection in medical images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 636–644. Springer, 2021.

Thomas Buddenkotte, Lorena Escudero Sanchez, Mireia Crispin-Ortuzar, Ramona Woitek, Cathal McCague, James D. Brenton, Ozan Öktem, Evis Sala, and Leonardo Rundo. Calibrating ensembles for scalable uncertainty quantification in deep learning-based medical image segmentation. *Computers in Biology and Medicine*, 163:107096, 2023. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2023.107096. URL https://www.sciencedirect.com/science/article/pii/S0010482523005619.

Yu Cai, Hao Chen, Xin Yang, Yu Zhou, and Kwang-Ting Cheng. Dual-distribution discrepancy for anomaly detection in chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 584–593. Springer, 2022.

João Caldeira and Brian Nord. Deeply uncertain: comparing methods of uncertainty quantification in deep learning algorithms. *Machine Learning: Science and Technology*, 2(1):015002, dec 2020. doi: 10.1088/2632-2153/aba6f3. URL https://dx.doi.org/10.1088/2632-2153/aba6f3.

Victor M Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021.

Vincent Casser, Kai Kang, Hanspeter Pfister, and Daniel Haehn. Fast mitochondria detection for connectomics. In *Medical Imaging with Deep Learning*, pages 111–120. PMLR, 2020.

Aobo Chen, Yangyi Li, Wei Qian, Kathryn Morse, Chenglin Miao, and Mengdi Huai. Modeling and understanding uncertainty in medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 557–567. Springer, 2024a.

Cheng Chen, Quande Liu, Yueming Jin, Qi Dou, and Pheng-Ann Heng. Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 225–235. Springer, 2021.

Feng Chen, Liqin Wang, Julie Hong, Jiaqi Jiang, and Li Zhou. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *Journal of the American Medical Informatics Association*, 31(5):1172–1183, 2024b.

Zizhang Chen, Peizhao Li, Xiaomeng Dong, and Pengyu Hong. Uncertainty quantification for clinical outcome predictions with (large) language models, 2024c. URL https://arxiv.org/abs/2411.03497.

Michelle Chua, Doyun Kim, Jongmun Choi, Nahyoung G Lee, Vikram Deshpande, Joseph Schwab, Michael H Lev, Ramon G Gonzalez, Michael S Gee, and Synho Do. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, 7(6):711–718, 2023.

Nathalie Conrad, Andrew Judge, Jenny Tran, Hamid Mohseni, Deborah Hedgecott, Abel Perez Crespillo, Moira Allison, Harry Hemingway, John G Cleland, John JV McMurray, et al. Temporal trends and patterns in heart failure incidence:

a population-based study of 4 million individuals. *The Lancet*, 391(10120):572–580, 2018.

Nico Daheim, Thomas Möllenhoff, Edoardo Maria Ponti, Iryna Gurevych, and Mohammad Emtiyaz Khan. Model merging by uncertainty-based gradient matching. *arXiv preprint arXiv:2310.12808*, 2023.

Weihang Dai, Xiaomeng Li, and Kwang-Ting Cheng. Semi-supervised deep regression with uncertainty consistency and variational model ensembling via bayesian neural networks. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i6.25890. URL https://doi.org/10.1609/aaai.v37i6.25890.

Abhijit Das, Vandan Gorade, Komal Kumar, Snehashis Chakraborty, Dwarikanath Mahapatra, and Sudipta Roy. Confidence-guided semi-supervised learning for generalized lesion localization in x-ray images. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 242–252, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72378-0.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6): 1293–1305, 2019.

Aysen Degerli, Morteza Zabihi, Serkan Kiranyaz, Tahir Hamid, Rashid Mazhar, Ridha Hamila, and Moncef Gabbouj. Early detection of myocardial infarction in low-quality echocardiography. *IEEE Access*, 9:34442–34453, 2021.

Danruo Deng, Guangyong Chen, Yang Yu, Furui Liu, and Pheng-Ann Heng. Uncertainty estimation by fisher information-based evidential deep learning. *arXiv preprint arXiv:2303.02045*, 2023. URL https://arxiv.org/abs/2303.02045.

Leon Deng, Hong Xiong, Feng Wu, Sanyam Kapoor, Soumya Gosh, Zach Shahn, and Li-wei Lehman. Uncertainty quantification for conditional treatment effect estimation under dynamic treatment regimes. In Stefan Hegselmann, Helen Zhou, Elizabeth Healey, Trenton Chang, Caleb Ellington, Vishwali Mhasawade, Sana Tonekaboni, Peniel Argaw, and Haoran Zhang, editors, *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pages 248–266. PMLR, 15–16 Dec 2025. URL https://proceedings.mlr.press/v259/deng25a.html.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15. Springer, 2000. URL https://link.springer.com/chapter/10.1007/3-540-45014-9_1.

Thomas G. Dietterich. Uncertainty quantification in machine learning. https://web.engr.oregonstate.edu/~tgd/talks/dietterich-uncertainty-quantification-in-machine-learning-final.pdf, 2024.

Ze Yang Ding, Junn Yong Loo, Vishnu Monn Baskaran, Surya Girinatha Nurzaman, and Chee Pin Tan. Predictive uncertainty estimation using deep learning for soft robot multimodal sensing. *IEEE Robotics and Automation Letters*, 6(2): 951–957, 2021. doi: 10.1109/LRA.2021.3056066.

Kien Do, Truyen Tran, and Svetha Venkatesh. Semi-supervised learning with variational bayesian inference and maximum uncertainty regularization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:7236–7244, 05 2021. doi: 10.1609/aaai.v35i8.16889.

Zhicheng Dong, Xiaodong Yue, Yufei Chen, Xujing Zhou, and Jiye Liang. Uncertainty-aware multiview learning for prostate cancer grading with dwi. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 739–748. Springer, 2024.

Emma Dugas, Jorge Jared, and Will Cukierski. Diabetic retinopathy detection (2015). *URL https://kaggle. com/competitions/diabetic-retinopathy-detection*, 7, 2015.

Joshua Durso-Finley, Jean-Pierre Falet, Raghav Mehta, Douglas L. Arnold, Nick Pawlowski, and Tal Arbel. Improving image-based precision medicine with uncertainty-aware causal models. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 472–481, Cham, 2023a. Springer Nature Switzerland. ISBN 978-3-031-43904-9.

Joshua Durso-Finley, Jean-Pierre Falet, Raghav Mehta, Douglas L. Arnold, Nick Pawlowski, and Tal Arbel. Improving image-based precision medicine with uncertainty-aware causal models, 2023b. URL https://arxiv.org/abs/2305.03829.

Michael W. Dusenberry, Dustin Tran, Edward Choi, Jonas Kemp, Jeremy Nixon, Ghassen Jerfel, Katherine Heller, and Andrew M. Dai. Analyzing the role of model uncertainty for electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 204–213, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370462. doi: 10.1145/3368555.3384457. URL https://doi.org/10.1145/3368555.3384457.

Shiladitya Dutta, Hongbo Wei, Lars van der Laan, and Ahmed M. Alaa. Estimating uncertainty in multimodal foundation models using public internet data. 2023. URL http://arxiv.org/abs/2310.09926. Publisher: arXiv.

Vineet Edupuganti, Morteza Mardani, Shreyas Vasanawala, and John Pauly. Uncertainty quantification in deep MRI reconstruction. 40(1):239–250, 2021. doi: 10.1109/TMI.2020.3025065.

EyePACS. Diabetic retinopathy detection dataset, 2015. URL https://www.kaggle.com/c/diabetic-retinopathy-detection. Accessed: 2025-01-15.

Boyuan Feng, Yuke Wang, and Yufei Ding. Uag: Uncertainty-aware attention graph neural network for defending adversarial attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35 (8):7404–7412, May 2021. doi: 10.1609/aaai.v35i8.16908. URL https://ojs.aaai.org/index.php/AAAI/article/view/16908.

Matteo Figini, Marco Riva, Mark Graham, Gian Marco Castelli, Bethania Fernandes, Marco Grimaldi, Giuseppe Baselli, Federico Pessina, Lorenzo Bello, Hui Zhang, et al. Prediction of isocitrate dehydrogenase genotype in brain gliomas with mri: single-shell versus multishell diffusion models. *Radiology*, 289(3):788–796, 2018.

Duarte Folgado, Marília Barandas, Lorenzo Famiglini, Ricardo Santos, Federico Cabitza, and Hugo Gamboa. Explainability meets uncertainty quantification: Insights from feature-based model fusion on multimodal time series. 100:101955, 2023. URL https://api.semanticscholar.org/CorpusID:260322702.

Andreas Foltyn and Jessica Deuschel. Towards reliable multimodal stress detection under distribution shift. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 329–333, 2021.

US Food, Drug Administration, et al. Transparency for machine learning-enabled medical devices: guiding principles, 2024.

Jia Fu, Guotai Wang, Tao Lu, Qiang Yue, Tom Vercauteren, Sébastien Ourselin, and Shaoting Zhang. Um-cam: Uncertainty-weighted multi-resolution class activation maps for weakly-supervised segmentation. *Pattern Recognition*, 160:111204, 2025.

Tianfan Fu, Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Hint: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns*, 3(4), 2022.

Moritz Fuchs, Camila Gonzalez, and Anirban Mukhopadhyay. Practical uncertainty quantification for brain tumor segmentation. In *Medical Imaging with Deep Learning*, 2021.

Joseph Futoma. *Gaussian process-based models for clinical time series in healthcare*. PhD thesis, Duke University, 2018.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Ido Galil and Ran El-Yaniv. Disrupting deep uncertainty estimation without harming accuracy. *Advances in Neural Information Processing Systems*, 34:21285–21296, 2021.

Lidia Garrucho, Claire-Anne Reidel, Kaisar Kushibar, Smriti Joshi, Richard Osuala, Apostolia Tsirikoglou, Maciej Bobowicz, Javier del Riego, Alessandro Catanese, Katarzyna Gwoździewicz, et al. Mama-mia: A large-scale multi-center breast cancer dce-mri benchmark dataset with expert segmentations. *arXiv preprint arXiv:2406.13844*, 2024.

Changran Geng, Harald Paganetti, and Clemens Grassberger. Prediction of treatment response for combined chemo-and radiation therapy for non-small cell lung cancer patients using a bio-mathematical model. *Scientific reports*, 7(1):13542, 2017.

Subhankar Ghosh, Taha Belkhouja, Yan Yan, and Janardhan Rao Doppa. Improving uncertainty quantification of deep classifiers via neighborhood conformal prediction: novel algorithm and theoretical analysis. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i6.25936. URL https://doi.org/10.1609/aaai.v37i6.25936.

Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178 (11):1544–1547, 2018.

Ralf Gold, Ludwig Kappos, Douglas L Arnold, Amit Bar-Or, Gavin Giovannoni, Krzysztof Selmaj, Carlo Tornatore, Marianne T Sweetser, Minhua Yang, Sarah I Sheikh, et al. Placebo-controlled phase 3 study of oral bg-12 for relapsing multiple sclerosis. *New England Journal of Medicine*, 367 (12):1098–1107, 2012.

Hao Gong, Lifeng Yu, Shuai Leng, Scott S. Hsieh, Joel G. Fletcher, and Cynthia H. McCollough. Patient-specific uncertainty and bias quantification of non-transparent convolutional neural network model through knowledge distillation and bayesian deep learning. In Lifeng Yu, Rebecca Fahrig, and John M. Sabol, editors, *Medical Imaging 2023*, Progress in Biomedical Optics and Imaging - Proceedings of SPIE. SPIE, 2023a. doi: 10.1117/12.2654318. Publisher Copyright: © COPYRIGHT SPIE. Downloading of the abstract is permitted for personal use only.; Medical Imaging 2023: Physics of Medical Imaging ; Conference date: 19-02-2023 Through 23-02-2023.

Xiaoyu Gong, Shuai Lü, Jiayu Yu, Sheng Zhu, and Zongze Li. Adaptive estimation q-learning with uncertainty and familiarity. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023b. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/417. URL https://doi.org/10.24963/ijcai.2023/417.

Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter, and Göran Kauermann. Sources of uncertainty in machine learning–a statisticians' view. *arXiv preprint arXiv:2305.16703*, 2023.

Ang Nan Gu, Michael Tsang, Hooman Vaseli, Teresa Tsang, and Purang Abolmaesumi. Reliable multi-view learning with conformal prediction for aortic stenosis classification in echocardiography. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 327–337, Cham, 2024a. Springer Nature Switzerland. ISBN 978-3-031-72378-0.

Yi Gu, Yi Lin, Kwang-Ting Cheng, and Hao Chen. Revisiting Deep Ensemble Uncertainty for Enhanced Medical Anomaly Detection . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15006. Springer Nature Switzerland, October 2024b.

Yingqi Gu, Akshay Zalkikar, Mingming Liu, Lara Kelly, Amy Hall, Kieran Daly, and Tomas Ward. Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data. *Scientific Reports*, 11(1):18961, 2021.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond. *arXiv preprint arXiv:2404.10136*, 2024.

Güney Gürsel. Healthcare, uncertainty, and fuzzy logic. *Digital Medicine*, 2(3):101–112, 2016.

Fatemeh Hamedani-KarAzmoudehFar, Reza Tavakkoli-Moghaddam, Amir Reza Tajally, and Seyed Sina Aria. Breast cancer classification by a new approach to assessing deep neural network-based uncertainty quantification methods. *Biomedical Signal Processing and Control*, 79:104057, 2023. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2022.104057. URL https://www.sciencedirect.com/science/article/pii/S1746809422005298.

Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20707–20717, 2022.

Russell C Hardie, Redha Ali, Manawaduge Supun De Silva, and Temesguen Messay Kebede. Skin lesion segmentation and classification for isic 2018 using traditional classifiers with hand-crafted features. *arXiv preprint arXiv:1807.07001*, 2018.

Michael P Harms, Leah H Somerville, Beau M Ances, Jesper Andersson, Deanna M Barch, Matteo Bastiani, Susan Y Bookheimer, Timothy B Brown, Randy L Buckner, Gregory C Burgess, et al. Extending the human connectome project across ages: Imaging protocols for the lifespan development and aging projects. *Neuroimage*, 183:972–984, 2018.

Stephen L Hauser, Amit Bar-Or, Giancarlo Comi, Gavin Giovannoni, Hans-Peter Hartung, Bernhard Hemmer, Fred Lublin, Xavier Montalban, Kottil W Rammohan, Krzysztof Selmaj, et al. Ocrelizumab versus interferon beta-1a in relapsing multiple sclerosis. *New England Journal of Medicine*, 376(3): 221–234, 2017.

Jakob Heiss, Jakob Weissteiner, Hanna Wutte, Sven Seuken, and Josef Teichmann. Nomu: Neural optimization-based model uncertainty. *URL https://arxiv. org/abs/2102.13640*, 8, 2021.

Achim Hekler, Titus J. Brinker, and Florian Buettner. Test time augmentation meets post-hoc calibration: Uncertainty quantification under real-world conditions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14856–14864, Jun. 2023. doi: 10.1609/aaai.v37i12.26735. URL https://ojs.aaai.org/index.php/AAAI/article/view/26735.

Thomas Heldt, Ramakrishna Mukkamala, George B Moody, and Roger G Mark. Cvsim: an open-source cardiovascular simulator for teaching and research. *The open pacing, electrophysiology & therapy journal*, 3:45, 2010.

Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021.

Lukas Herzog, Marco Fraccaro, and Casper Kaae Sønderby. Bayesian neural networks for uncertainty estimation in clinical applications. *NeurIPS Workshop on Machine Learning for Healthcare*, 2020. URL https://arxiv.org/abs/2002.10077.

Konstantin Hess, Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bayesian neural controlled differential equations for treatment effect estimation, 2024. URL https://arxiv.org/abs/2310.17463.

W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020.

Shunsuke Horii and Yoichi Chikahara. Uncertainty quantification in heterogeneous treatment effect estimation with gaussian-process-based partially linear model, 2023. URL https://arxiv.org/abs/2312.10435.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. *arXiv preprint arXiv:2311.08718*, 2023.

Leland S. Hu, Lujia Wang, Andrea Hawkins-Daarud, Jennifer M. Eschbacher, Kyle W. Singleton, Pamela R. Jackson, Kamala Clark-Swanson, Christopher P. Sereduk, Sen Peng, Panwen Wang, Junwen Wang, Leslie C. Baxter, Kris A. Smith, Gina L. Mazza, Ashley M. Stokes, Bernard R. Bendok, Richard S. Zimmerman, Chandan Krishna, Alyx B. Porter, Maciej M. Mrugala, Joseph M. Hoxworth, Teresa Wu, Nhan L. Tran, Kristin R. Swanson, and Jing Li. Uncertainty quantification in radiogenomics: Egfr amplification in glioblastoma. *medRxiv*, 2020. doi: 10.1101/2020.05.22.20110288. URL https://www.medrxiv.org/content/early/2020/05/26/2020.05.22.20110288.

Shi Hu, Nicola Pezzotti, and Max Welling. Learning to predict error for mri reconstruction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 604–613. Springer, 2021a.

Shi Hu, Nicola Pezzotti, and Max Welling. Learning to predict error for MRI reconstruction. pages 604–613. Springer International Publishing, 2021b. doi: 10.1007/978-3-030-87199-4_57.

Yibo Hu, Yuzhe Ou, Xujiang Zhao, Jin-Hee Cho, and Feng Chen. Multidimensional uncertainty-aware evidential neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35: 7815–7822, 05 2021c. doi: 10.1609/aaai.v35i9.16954.

Ling Huang, Su Ruan, Yucheng Xing, and Mengling Feng. A review of uncertainty quantification in medical image analysis: probabilistic and non-probabilistic methods. *Medical Image Analysis*, page 103223, 2024a.

Ling Huang, Yucheng Xing, Swapnil Mishra, Thierry Denoeux, and Mengling Feng. Evidential time-to-event prediction with calibrated uncertainty quantification, 2024b. URL https://arxiv.org/abs/2411.07853.

Ling Huang, Su Ruan, Pierre Decazes, and Thierry Denœux. Deep evidential fusion with uncertainty quantification and reliability learning for multimodal medical image segmentation. *Information Fusion*, 113:102648, 2025. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2024.102648. URL https://www.sciencedirect.com/science/article/pii/S1566253524004263.

Yingsong Huang, Bing Bai, Shengwei Zhao, Kun Bai, and Fei Wang. Uncertainty-aware learning against label noise on imbalanced datasets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36: 6960–6969, 06 2022a. doi: 10.1609/aaai.v36i6.20654.

Zhe Huang, Gary Long, Benjamin Wessler, and Michael C Hughes. Tmed 2: a dataset for semi-supervised classification of echocardiograms. In *In DataPerf: Benchmarking Data for Data-Centric AI Workshop*, 2022b.

Alex Ling Yu Hung, Haoxin Zheng, Kai Zhao, Kaifeng Pang, Demetri Terzopoulos, and Kyunghyun Sung. Cross-slice attention and evidential critical loss for uncertainty-aware prostate cancer detection. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 113–123, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72111-3.

Antonella Iuliano, Pietro Liò, Gilda Manfredi, and Federico Romaniello. Denoising probabilistic diffusion models for synthetic healthcare image generation. In *2024 IEEE International Workshop on Metrology for Living Environment (MetroLivEnv)*, pages 415–420. IEEE, 2024.

Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

V. Jahmunah, E.Y.K. Ng, Ru-San Tan, Shu Lih Oh, and U. Rajendra Acharya. Uncertainty quantification in densenet model using myocardial infarction ecg signals. *Computer Methods and Programs in Biomedicine*, 229:107308, 2023. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2022.107308. URL https://www.sciencedirect.com/science/article/pii/S0169260722006897.

Andras Janosi, William Steinbrunn, Matthias Pfisterer, Robert Detrano, et al. Heart disease data set. *The UCI KDD Archive*, 1988.

Andrew et al. Janowczyk. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7(1):29, 2016.

Minjae Jeong, Hyuna Cho, Sungyoon Jung, and Won Hwa Kim. Uncertainty-aware diffusion-based adversarial attack for realistic colonoscopy image synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 647–658. Springer, 2024.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, pages 451–462. Springer, 2020.

Xinwei Ji, Xiaomin Chang, Wei Li, and Albert Y Zomaya. Unraveling pain levels: A data-uncertainty guided approach for effective pain assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22167–22175, 2024.

Lisa Johansson, Daniel Sundh, Per Magnusson, Komagal Rukmangatharajan, Dan Mellstr"om, Anna G Nilsson, and Mattias Lorentzon. Grade 1 vertebral fractures identified by densitometric lateral spine imaging predict incident major osteoporotic fracture independently of clinical risk factors and bone mineral density in older women. *Journal of Bone and Mineral Research*, 35(10): 1942–1951, 2020.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Liwei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3 (1):1–9, 2016.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1, 2023a.

Daniel D Johnson, Daniel Tarlow, and Christian Walder. Ru-sure? uncertainty-aware code suggestions by maximizing utility across random user intents. *arXiv preprint arXiv:2303.00732*, 2023b.

Thierry Judge, Olivier Bernard, Mihaela Porumb, Agisilaos Chartsias, Arian Beqiri, and Pierre-Marc Jodoin. Crisp-reliable uncertainty estimation for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 492–502. Springer, 2022a.

Thierry Judge, Olivier Bernard, Mihaela Porumb, Agisilaos Chartsias, Arian Beqiri, and Pierre-Marc Jodoin. Crisp - reliable uncertainty estimation for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, page 492–502, Berlin, Heidelberg, 2022b. Springer-Verlag. ISBN 978-3-031-16451-4. doi: 10.1007/978-3-031-16452-1_47. URL https://doi.org/10.1007/978-3-031-16452-1_47.

Thierry Judge, Olivier Bernard, Woo-Jin Cho Kim, Alberto Gomez, Agisilaos Chartsias, and Pierre-Marc Jodoin. Asymmetric contour uncertainty estimation for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 210–220. Springer, 2023a.

Thierry Judge, Olivier Bernard, Woo-Jin Cho Kim, Alberto Gomez, Agisilaos Chartsias, and Pierre-Marc Jodoin. Asymmetric contour uncertainty estimation for medical image segmentation. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 210–220, Cham, 2023b. Springer Nature Switzerland. ISBN 978-3-031-43898-1.

Myong Chol Jung, He Zhao, Joanna Dipnall, and Lan Du. Beyond unimodal: Generalising neural processes for multimodal uncertainty estimation. *Advances in Neural Information Processing Systems*, 36, 2024.

Mira Jürgens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods? *arXiv preprint arXiv:2402.09056*, 2024.

Kim-Celine Kahl, Carsten T Lüth, Maximilian Zenk, Klaus Maier-Hein, and Paul F Jaeger. Values: A framework for systematic validation of uncertainty estimation in semantic segmentation. *arXiv preprint arXiv:2401.08501*, 2024.

Dae Kang, Pamela Deyoung, Justin Tantiongloc, Todd Coleman, and Robert Owens. Statistical uncertainty quantification to augment clinical decision support: a first implementation in sleep medicine. *npj Digital Medicine*, 4, 12 2021. doi: 10.1038/s41746-021-00515-3.

Maggie Karthik and Sohier Dane. Aptos 2019 blindness detection. *Kaggle https://kaggle.com/competitions/aptos2019-blindness-detection Go to reference in*, page 5, 2019.

Mohammad Mahdi Kazemi Esfeh, Zahra Gholami, Christina Luong, Teresa Tsang, and Purang Abolmaesumi. Deue: Delta ensemble uncertainty estimation for a more robust estimation of ejection fraction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–534. Springer, 2022.

Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baeßler, Sebastian Foersch, et al. Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 13(1):7303, 2023.

Sirvan Khalighi, Kartik Reddy, Abhishek Midya, Krunal Balvantbhai Pandav, Anant Madabhushi, and Malak Abedalthagafi. Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. *NPJ precision oncology*, 8(1):80, 2024.

William A Knaus, Frank E Harrell, Joanne Lynn, Lee Goldman, Russell S Phillips, Alfred F Connors, Neal V Dawson, William J Fulkerson, Robert M Califf, Norman Desbiens, et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3):191–203, 1995.

Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

Emir Konuk, Robert Welch, Filip Christiansen, Elisabeth Epstein, and Kevin Smith. A framework for assessing joint human-ai systems based on uncertainty estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–12. Springer, 2024.

Anna-Kathrin Kopetzki, Bertrand Charpentier, Daniel Zügner, Sandhya Giri, and Stephan Günnemann. Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In *International Conference on Machine Learning*, pages 5707–5718. PMLR, 2021.

Bart Kosko and Satoru Isaka. Fuzzy logic. *Scientific American*, 269(1):76–81, 1993.

Nikita Kotelevskii, Samuel Horváth, Karthik Nandakumar, Martin Takáč, and Maxim Panov. Dirichlet-based uncertainty quantification for personalized federated learning with improved posterior networks, 2024. URL https://openreview.net/forum?id=PEoBvQWzHo.

Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11):2556–2568, 2019.

Volodymyr Kuleshov and Shachi Deshpande. Calibrated and sharp uncertainties in deep learning via density estimation. In *International Conference on Machine Learning*, pages 11683–11693. PMLR, 2022.

Alexander Kurz, Katja Hauser, Hendrik Alexander Mehrtens, Eva Krieghoff-Henning, Achim Hekler, Jakob Nikolas Kather, Stefan Fröhling, Christof von Kalle, Titus Josef Brinker, et al. Uncertainty estimation in medical image classification: systematic review. *JMIR Medical Informatics*, 10(8): e36427, 2022.

Kaisar Kushibar, Victor Campello, Lidia Garrucho, Akis Linardos, Petia Radeva, and Karim Lekadir. Layer ensembles: A single-pass uncertainty estimation in deep learning for segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 514–524. Springer, 2022.

Adrien Lafage, Mathieu Barbier, Gianni Franchi, and David Filliat. Hierarchical light transformer ensembles for multimodal trajectory forecasting. abs/2403.17678, 2024. URL https://api.semanticscholar.org/CorpusID:268691916.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Benjamin Lambert, Florence Forbes, Senan Doyle, Harmonie Dehaene, and Michel Dojat. Trustworthy clinical ai solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine*, page 102830, 2024.

Jessica Lanini, Minh Tam Davide Huynh, Gaetano Scebba, Nadine Schneider, and Raquel Rodríguez-Pérez. Unique: A framework for uncertainty quantification benchmarking. *Journal of Chemical Information and Modeling*, 64(22):8379–8386, 2024. doi: 10.1021/acs.jcim.4c01578. URL https://doi.org/10.1021/acs.jcim.4c01578. PMID: 39542432.

Agostina J. Larrazabal, César Martínez, Jose Dolz, and Enzo Ferrante. Maximum entropy on erroneous predictions: Improving model calibration for medical image segmentation. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 273–283, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43898-1.

Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

Jae-Hun Lee, Yoonho Nam, Dong-Hyun Kim, and Kanghyun Ryu. Diffusion probabilistic models-based noise reduction for enhancing the quality of medical images. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1661–1666. IEEE, 2023.

JoonHo Lee, Jae Oh Woo, Juree Seok, Parisa Hassanzadeh, Wooseok Jang, JuYoun Son, Sima Didari, Baruch Gutow, Heng Hao, Hankyu Moon, et al. Improving instruction following in language models through proxy-based uncertainty estimation. *arXiv preprint arXiv:2405.06424*, 2024.

Christian Leibig, Vaneeda Allken, M. Ayhan, Philipp Berens, and S. Wahl. Leveraging uncertainty information from deep neural networks for disease detection. 2016. doi: 10.1038/S41598-017-17876-Z.

Christian Leibig, Moritz Brehmer, Stefan Bunk, Danalyn Byng, Katja Pinker, and Lale Umutlu. Combining the strengths of radiologists and ai for breast cancer screening: a retrospective analysis. *The Lancet Digital Health*, 4(7):e507–e519, 2022.

Jonathan Lennartz and Thomas Schultz. Segmentation distortion: Quantifying segmentation uncertainty under domain shift via the effects of anomalous activations. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 316–325, Cham, 2023a. Springer Nature Switzerland. ISBN 978-3-031-43898-1.

Jonathan Lennartz and Thomas Schultz. Segmentation distortion: Quantifying segmentation uncertainty under domain shift via the effects of anomalous activations. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 316–325, Cham, 2023b. Springer Nature Switzerland. ISBN 978-3-031-43898-1.

Changbin Li, Kangshuo Li, Yuzhe Ou, Lance M Kaplan, Audun Jøsang, Jin-Hee Cho, Dong Hyun Jeong, and Feng Chen. Hyper evidential deep learning to quantify composite classification uncertainty. *arXiv preprint arXiv:2404.10980*, 2024.

Chaoyi Li, Meng Li, Can Peng, and Brian C. Lovell. Dynamic curriculum learning via in-domain uncertainty for medical image classification. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 747–757, Cham, 2023a. Springer Nature Switzerland. ISBN 978-3-031-43904-9.

Chen Li, Xiaoling Hu, Shahira Abousamra, and Chao Chen. Calibrating Uncertainty for Semi-Supervised Crowd Counting . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16685–16695, Los Alamitos, CA, USA, October 2023b. IEEE Computer Society. doi: 10.1109/ICCV51070.2023.01534. URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01534.

Kevin Li, Abhishek Gupta, Ashwin Reddy, Vitchyr H Pong, Aurick Zhou, Justin Yu, and Sergey Levine. Mural: Meta-learning uncertainty-aware rewards for outcome-driven reinforcement learning. In *International conference on machine learning*, pages 6346–6356. PMLR, 2021a.

Liu Li, Mai Xu, Xiaofei Wang, Lai Jiang, and Hanruo Liu. Attention based glaucoma detection: a large-scale database and cnn model. 2019 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10563–10572, 2019.

Rui Li, Stephanie Hu, Mingyu Lu, Yuria Utsumi, Prithwish Chakraborty, Daby M. Sow, Piyush Madan, Jun Li, Mohamed Ghalwash, Zach Shahn, and Li-wei Lehman. G-net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime. In Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer, editors, *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 282–299. PMLR, 04 Dec 2021b. URL https://proceedings.mlr.press/v158/li21a.html.

Xiangyu Li, Xinjie Liang, Gongning Luo, Wei Wang, Kuanquan Wang, and Shuo Li. Ultra: Uncertainty-aware label distribution learning for breast tumor cellularity assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 303–312. Springer, 2022a.

Yi Li, Yiqun Duan, Zhanghui Kuang, Yimin Chen, Wayne Zhang, and Xiaomeng Li. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. 36(2):1447–1455, 2022b. doi: 10.1609/aaai.v36i2.20034. URL https://ojs.aaai.org/index.php/AAAI/article/view/20034.

Yikuan Li, Shishir Rao, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Gholamreza Salimi-Khorshidi, Mohammad Mamouei, Thomas Lukasiewicz, and Kazem Rahimi. Deep bayesian gaussian processes for uncertainty estimation in electronic health records. 11(1):20685, 2021c. doi: 10.1038/s41598-021-00144-6. URL https://www.nature.com/articles/s41598-021-00144-6.

Christie Lin, Tyler Bradshaw, Timothy Perk, Stephanie Harmon, Jens Eickhoff, Ngoneh Jallow, Peter L Choyke, William L Dahut, Steven Larson, John Laurence Humm, et al. Repeatability of quantitative 18f-naf pet: a multicenter study. *Journal of Nuclear Medicine*, 57(12):1872–1879, 2016.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018a.

Guiliang Liu, Yudong Luo, Oliver Schulte, and Pascal Poupart. Uncertainty-aware reinforcement learning for risk-sensitive player evaluation in sports game. *Advances in Neural Information Processing Systems*, 35:20218–20231, 2022.

Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. When gaussian process meets big data: A review of scalable gps. *arXiv preprint arXiv:1807.01065*, 2018b. URL https://arxiv.org/abs/1807.01065.

Kangning Liu, Brian L Price, Jason Kuen, Yifei Fan, Zijun Wei, Luis Figueroa, Krzysztof J Geras, and Carlos Fernandez-Granda. Uncertainty-aware finetuning of segmentation foundation models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Pei Liu and Luping Ji. Weakly-supervised residual evidential learning for multi-instance uncertainty estimation. *arXiv preprint arXiv:2405.04405*, 2024.

Tyler J. Loftus, Benjamin Shickel, Matthew M. Ruppert, Jeremy A. Balch, Tezcan Ozrazgat-Baslanti, Patrick J. Tighe, Philip A. Efron, William R. Hogan, Parisa Rashidi, Gilbert R. Upchurch, and

Azra Bihorac. Uncertainty-aware deep learning in healthcare: A scoping review. 1(8):e0000085, 2022. doi: 10.1371/journal.pdig.0000085. URL https://dx.plos.org/10.1371/journal.pdig.0000085.

Imperial College London. Ixi dataset, 2008. URL https://brain-development.org/ixi-dataset/.

Charles Lu, Andreanne Lemay, Ken Chang, Katharina Hoebel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. *arXiv preprint arXiv:2109.04392*, 2021. URL https://arxiv.org/abs/2109.04392.

Wenjing Lu, Jiahao Lei, Peng Qiu, Rui Sheng, Jinhua Zhou, Xinwu Lu, and Yang Yang. Upcol: uncertainty-informed prototype consistency learning for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 662–672. Springer, 2023.

Yingzhou Lu, Tianyi Chen, Nan Hao, Capucine Van Rechem, Jintai Chen, and Tianfan Fu. Uncertainty quantification and interpretability for clinical trial approval prediction. 4:0126, 2024. doi: 10.34133/hds.0126. URL https://spj.science.org/doi/10.34133/hds.0126.

Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 318–329. Springer, 2021.

Chenxi Ma. Uncertainty-aware GAN for single image super resolution. 38(5):4071–4079, 2024. doi: 10.1609/aaai.v38i5.28201. URL https://ojs.aaai.org/index.php/AAAI/article/view/28201.

Samual MacDonald, Helena Foley, Melvyn Yap, Rebecca L Johnston, Kaiah Steven, Lambros T Koufariotis, Sowmya Sharma, Scott Wood, Venkateswar Addala, John V Pearson, et al. Generalising uncertainty improves accuracy and safety of deep learning analytics applied to oncology. *Scientific Reports*, 13(1):7395, 2023.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

Hila Manor and Tomer Michaeli. On the posterior distribution in denoising: Application to uncertainty quantification. *arXiv preprint arXiv:2309.13598*, 2023.

Yamin Mao, Zhihua Liu, Weiming Li, Yuchao Dai, Qiang Wang, Yun-Tae Kim, and Hong-Seok Lee. Uasnet: Uncertainty adaptive sampling network for deep stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6311–6319, 2021a.

Yamin Mao, Zhihua Liu, Weiming Li, Yuchao Dai, Qiang Wang, Yun-Tae Kim, and Hong-Seok Lee. Uasnet: Uncertainty adaptive sampling network for deep stereo matching. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6291–6299, 2021b. doi: 10.1109/ICCV48922.2021.00625.

Taylor R Moen, Baiyu Chen, David R Holmes III, Xinhui Duan, Zhicong Yu, Lifeng Yu, Shuai Leng, Joel G Fletcher, and Cynthia H McCollough. Low-dose ct image and projection dataset. *Medical physics*, 48(2):902–911, 2021.

Pablo Morales-Álvarez, Daniel Hernández-Lobato, Rafael Molina, and José Miguel Hernández-Lobato. ACTIVATION-LEVEL UNCERTAINTY IN DEEP NEURAL NETWORKS. 2021.

Daniel Cardoso Moura, Miguel Angel Guevara López, Pedro Cunha, Naimy González de Posada, Raúl Ramos Pollan, Isabel Ramos, Joana Pinheiro Loureiro, Inês C Moreira, Bruno M Ferreira de Araújo, and Teresa Cardoso Fernandes. Benchmarking datasets for breast cancer computer-aided diagnosis (cadx). In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part I 18*, pages 326–333. Springer, 2013.

Elias Nehme, Omer Yair, and Tomer Michaeli. Uncertainty quantification via neural posterior principal components. *Advances in Neural Information Processing Systems*, 36:37128–37141, 2023.

Dinh C Nguyen, Quoc-Viet Pham, Pubudu N Pathirana, Ming Ding, Aruna Seneviratne, Zihuai Lin, Octavia Dobre, and Won-Joo Hwang. Federated learning for smart healthcare: A survey. *ACM Computing Surveys (Csur)*, 55(3):1–37, 2022.

Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2020.

Thanh Nguyen, Abbas Khosravi, Douglas Creighton, and Saeid Nahavandi. Classification of healthcare data using genetic fuzzy logic system and wavelets. *Expert Systems with Applications*, 42(4): 2184–2197, 2015.

Msoud Nickparvar. Brain tumor mri dataset, 2021. URL https://www.kaggle.com/dsv/2645886.

Frank Nussbaum, Jakob Gawlikowski, and Julia Niebling. Structuring uncertainty for fine-grained sampling in stochastic segmentation networks. *Advances in Neural Information Processing Systems*, 35:27678–27691, 2022.

Philipp Oberdiek, Gernot Fink, and Matthias Rottmann. Uqgan: A unified model for uncertainty quantification of deep classifiers trained via conditional gans. *Advances in Neural Information Processing Systems*, 35:21371–21385, 2022.

Seok-Hwan Oh, Guil Jung, Sang-Yun Kim, Myeong-Gee Kim, Young-Min Kim, Hyeon-Jik Lee, Hyuk-Sool Kwon, and Hyeon-Min Bae. Uncertainty-aware meta-weighted optimization framework for domain-generalized medical image segmentation. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 775–785, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72083-3.

Jaya Ojha, Oriana Presacan, Pedro Goncalves Lind, Eric Monteiro, and Anis Yazidi. Navigating uncertainty: A user-perspective survey of trustworthiness of ai in healthcare. *ACM Trans. Comput. Healthcare*, February 2025. doi: 10.1145/3716317. URL https://doi.org/10.1145/3716317. Just Accepted.

Valentin Oreiller, Vincent Andrearczyk, Mario Jreige, Sarah Boughdad, Hesham Elhalawani, Joel Castelli, Martin Vallières, Simeng Zhu, Juanying Xie, Ying Peng, et al. Head and neck tumor segmentation in pet/ct: the hecktor challenge. *Medical image analysis*, 77:102336, 2022.

José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020.

Clemens Oszkinat, Susan E Luczak, and IG Rosen. Uncertainty quantification in estimating blood alcohol concentration from transdermal alcohol level with physics-informed neural networks. *IEEE transactions on neural networks and learning systems*, 34(10):8094–8101, 2023.

David Ouyang, Bryan He, Amirata Ghorbani, Curt P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, and James Y Zou. Interpretable ai for beat-to-beat cardiac function assessment. *medRxiv*, page 19012419, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Andre G. C. Pacheco, Chandramouli S. Sastry, Thomas Trappenberg, Sageev Oore, and Renato A. Krohling. On out-of-distribution detection algorithms with deep neural skin cancer classifiers. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3152–3161, 2020a. doi: 10.1109/CVPRW50498.2020.00374.

Andre G.C. Pacheco, Gustavo R. Lima, Amanda S. Salomão, Breno Krohling, Igor P. Biral, Gabriel G. de Angelo, Fábio C.R. Alves Jr, José G.M. Esgario, Alana C. Simora, Pedro B.C. Castro, Felipe B. Rodrigues, Patricia H.L. Frasson, Renato A. Krohling, Helder Knidel, Maria C.S. Santos, Rachel B. do Espírito Santo, Telma L.S.G. Macedo, Tania R.P. Canuto, and Luíz F.S. de Barros. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:106221, 2020b. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2020.106221. URL https://www.sciencedirect.com/science/article/pii/S235234092031115X.

Andre GC Pacheco, Chandramouli S Sastry, Thomas Trappenberg, Sageev Oore, and Renato A Krohling. On out-of-distribution detection algorithms with deep neural skin cancer classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 732–733, 2020c.

Chengwei Pan, Gangming Zhao, Junjie Fang, Baolian Qi, Jiaheng Liu, Chaowei Fang, Dingwen Zhang, Jinpeng Li, and Yizhou Yu. Computer-aided tuberculosis diagnosis with attribute reasoning assistance. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 623–633. Springer, 2022.

Aris T Papageorghiou, Stephen H Kennedy, Laurent J Salomon, Douglas G Altman, Eric O Ohuma, William Stones, Michael G Gravett, Fernando C Barros, Cesar Victora, Manorama Purwar, et al. The intergrowth-21st fetal growth standards: toward the global integration of pregnancy and pediatric care. *American journal of obstetrics and gynecology*, 218(2):S630–S640, 2018.

Christina Papangelou, Konstantinos Kyriakidis, Pantelis Natsiavas, Ioanna Chouvarda, and Andigoni Malousi. Reliable machine learning models in genomic medicine using conformal prediction. *medRxiv*, pages 2024–09, 2024.

Jiahuan Pei, Cheng Wang, and György Szarvas. Transformer uncertainty estimation with hierarchical stochastic attention. 36(10):11147–11155, 2022. doi: 10.1609/aaai.v36i10.21364. URL https://ojs.aaai.org/index.php/AAAI/article/view/21364.

Alina Peluso, Ioana Danciu, Hong-Jun Yoon, Jamaludin Mohd Yusof, Tanmoy Bhattacharya, Adam Spannaus, Noah Schaefferkoetter, Eric B. Durbin, Xiao-Cheng Wu, Antoinette Stroup, Jennifer Doherty, Stephen Schwartz, Charles Wiggins, Linda Coyle, Lynne Penberthy, Georgia D. Tourassi, and Shang Gao. Deep learning uncertainty quantification for clinical text classification. 149:104576, 2024. doi: 10.1016/j.jbi.2023.104576.

Weiwen Peng, Zhi-Sheng Ye, and Nan Chen. Bayesian deep-learning-based health prognostics toward prognostics uncertainty. 67(3):2283–2293, 2020. doi: 10.1109/TIE.2019.2907440. URL https://ieeexplore.ieee.org/abstract/document/8681720.

Nicholas Petrick, Shazia Akbar, Kenny H Cha, Sharon Nofech-Mozes, Berkman Sahiner, Marios A Gavrielides, Jayashree Kalpathy-Cramer, Karen Drukker, Anne L Martel, and for the BreastPathQ Challenge Group. Spie-aapm-nci breastpathq challenge: an image analysis challenge for quantitative tumor cellularity assessment in breast cancer histology images following neoadjuvant treatment. *Journal of Medical Imaging*, 8(3):034501–034501, 2021.

Henry Pinkard, Cherry Liu, Fanice Nyatigo, Daniel A Fletcher, and Laura Waller. The berkeley single cell computational microscopy (bsccm) dataset. *arXiv preprint arXiv:2402.06191*, 2024.

Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In *International Conference on Machine Learning*, pages 27907–27947. PMLR, 2023.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Teodora Popordanoska, Jeroen Bertels, Dirk Vandermeulen, Frederik Maes, and Matthew B Blaschko. On the relationship between calibrated predictors and unbiased volume estimation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 678–688. Springer, 2021.

Chetanya Puri, Gerben Kooijman, Bart Vanrumste, and Stijn Luca. Forecasting time series in healthcare with gaussian processes and dynamic time warping based subset selection. *IEEE Journal of*

Biomedical and Health Informatics, 26(12):6126–6137, 2022.

Lorena Qendro, Alexander Campbell, Pietro Lio, and Cecilia Mascolo. Early exit ensembles for uncertainty quantification. In Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer, editors, *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 181–195. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/qendro21a.html.

Chao Qu, Wenxin Liu, and Camillo J Taylor. Bayesian deep basis fitting for depth completion with uncertainty. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16147–16157, 2021.

Jayroop Ramesh, Nicola Dinsdale, Pak-Hei Yeung, and Ana I. L. Namburete. Geometric transformation uncertainty for improving 3d fetal brain pose prediction from freehand 2d ultrasound videos. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 419–429, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72378-0.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pages 108–109. IEEE, 2012.

Kai Ren, Ke Zou, Xianjie Liu, Yidi Chen, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu. Uncertainty-informed mutual learning for joint medical image classification and segmentation, 2023. URL https://arxiv.org/abs/2303.10049.

Zhihang Ren, Yunqi Li, Xinyu Li, Xinrong Xie, Erik P. Duhaime, Kathy Fang, Tapabrata Chakraborti, Yunhui Guo, Stella X. Yu, and David Whitney. Skincon: Towards consensus for the uncertainty of skin cancer sub-typing through distribution regularized adaptive predictive sets (draps). In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 405–415, Cham, 2024a. Springer Nature Switzerland. ISBN 978-3-031-72378-0.

Zhihang Ren, Yunqi Li, Xinyu Li, Xinrong Xie, Erik P. Duhaime, Kathy Fang, Tapabrata Chakraborty, Yunhui Guo, Stella X. Yu, and David Whitney. SkinCON: Towards consensus for the uncertainty of skin cancer sub-typing through distribution regularized adaptive predictive sets (DRAPS) . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15001. Springer Nature Switzerland, October 2024b.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=-ODN6SbiUU.

Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1):1–39, 2010.

Holger R. Roth, Amal Farag, Le Lu, Evrim B. Turkbey, and Ronald M. Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9349 of *Lecture Notes in Computer Science*, pages 556–564. Springer, 2015. doi: 10.1007/978-3-319-24553-9_68. URL https://doi.org/10.1007/978-3-319-24553-9_68.

Patrick Royston and Douglas G Altman. External validation of a cox prognostic model: principles and methods. *BMC medical research methodology*, 13:1–15, 2013.

Tim GJ Rudner, Zonghao Chen, Yee Whye Teh, and Yarin Gal. Tractable function-space variational inference in bayesian neural networks. *Advances in Neural Information Processing Systems*, 35:22686–22698, 2022.

Anindo Saha, Joeran Bosma, Jasper Twilt, Bram van Ginneken, Derya Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen F"utterer, Maarten de Rooij, et al. Artificial intelligence and radiologists at prostate cancer detection in mri—the pi-cai challenge. In *Medical Imaging with Deep Learning, short paper track*, 2023.

Jaakko Sahlsten, Joel Jaskari, Kareem A. Wahid, Sara Ahmed, Enrico Glerean, Renjie He, Benjamin H. Kann, Antti Mäkitie, Clifton D. Fuller, Mohamed A. Naser, and Kimmo Kaski. Application of simultaneous uncertainty quantification and segmentation for oropharyngeal cancer use-case with bayesian deep learning. 4(1):110, 2024. doi: 10.1038/s43856-024-00528-5. URL https://www.nature.com/articles/s43856-024-00528-5.

EB Saldich, C Wang, IG Rosen, L Goldstein, J Bartroff, RM Swift, and SE Luczak. Obtaining high-resolution multi-biosensor data for modeling transdermal alcohol concentration data. In *ALCOHOLISM-CLINICAL AND EXPERIMENTAL RESEARCH*, volume 44, pages 181–181. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2020.

Aven Samareh and Shuai Huang. UQ-CHI: An uncertainty quantification-based contemporaneous health index for degenerative disease monitoring. 2019.

Elisa Scalco, Silvia Pozzi, Giovanna Rizzo, and Ettore Lanzarone. Uncertainty quantification in multi-class segmentation: Comparison between bayesian and non-bayesian approaches in a clinical perspective. 2024. doi: 10.1002/mp.17189.

Jelena Schmidt, Nienke M Schutte, Stefan Buttigieg, David Novillo-Ortiz, Eric Sutherland, Michael Anderson, Bart de Witte, Michael Peolsson, Brigid Unim, Milena Pavlova, et al. Mapping the regulatory landscape for artificial intelligence in health within the european union. *npj Digital Medicine*, 7(1):229, 2024.

Dominik Schnaus, Jongseok Lee, Daniel Cremers, and Rudolph Triebel. Learning expressive priors for generalization and uncertainty estimation in neural networks. In *International Conference on Machine Learning*, pages 30252–30284. PMLR, 2023.

Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd M"unzer, Manfred J"urgen Primus, and Doris Putzgruber. Cataract-101: video dataset of 101 cataract surgeries. In *Proceedings of the 9th ACM multimedia systems conference*, pages 421–425, 2018.

Brayden Schott, Dmitry Pinchuk, Victor Santoro-Fernandes, Žan Klaneček, Luciano Rivetti, Alison Deatsch, Scott Perlman, Yixuan Li, and Robert Jeraj. Uncertainty quantification via localized gradients for deep learning-based medical image assessments. 69(15), 2024. doi: 10.1088/1361-6560/ad611d.

Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, Günter Klambauer, and Sepp Hochreiter. Quantification of uncertainty with adversarial models. *Advances in Neural Information Processing Systems*, 36:19446–19484, 2023.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, pages 3179–3189, 2018. URL https://arxiv.org/abs/1806.01768.

Silvia Seoni, Vicnesh Jahmunah, Massimo Salvi, Prabal Datta Barua, Filippo Molinari, and U. Rajendra Acharya. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). *Computers in Biology and Medicine*, 165:107441, 2023a. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2023.107441. URL https://www.sciencedirect.com/science/article/pii/S001048252300906X.

Silvia Seoni, Vicnesh Jahmunah, Massimo Salvi, Prabal Datta Barua, Filippo Molinari, and U. Rajendra Acharya. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013-2023). 165:107441, 2023b. doi: 10.1016/j.compbiomed.2023.107441.

Uday Shankar Shanthamallu, Jayaraman J. Thiagarajan, and Andreas Spanias. Uncertainty-matching graph neural networks to defend against poisoning attacks. 35(11):9524–9532, 2021. doi: 10.1609/aaai.v35i11.17147. URL https://ojs.aaai.org/index.php/AAAI/article/view/17147.

Saurabh Sharma, Atul Kumar, and Joydeep Chandra. Confidence matters: Enhancing medical image classification through uncertainty-driven contrastive self-distillation. In *Medical Image Computing and Computer Assisted Intervention – MIC-*

CAI 2024: 27th International Conference, Marrakesh, Morocco, October 6–10, 2024, Proceedings, Part X, page 133–142, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72116-8. doi: 10.1007/978-3-031-72117-5_13. URL https://doi.org/10.1007/978-3-031-72117-5_13.

Maohao Shen, Yuheng Bu, Prasanna Sattigeri, Soumya Ghosh, Subhro Das, and Gregory Wornell. Post-hoc uncertainty learning using a dirichlet meta-model. 37(8):9772–9781, 2023. doi: 10.1609/aaai.v37i8.26167. URL https://ojs.aaai.org/index.php/AAAI/article/view/26167.

Ruohua Shi, Lingyu Duan, Tiejun Huang, and Tingting Jiang. Evidential uncertainty-guided mitochondria segmentation for 3d EM images. 38(5):4847–4855, 2024. doi: 10.1609/aaai.v38i5.28287. URL https://ojs.aaai.org/index.php/AAAI/article/view/28287.

George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1): e180041, 2019.

Smadar Shilo, Hagai Rossman, and Eran Segal. Applications of machine learning in healthcare. *Journal of the American Medical Association*, 323(9): 891–892, 2020. URL https://jamanetwork.com/journals/jama/article-abstract/2761828.

Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Kenichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American journal of roentgenology*, 174(1):71–74, 2000.

Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9:283–293, 2014.

Victor Wåhlstrand Skärström, Lisa Johansson, Jennifer Alvén, Mattias Lorentzon, and Ida Häggström. Explainable vertebral fracture analysis with uncertainty estimation using differentiable rule-based classification, 2024. URL https://arxiv.org/abs/2407.02926.

Roberto Souza, Oeslle Lucena, Julia Garrafa, David Gobbi, Marina Saluzzi, Simone Appenzeller, Letícia Rittner, Richard Frayne, and Roberto Lotufo. An open, multi-vendor, multi-field-strength brain mr dataset and analysis of publicly available skull stripping methods agreement. *NeuroImage*, 170:482–494, 2018.

Alex C. Stutts, Danilo Erricolo, Sathya Ravi, Theja Tulabandhula, and Amit Ranjan Trivedi. Mutual information-calibrated conformal feature fusion for uncertainty-aware multimodal 3d object detection at the edge, 2023. URL https://arxiv.org/abs/2309.09593.

Pegah Tabarisaadi, Abbas Khosravi, Saeid Nahavandi, Miadreza Shafie-Khah, and João P. S. Catalão. An optimized uncertainty-aware training framework for neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5): 6928–6935, 2024. doi: 10.1109/TNNLS.2022.3213315.

Ryutaro Tanno, Daniel E. Worrall, Enrico Kaden, Aurobrata Ghosh, Francesco Grussu, A. Bizzi, S. Sotiropoulos, A. Criminisi, and D. Alexander. Uncertainty quantification in deep learning for safer neuroimage enhancement. 2019.

Marvin Tom Teichmann, Manasi Datar, Lisa Kratzke, Fernando Vega, and Florin C Ghesu. Towards integrating epistemic uncertainty estimation into the radiotherapy workflow. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2024.

Ponkrshnan Thiagarajan, Pushkar Khairnar, and Susanta Ghosh. Explanation and use of uncertainty quantified by bayesian neural network classifiers for breast histopathology images. 2022. doi: 10.1109/TMI.2021.3123300.

Jordan Thompson, Ronald Koe, Anthony Le, Gabriella Goodman, Daniel S Brown, and Alan Kuntz. Early failure detection in autonomous surgical soft-tissue manipulation via uncertainty quantification. *arXiv preprint arXiv:2501.10561*, 2025.

Giovanni Tripepi, Kitty J Jager, Friedo W Dekker, and Carmine Zoccali. Selection bias and information bias in clinical research. *Nephron Clinical Practice*, 115(2):c94–c99, 2010.

Puja Trivedi, Mark Heimann, Rushil Anirudh, Danai Koutra, and Jayaraman J Thiagarajan. Accurate and scalable estimation of epistemic uncertainty for graph neural networks. *arXiv preprint arXiv:2401.03350*, 2024.

Krasimira Tsaneva-Atanasova, Giulia Pederzanil, and Marianna Laviola. Decoding uncertainty for clinical decision-making. *Philosophical Transactions A*, 383(2292):20240207, 2025.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Malte Tölle, Fernando Navarro, Sebastian Eble, Ivo Wolf, Bjoern Menze, and Sandy Engelhardt. Funavg: Federated uncertainty weighted averaging for datasets with diverse labels, 2024. URL https://arxiv.org/abs/2407.07488.

Uddeshya Upadhyay, Yanbei Chen, Tobias Hepp, Sergios Gatidis, and Zeynep Akata. Uncertainty-guided progressive gans for medical image translation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 614–624. Springer, 2021.

Matias Valdenegro-Toro, Ivo Pascal de Jong, and Marco Zullich. Unified uncertainties: Combining input, data and model uncertainty into a single formulation. *arXiv preprint arXiv:2406.18787*, 2024.

Menno Valk, Ryutaro Tanno, Jo Schlemper, Erik Dam, Max Welling, Yee Whye Teh, and Sebastien Ourselin. Uncertainty quantification in deep learning for medical imaging: A survey. *arXiv preprint arXiv:2106.00473*, 2021. URL https://arxiv.org/abs/2106.00473.

Thomas LA van den Heuvel, Dagmar de Bruijn, Chris L de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2d ultrasound images. *PloS one*, 13(8): e0200412, 2018.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Hooman Vaseli, Ang Nan Gu, S Neda Ahmadi Amiri, Michael Y Tsang, Andrea Fung, Nima Kondori, Armin Saadat, Purang Abolmaesumi, and Teresa SM Tsang. Protoasnet: Dynamic prototypes for inherently interpretable and uncertainty-aware aortic stenosis classification in echocardiography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 368–378. Springer, 2023.

Amol A Verma, Joshua Murray, Russell Greiner, Joseph Paul Cohen, Kaveh G Shojania, Marzyeh Ghassemi, Sharon E Straus, Chloe Pou-Prom, and Muhammad Mamdani. Implementing machine learning in medicine. *Cmaj*, 193(34):E1351–E1357, 2021.

TL Vollmer, PS Sorensen, K Selmaj, F Zipp, E Havrdova, JA Cohen, N Sasson, Y Gilgun-Sherki, DL Arnold, and BRAVO Study Group. A randomized placebo-controlled phase iii trial of oral laquinimod for multiple sclerosis. *Journal of neurology*, 261:773–783, 2014.

Jeroen F Vranken, Rutger R van de Leur, Deepak K Gupta, Luis E Juarez Orozco, Rutger J Hassink, Pim van der Harst, Pieter A Doevendans, Sadaf Gulshad, and René van Es. Uncertainty estimation for deep learning-based automated analysis of 12-lead electrocardiograms. 2(3):401–415, 2021. doi: 10.1093/ehjdh/ztab045. URL https://doi.org/10.1093/ehjdh/ztab045.

Neeraj Wagh, Jionghao Wei, Samarth Rawal, Brent M Berry, and Yogatheesan Varatharajah. Evaluating latent space robustness and uncertainty of eeg-ml models under realistic distribution shifts. *Advances in Neural Information Processing Systems*, 35:21142–21156, 2022.

Haoxuan Wang, Zhiding Yu, Yisong Yue, Animashree Anandkumar, Anqi Liu, and Junchi Yan. Learning calibrated uncertainties for domain shift: a distributionally robust learning approach. In *Proceedings of the Thirty-Second International Joint Con-*

ference on Artificial Intelligence, IJCAI '23, 2023. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/162. URL https://doi.org/10.24963/ijcai.2023/162.

Hu Wang, Guansong Pang, Chunhua Shen, and Congbo Ma. Unsupervised representation learning by predicting random distances. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2950–2956. ijcai.org, 2020a.

Hu Wang, Jianpeng Zhang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Uncertainty-aware multi-modal learning via cross-modal random network prediction. In *European Conference on Computer Vision*, pages 200–217. Springer, 2022a.

Jiaqi Wang, Chenxu Zhao, Lingjuan Lyu, Quanzeng You, Mengdi Huai, and Fenglong Ma. Bridging model heterogeneity in federated learning via uncertainty-based asymmetrical reciprocity learning. *arXiv preprint arXiv:2407.03247*, 2024.

Wenkong Wang, Weijie Huang, Quanli Lu, Jiyang Chen, Menghua Zhang, Jia Qiao, and Yong Zhang. Attention mechanism-based deep learning method for hairline fracture detection in hand x-rays. *Neural Computing and Applications*, 34(21):18773–18785, 2022b.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Yezhen Wang, Bo Li, Tong Che, Kaiyang Zhou, Ziwei Liu, and Dongsheng Li. Energy-based open-world uncertainty modeling for confidence calibration. *CoRR*, abs/2107.12628, 2021. URL https://arxiv.org/abs/2107.12628.

Yuexi Wang and Veronika Rockova. Uncertainty quantification for sparse deep learning. pages 298–308. PMLR, 2020. URL https://proceedings.mlr.press/v108/wang20b.html.

Zhiqing Wang, Zihang Dai, Zhenwen Hu, and Mark JF Gales. Bayesian deep learning: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020b. URL https://arxiv.org/abs/2011.08588.

Frederik Warburg, Martin Jørgensen, Javier Civera, and Søren Hauberg. Bayesian triplet loss: Uncertainty quantification in image retrieval. *CoRR*, abs/2011.12663, 2020. URL https://arxiv.org/abs/2011.12663.

Donglai Wei, Zudi Lin, Daniel Franco-Barranco, Nils Wendt, Xingyu Liu, Wenjie Yin, Xin Huang, Aarush Gupta, Won-Dong Jang, Xueying Wang, et al. Mitoem dataset: Large-scale 3d mitochondria instance segmentation from em images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 66–76. Springer, 2020.

Philipp Werner, Ayoub Al-Hamadi, and Steffen Walter. Analysis of facial expressiveness during experimentally induced heat pain. In *2017 Seventh international conference on affective computing and intelligent interaction workshops and demos (ACIIW)*, pages 176–180. IEEE, 2017.

Veit David Wild, Sahra Ghalebikesabi, Dino Sejdinovic, and Jeremias Knoblauch. A rigorous link between deep ensembles and (variational) bayesian methods. *arXiv preprint arXiv:2305.15027*, 2023. URL https://arxiv.org/abs/2305.15027.

Ashia C. Wilson and Pavel Izmailov. Bayesian deep ensembles via the neural tangent kernel. *arXiv preprint arXiv:2007.05864*, 2020. URL https://arxiv.org/abs/2007.05864.

Stefan Winzeck, Arsany Hakim, Richard McKinley, José AADSR Pinto, Victor Alves, Carlos Silva, Maxim Pisov, Egor Krivov, Mikhail Belyaev, Miguel Monteiro, et al. Isles 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral mri. *Frontiers in neurology*, 9:679, 2018.

Street Nick Wolberg William, Mangasarian Olvi and Street W. Breast cancer wisconsin (diagnostic). UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5DW2B.

Tom Wollschläger, Nicholas Gao, Bertrand Charpentier, Mohamed Amine Ketata, and Stephan Günnemann. Uncertainty estimation for molecules: Desiderata and methods. In *International conference on machine learning*, pages 37133–37156. PMLR, 2023.

Jianghao Wu, Guotai Wang, Ran Gu, Tao Lu, Yinan Chen, Wentao Zhu, Tom Vercauteren, Sébastien Ourselin, and Shaoting Zhang. Upl-sfda: Uncertainty-aware pseudo label guided source-free domain adaptation for medical image segmentation. *IEEE transactions on medical imaging*, 2023a.

Jiawei Wu, Changqing Zhang, Zuoyong Li, Huazhu Fu, Xi Peng, and Joey Tianyi Zhou. dugmatting: decomposed-uncertainty-guided matting. *arXiv preprint arXiv:2306.01452*, 2023b.

Junyang Wu, Rong Tao, and Guoyan Zheng. Nonlinear regression of remaining surgical duration via bayesian lstm-based deep negative correlation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 421–430. Springer, 2022.

Yue Wu, Shuangfei Zhai, Nitish Srivastava, Joshua Susskind, Jian Zhang, Ruslan Salakhutdinov, and Hanlin Goh. Uncertainty weighted actor-critic for offline reinforcement learning. *arXiv preprint arXiv:2105.08140*, 2021.

Tong Xia, Jing Han, and Cecilia Mascolo. *Benchmarking Uncertainty Quantification on Biosignal Classification Tasks Under Dataset Shift*, pages 347–359. Springer International Publishing, Cham, 2023. ISBN 978-3-031-14771-5. doi: 10.1007/978-3-031-14771-5_25. URL https://doi.org/10.1007/978-3-031-14771-5_25.

Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama. Sample selection with uncertainty of losses for learning with noisy labels. *arXiv preprint arXiv:2106.00445*, 2021.

Jinyi Xiang, Peng Qiu, and Yang Yang. Fussnet: Fusing two sources of uncertainty for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 481–491. Springer, 2022a.

Jinyi Xiang, Peng Qiu, and Yang Yang. FUSS-Net: Fusing two sources of uncertainty for semi-supervised medical image segmentation. pages 481–491. Springer Nature Switzerland, 2022b. doi: 10.1007/978-3-031-16452-1_46.

Annie Xie, Shagun Sodhani, Chelsea Finn, Joelle Pineau, and Amy Zhang. Robust policy learning over multiple uncertainty sets. In *International Conference on Machine Learning*, pages 24414–24429. PMLR, 2022a.

Yanting Xie, Hongen Liao, Daoqiang Zhang, and Fang Chen. Uncertainty-aware cascade network for ultrasound image segmentation with ambiguous boundary. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 268–278. Springer, 2022b.

Yanting Xie, Hongen Liao, Daoqiang Zhang, and Fang Chen. Uncertainty-aware cascade network for ultrasound image segmentation with ambiguous boundary. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 268–278. Springer, 2022c.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical image analysis*, 67:101832, 2021.

Pengxiang Yan, Ziyi Wu, Mengmeng Liu, Kun Zeng, Liang Lin, and Guanbin Li. Unsupervised domain adaptive salient object detection through uncertainty-aware pseudo-label learning. 36(3): 3000–3008, 2022. doi: 10.1609/aaai.v36i3.20206. URL https://ojs.aaai.org/index.php/AAAI/article/view/20206.

Chen Yang, Xiaoqing Guo, Zhen Chen, and Yixuan Yuan. Source free domain adaptation for medical image segmentation with fourier style mining. *Medical Image Analysis*, 79:102457, 2022a.

Han Yang, Lu Shen, Mengke Zhang, and Qiuli Wang. Uncertainty-guided lung nodule segmentation with feature-aware attention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 44–54. Springer, 2022b.

Hongzheng Yang, Cheng Chen, Yueyao Chen, Hon Chi Yip, and DOU QI. Uncertainty estimation for safety-critical scene segmentation via fine-grained reward maximization. *Advances in Neural Information Processing Systems*, 36:36238–36249, 2023a.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023b.

Lin Yang, Junlong Lyu, Wenlong Lyu, and Zhitang Chen. Efficient robust bayesian optimization for arbitrary uncertain inputs. *Advances in Neural Information Processing Systems*, 36, 2024.

Yuling Yao, Aki Vehtari, and Andrew Gelman. Stacking for non-mixing bayesian computations: The curse and blessing of multimodal posteriors. *J. Mach. Learn. Res.*, 23:79:1–79:45, 2020. URL https://api.semanticscholar.org/CorpusID:219966748.

Linlin Yu, Yifei Lou, and Feng Chen. Uncertainty-aware graph-based hyperspectral image classification. In *The Twelfth International Conference on Learning Representations*, 2023.

Han Yuan. Toward real-world deployment of machine learning for health care: external validation, continual monitoring, and randomized clinical trials. *Health Care Science*, 3(5):360, 2024.

Lotfi Asker Zadeh. Fuzzy logic. *Computer*, 21(4): 83–93, 1988.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

Kilian Zepf, Selma Wanna, Marco Miani, Juston Moore, Jes Frellsen, Søren Hauberg, Frederik Warburg, and Aasa Feragen. Laplacian segmentation networks improve epistemic uncertainty quantification, 2024. URL https://arxiv.org/abs/2303.13123.

Angela Zhang, Lei Xing, James Zou, and Joseph C Wu. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature biomedical engineering*, 6(12):1330–1345, 2022a.

Dawei Zhang, Yanwei Fu, and Zhonglong Zheng. Uast: uncertainty-aware siamese tracking. In *International Conference on Machine Learning*, pages 26161–26175. PMLR, 2022b.

Junkai Zhang, Weitong Zhang, Dongruo Zhou, and Quanquan Gu. Uncertainty-aware reward-free exploration with general function approximation. *arXiv preprint arXiv:2406.16255*, 2024a.

Quan Zhang, Jian-Huang Lai, Zhanxiang Feng, and Xiaohua Xie. Uncertainty modeling with second-order transformer for group re-identification. 36(3): 3318–3325, 2022c. doi: 10.1609/aaai.v36i3.20241. URL https://ojs.aaai.org/index.php/AAAI/article/view/20241.

Wang Zhang, Ziwen Martin Ma, Subhro Das, Tsui-Wei Lily Weng, Alexandre Megretski, Luca Daniel, and Lam M. Nguyen. One step closer to unbiased aleatoric uncertainty estimation. 38(15):16857–16864, 2024b. doi: 10.1609/aaai.v38i15.29627. URL https://ojs.aaai.org/index.php/AAAI/article/view/29627.

Xiaoran Zhang, Daniel H. Pak, Shawn S. Ahn, Xiaoxiao Li, Chenyu You, Lawrence H. Staib, Albert J. Sinusas, Alex Wong, and James S. Duncan. Heteroscedastic Uncertainty Estimation Framework for Unsupervised Registration . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15002. Springer Nature Switzerland, October 2024c.

Xinyu Zhang, Zhiwei Li, Zhenhong Zou, Xin Gao, Yijin Xiong, Dafeng Jin, Jun Li, and Huaping Liu. Informative data selection with uncertainty for multimodal object detection. pages 1–13, 2023a. doi: 10.1109/TNNLS.2023.3270159. URL https://ieeexplore.ieee.org/document/10132437/.

Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34:17616–17627, 2021.

Yuwei Zhang, Tong Xia, Abhirup Ghosh, and Cecilia Mascolo. Uncertainty quantification in federated learning for heterogeneous health data. In *International Workshop on Federated Learning for Distributed Data Mining*, 2023b.

Yidong Zhao, Changchun Yang, Artur Schweidtmann, and Qian Tao. Efficient bayesian uncertainty estimation for nnu-net. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 535–544. Springer, 2022.

Yidong Zhao, Yi Zhang, Orlando Simonetti, Yuchi Han, and Qian Tao. Lost in tracking: Uncertainty-guided cardiac cine mri segmentation at right ventricle base. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 415–424. Springer, 2024a.

Yidong Zhao, Yi Zhang, Orlando Simonetti, Yuchi Han, and Qian Tao. Lost in tracking: Uncertainty-guided cardiac cine MRI segmentation at right ventricle base. pages 415–424. Springer Nature Switzerland, 2024b. doi: 10.1007/978-3-031-72114-4_40.

Zhiwei Zhao, Bin Liu, Yan Lu, Qi Chu, and Nenghai Yu. Unifying multi-modal uncertainty modeling and semantic alignment for text-to-image person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7534–7542, 2024c.

Ervine Zheng, Qi Yu, Rui Li, Pengcheng Shi, and Anne Haake. A continual learning framework for uncertainty-aware interactive image segmentation. 35(7):6030–6038, 2021. doi: 10.1609/aaai.v35i7.16752. URL https://ojs.aaai.org/index.php/AAAI/article/view/16752.

Zhihao Zheng, J Scott Lauritzen, Eric Perlman, Camenzind G Robinson, Matthew Nichols, Daniel Milkie, Omar Torrens, John Price, Corey B Fisher, Nadiya Sharifi, et al. A complete electron microscopy volume of the brain of adult drosophila melanogaster. *Cell*, 174(3):730–743, 2018.

Hao Zhou, Yanze Zhang, and Wenhao Luo. Safety-critical control with uncertainty quantification using adaptive conformal prediction. In *2024 American Control Conference (ACC)*, pages 574–580. IEEE, 2024.

Tongxue Zhou and Shan Zhu. Uncertainty quantification and attention-aware fusion guided multi-modal MR brain tumor segmentation. 163:107142, 2023. doi: 10.1016/j.compbiomed.2023.107142.

Zhi-Hua Zhou. Ensemble methods: Foundations and algorithms. *CRC Press*, 2012.

Yufan Zhu, Weisheng Dong, Leida Li, Jinjian Wu, Xin Li, and Guangming Shi. Robust depth completion with uncertainty-driven loss functions. 36(3):3626–3634, 2022. doi: 10.1609/aaai.v36i3.20275. URL https://ojs.aaai.org/index.php/AAAI/article/view/20275.

Ke Zou, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu. Tbrats: Trusted brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 503–513. Springer, 2022.

Ke Zou, Zhihao Chen, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu. A review of uncertainty estimation and its application in medical imaging. *Meta-Radiology*, page 100003, 2023.

Martin Žukovec, Lara Dular, and Žiga Špiclin. Modeling multi-annotator uncertainty as multi-class segmentation problem. In *International MICCAI Brainlesion Workshop*, pages 112–123. Springer, 2021.

Appendix A. Foundations of Uncertainty Quantification in Machine Learning

This appendix provides a brief overview of key UQ methods developed across different machine learning applications. We do not aim at a comprehensive review, as previous work already covered the theoretical foundations of UQ in great depth (Gruber et al., 2023; Liu et al., 2018b; Seoni et al., 2023b). Instead, we focus on presenting key methods and trade-offs relevant for readers interested in applying UQ techniques, particularly in healthcare contexts as described in Section 2. Figure A1 summarizes the common UQ approaches mapped to each stage of the machine learning pipeline.

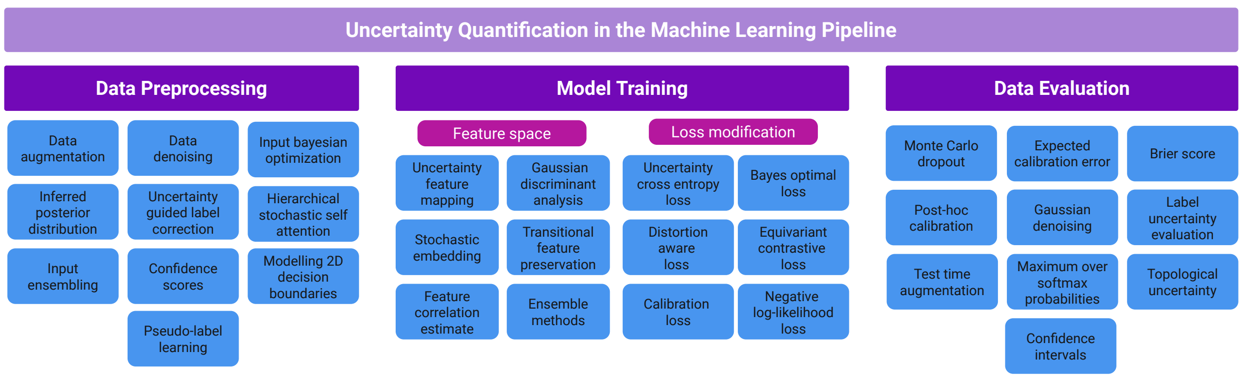

Figure A1: **Uncertainty quantification across the ML pipeline**. Key UQ methods from different domains applied at each stage: data processing, model training, and evaluation.

A.1. Probabilistic Methods

These approaches represent uncertainty using probability distributions and statistical models, providing a mathematical framework for modeling variability and randomness in the data and model architecture.

Bayesian Neural Networks (BNNs). BNNs learn distributions over network weights rather than relying on fixed point estimates, providing a principled Bayesian framework for modeling uncertainty. Training is computationally expensive due to intractable posterior inference, often requiring approximations that may affect calibration (Blundell et al., 2015; Gal and Ghahramani, 2016). Inference can also be slow depending on the approximation method used. BNNs have been applied to clinical tasks such as outcome prediction and survival analysis, where quantifying model confidence is critical (Wang et al., 2020b; Herzog et al., 2020).

Gaussian Processes (GPs). GPs are non-parametric models that define distributions over functions fitting the data, offering strong uncertainty estimates, especially for small datasets. They provide exact Bayesian inference but scale poorly ($O(n^3)$ complexity), limiting applicability to large-scale problems (Dietterich, 2024; Liu et al., 2018b). GPs have been employed in modeling disease trajectories and personalized medicine applications (Futoma, 2018; Puri et al., 2022).

Ensemble Methods. Ensemble techniques improve predictive performance and uncertainty estimation by aggregating outputs from multiple models (Dietterich, 2000). Approaches include deep ensembles (Lakshminarayanan et al., 2017), bagging (Rokach, 2010), and snapshot ensembles (Zhou, 2012). Ensembles require training multiple models independently, resulting in high memory and training costs, and inference overhead from multiple forward passes (Dietterich, 2000). They have demonstrated success in diagnostic tasks such as pneumonia detection from chest X-rays and sepsis prediction (Shilo et al., 2020; Valk et al., 2021).

A.2. Non-Probabilistic Methods

These methods quantify uncertainty without relying on explicit probability distributions, often using bounded sets or evidence-based frameworks.

Evidential Deep Learning (EDL). EDL incorporates uncertainty estimation into the learning process by modeling evidence through Dirichlet distributions. EDL methods typically avoid sampling-based inference, leading to faster training and single-pass inference while maintaining theoretical rigor (Sensoy et al., 2018). They have been applied to disease classification tasks, enabling models to flag uncertain predictions for clinical review and improving AI-driven diagnostic safety (Deng et al., 2023).

Fuzzy Logic. Fuzzy logic captures uncertainty by modeling partial membership across multiple classes, useful for handling imprecise or vague information. Training complexity varies with rule complexity but is generally lower than probabilistic approaches; inference is fast but lacks probabilistic confidence estimates (Zadeh, 1988; Kosko and Isaka, 1993). In healthcare, fuzzy logic has been applied to clinical decision systems where test results or symptom categories are inherently ambiguous (Nguyen et al., 2015; Gürsel, 2016).

A.3. Hybrid Methods

These methods are combinations of different probabilistic and non-probabilistic approaches for UQ that leverage the strengths of each framework.

Bayesian Deep Ensembles. Combine the strengths of ensemble learning and Bayesian learning, leveraging the diversity of multiple indepoendently trained models with Bayesian principles while incorporating priors or randomized initialization (Wild et al., 2023; Abulawi et al., 2024). They have been used in the diagnosis of chronic diseases (Abdollahi et al., 2021) and prediction of medication adherence (Gu et al., 2021).

Conformal Prediction. By constructing prediction sets or intervals that contain the true label with a user-specified confidence level, conformal predictors offer a formal measure of uncertainty with guaranteed coverage probabilities (Angelopoulos and Bates, 2021). This approach has been employed to enhance the confidence in predictions for skin lesions (Lu et al., 2021) and genomics (Papangelou et al., 2024).

A.4. Key Domains

In reviewing SOTA UQ methods, several key application domains emerge and can be summarized as follows.

Mathematical Foundations. This represents the backbone of the UQ field, focusing on developing rigorous theoretical frameworks, probabilistic models, and algorithmic proofs to enhance uncertainty modeling, model calibration, and learning stability (Wang et al., 2021; Pei et al., 2022; Ghosh et al., 2023; Arora et al., 2024). Innovative approaches include normalizing flows, function space priors, and infinite-width neural networks to improve both epistemic and aleatoric uncertainty estimation (Bae et al., 2021; Adlam et al., 2020; Berry and Meger, 2023; Schnaus et al., 2023). Studies aiming to produce more reliable and interpretable predictions focus on methods such as conformal prediction, test time augmentation, mutual information, temperature scaling, and ensemble learning (Kuleshov and Deshpande, 2022; Hekler et al., 2023; Li et al., 2023b; Wang et al., 2023; Stutts et al., 2023).

Optimization. Closely linked to mathematical foundations, optimization is a widely studied domain to refine model performance, calibration, and generalization. Techniques such as gradient-based optimization, regularization strategies, and novel loss functions are employed to mitigate overfitting and calibrate predictions during training (Heiss et al., 2021; Xia et al., 2021; Dai et al., 2023; Daheim et al., 2023).

Computer Vision (CV). CV is a major application domain where UQ methods are applied to critical tasks such as crowd counting, image segmentation, depth estimation, image denoising, multi-view stereo, and medical imaging (Qu et al., 2021; Mao et al., 2021a; Manor and Michaeli, 2023; Li et al., 2023b; Kahl et al., 2024; Wang et al., 2022a). Other applications focus on identifying ambiguous or out-of-distribution data, object tracking, image-to-image regression, vision-matting and facial expression recognition (Wang et al., 2020a; Zhang et al., 2021; Nussbaum et al., 2022; Angelopoulos et al., 2022; Zhang et al., 2022b; Wu et al., 2023b).

Natural Language Processing (NLP). Relevant methods in NLP domain focus on uncertainty at the level of token-level prediction, text generation, dialogue retrieval, code generation, and LLM fine-tuning, addressing challenges related to confidence estimation and calibration both during data processing and model training (Malinin and Gales, 2020; Hou et al., 2023; Xiong et al., 2023; Johnson et al., 2023b; Gupta et al., 2024; Lee et al., 2024; Liu et al., 2024).

Reinforcement Learning (RL). Primarily used to optimize the exploration-exploitation trade-off, improve policy robustness, enable multitasking in offline RL, and enhance Q-learning in uncertain environments (Wu et al., 2021; Liu et al., 2022; Xie et al., 2022a; Bai et al., 2024). Other methods include Bayesian RL, meta-learning, and UQ exploration, which improve the reliability of adaptive learning in dynamic environments (Li et al., 2021a; Gong et al., 2023b; Zhang et al., 2024a).

Graph Neural Networks (GNNs). GNNs, known for modeling complex relational data, are explored for UQ in applications such as molecular modeling, adversarial robustness and social network analysis (Shanthamallu et al., 2021; Feng et al., 2021; Yu et al., 2023; Wollschläger et al., 2023; Trivedi et al., 2024).

Multimodal Learning. The development of UQ methods in multimodal learning has been limited by the lack of high-quality, large-scale multimodal datasets, resulting in being constrained to highly specialized applications such as text-to-image person reidentification, sensing in soft robotics systems, malware detection and traffic trajectory planning (Brown et al., 2020; Ding et al., 2021; Zhao et al., 2024c; Lafage et al., 2024). In healthcare, current studies include depression and stress detection, sentiment analysis, mortality prediction, clinical imaging segmentation, and mRNA classification (Foltyn and Deuschel, 2021; Han et al., 2022; Ahmed et al., 2023; Bezirganyan, 2023; Huang et al., 2025).

Emerging Fields. The application of UQ remains limited in several research areas, including: (1) Generative adversarial networks (GANs) in addressing vulnerabilities to adversarial attacks and enhancing resilience against uncertainty manipulation (Hu et al., 2021c; Galil and El-Yaniv, 2021; Schweighofer et al., 2023), (2) federated learning and (3) contrastive learning, being used independently in handling noisy, heterogeneous data while ensuring data privacy and model generalizability (Plassier et al., 2023; Kotelevskii et al., 2024; Wang et al., 2024). Additionally, evidential deep learning has gained increasing attention in recent years for its potential integration into clinical decision support systems to enhance medical diagnostics and risk assessment (Deng et al., 2023; Ashfaq et al., 2023; Li et al., 2024; Jürgens et al., 2024; Liu and Ji, 2024). These research areas reflect a dynamic, interdisciplinary effort to develop safe and robust ML models. Given the importance of UQ in high-stakes decision-making, we focus on its development and impact in the healthcare domain across each stage of the ML pipeline.

Appendix B. Medical Datasets for Uncertainty Quantification in Healthcare

A wide range of medical datasets has been developed to support machine learning research across diverse clinical tasks and conditions. These datasets serve as critical benchmarks for training, validation, and evaluation, enabling the development and assessment of uncertainty quantification models in healthcare. Table B1 presents a structured overview of open-access datasets, categorized by medical domain, along with their frequency of use across the reviewed studies. A key observation from our analysis is the strong reliance on standardized clinical datasets, largely driven by the practical constraints of medical data accessibility, suggesting that dataset availability often outweighs theoretical considerations in shaping UQ research directions. The table also highlights the primary clinical task associated with each dataset, offering a comprehensive reference for researchers integrating UQ methods into medical applications.

Private Datasets. In addition to public datasets, many studies leverage institution-specific private datasets, particularly for imaging-based applications and rare disease research. Notable examples include endoscopic submucosal dissection procedures, knee MRI for musculoskeletal analysis (Browning et al., 2021), MRI-to-PET translation (Upadhyay et al., 2021), fetal brain MRI for neurodevelopmental assessment (Fu et al., 2025), sleep pattern monitoring (Kang et al., 2021), and specialized cardiology (Adams and Elhabian, 2023) and oncology datasets, such as ovarian and prostate cancer imaging (Konuk et al., 2024; Dong et al., 2024). Although not openly accessible, these datasets provide valuable insights into specialized clinical domains, where UQ methods contribute to enhancing diagnostic confidence and decision support.

Table B1: Summary of Open-Access Healthcare Datasets for Uncertainty Quantification Research

Dataset	Clinical Task	No. of Papers*
Multi-domain		
MIMIC-III (Johnson et al., 2016)	Electronic health records	1
eICU (Pollard et al., 2018)	Multi-center critical care database	1
FastMRI (Zbontar et al., 2018)	Knee, brain, prostate, breast classification	2
MedMNIST (Yang et al., 2023b)	Biomedical classification & segmentation	1
CPRD (Conrad et al., 2018)	Electronic health records	1
TOP (Fu et al., 2022)	Clinical trial outcome prediction	1
BioVid (Werner et al., 2017)	Pain assessment	1
PAMAP2 (Reiss and Stricker, 2012)	Activity monitoring	1
USC Alcohol Concentration (Saldich et al., 2020)	Blood alcohol concentration estimation	1
MIMIC-IV (Johnson et al., 2023a)	Time-to-Event Prediction, CATE Estimation	2
Microscopy		
TEMCA2 (Zheng et al., 2018)	Electron microscopy of adult fly brain	1
BSCCM (Pinkard et al., 2024)	Single white cell microscopy	1
MitoEM (Wei et al., 2020)	3D Mitochondria instance segmentation	1
Kasthuri++ (Casser et al., 2020)	Mitochondria segmentation	1
Lucchi++ (Casser et al., 2020)	Mitochondria segmentation	1
Cardiology		
ACDC (Bernard et al., 2018)	MRI segmentation	5
HMC-QU (Degerli et al., 2021)	Myocardial infarction detection in ECG	2
Echonet-Dynamic (Ouyang et al., 2019)	Cardiac cycle assessment	2
ECG5000 (Dau et al., 2019)	Congestive heart failure detection	1
PhysioNet/CinC Challenge 2020 (Alday et al., 2020)	Cardiac abnormality detection in ECG	1
M&Ms (Campello et al., 2021)	Multi-disease cardiac segmentation	4
CPSC2018 (Liu et al., 2018a)	ECG classification	1

Dataset	Clinical Task	No. of Papers*
TMED 2 (Huang et al., 2022b)	ECG classification	2
PTB ECG Database (Bousseljot et al., 1995)	Myocardial infarction detection in ECG	1
Atrial Segmentation Challenge (Xiong et al., 2021)	Atrial segmentation	1
CAMUS (Leclerc et al., 2019)	Echocardiographic Image Segmentation	6
UPL (Wu et al., 2023a)	Heart MRI segmentation	1
UCI Heart-Disease (Janosi et al., 1988)	Heart disease classification	1
UCR Time Series Archive (Dau et al., 2019)	Heart Disease Detection in ECG	1
CVSim (Heldt et al., 2010)	Simulating the Dynamics of the Human Cardiovascular System	1
Gastroenterology		
CholecSEG8k (Hong et al., 2020)	Cholecystectomy segmentation	1
DeepOrgan (Roth et al., 2015)	Pancreas segmentation	1
Kvasir-Seg (Jha et al., 2020)	Colorectal polyp segmentation	1
PolypDB (Silva et al., 2014)	Wireless capsule endoscopy detection	1
Neurology		
IXI (London, 2008)	Brain MR Images from Healthy Subjects	1
ISLES 2018 (Winzeck et al., 2018)	Ischemic stroke lesion segmentation	1
WMH Segmentation (Kuijf et al., 2019)	White matter hyperintensities segmentation	1
Calgary-Campinas-359 (Souza et al., 2018)	Brain segmentation	1
BRAVO (Vollmer et al., 2014)	Evaluating laquinimod in RRMS	1
OPERA1 (Hauser et al., 2017)	Evaluating Treatment Effects in RMS	1
DEFINE (Gold et al., 2012)	Evaluating BG-12 in RMS	1
WU-Minn HCP (Van Essen et al., 2013)	Characterization of brain connectivity	1
HCP Lifespan Studies (Harms et al., 2018)	Diffusion MRI images	1
INTERGROWTH (Papageorghiou et al., 2018)	3D Ultrasound fetal brain volumes	1
Prisma (Alexander et al., 2017)	MRI Image Enhancement	1
IDH-Glioma-MRI (Figini et al., 2018)	IDH Prediction in Brain MRI Images	1
Oncology		
BCDR (Moura et al., 2013)	Benchmarking for Breast Cancer Diagnosis	1
BreastPathQ (Petrick et al., 2021)	Breast tumor cellularity assessment	1
ISIC 2018-2019 (Hardie et al., 2018)	Skin lesion detection	7
Breast Histopathology (Kaggle) (Janowczyk, 2016)	Breast Tumor Histopathology	1
KiTS19-21 (Heller et al., 2021)	Kidney CT segmentation	2
WDBC (Wolberg William and W, 1993)	Breast cancer classification	1
LiTS (Bilic et al., 2023a)	Liver tumor segmentation	1
Bone Metastates (Lin et al., 2016)	Bone tumor segmentation	1
HAM10000 (Tschandl et al., 2018)	Skin lesion detection	2
BraTS 2018-2019 (Hardie et al., 2018)	Brain tumor segmentation	4
The Cancer Genome Atlas (Abeshouse et al., 2017)	Different Types of Tumor Detection	1
ISPY I (Garrucho et al., 2024)	Breast Cancer Tumor Segmentation	1
OrganCMNIST (Bilic et al., 2023b)	Liver Tumor Segmentation	1
Derm-Skin (DERM) (Pacheco et al., 2020c)	Skin Cancer Detection	1
SkinCon (Ren et al., 2024b)	Skin Cancer Detection	1
ClinSkin (Pacheco et al., 2020a)	Skin Cancer Detection	1
PAD-UFES-20 (Pacheco et al., 2020b)	Skin Cancer Detection	1
QUBIQ 2021 (Žukovec et al., 2021)	Skin Cancer Detection	1

Dataset	Clinical Task	No. of Papers*
BrainMRI (Nickparvar, 2021)	Brain Tumor Detection	1
HECKTOR (Oreiller et al., 2022)	Head & neck tumor segmentation in PET/CT	1
Tumor Growth Model Geng et al. (2017)	Time-to-Event Prediction, CATE Estimation	1
GBSG Royston and Altman (2013)	Time-to-Event Prediction, CATE Estimation	1
SUPPORT Knaus et al. (1995)	Time-to-Event Prediction, ITE Estimation	1
Ophthalmology		
Cataract-101 (Schoeffmann et al., 2018)	Cataract Surgery Videos	1
EyePACS (EyePACS, 2015)	Diabetic retinopathy detection	2
APTOS 2019 (Karthik and Dane, 2019)	Diabetic retinopathy detection	2
REFUGE (Orlando et al., 2020)	Glaucoma assessment	1
DPL (Chen et al., 2021)	Fungus Image Segmentation	1
LAG (Li et al., 2019)	Glaucoma Detection	1
Diabetic Retinopathy (Kaggle) (Dugas et al., 2015)	High-resolution retina images	1
Pulmonology		
ChestX-ray8 (Wang et al., 2017)	Pulmonary disease detection	2
TBX11K (Pan et al., 2022)	Tuberculosis Diagnosis	1
Shenzhen Chest X-ray (Jaeger et al., 2014)	Pulmonary disease detection	1
LIDC-IDRI (Armato III et al., 2011)	Lung nodule detection & segmentation	2
JSRT (Shiraishi et al., 2000)	Lung Nodules Classification	2
RSNA (Shih et al., 2019)	Pneumonia Detection	1
VinDr-CXR (Nguyen et al., 2020)	Chest X-Ray Disease Detection	2
CXAD (Cai et al., 2022)	Chest X-Ray Disease Detection	1
Radiology		
SUPERB (Johansson et al., 2020)	Vertebral Fractures Diagnosis	1
HC18 (van den Heuvel et al., 2018)	Fetal Head Circumference Measurement	1
TN-SCUI (Xie et al., 2022b)	Thyroid Segmentation & Classification	1
BloodMNIST (Acevedo et al., 2020)	Disease Classification	1
Kvasir-SEG (Jha et al., 2020)	Polyp Segmentation	1
FSM (Yang et al., 2022a)	Polyp Segmentation	1
PICAI (Saha et al., 2023)	Prostate Cancer Detection	1
B-Fract (Wang et al., 2022b)	Hairline fracture detection	1
Low-Dose CT Images (Moen et al., 2021)	Low-dose CT denoising	1

* Number of papers in our survey that use the dataset.

Appendix C. Healthcare Studies Organized by ML Pipeline Stage

Table C1 summarizes the UQ healthcare studies reviewed in this survey. Each study is grouped by medical domain and annotated with the corresponding ML pipeline stage at which UQ methods are applied. We also report the specific clinical tasks addressed and the datasets utilized. Notably, most studies implement UQ at the model training stage, with a predominant focus on image classification and segmentation tasks.

Table C1: Summary of Healthcare Studies Implementing Uncertainty Quantification Methods

Reference	Data	Train	Eval	UQ Method	Task	Datasets
			Cardiology			
(Gu et al., 2024a)	✓	-	-	Conformal Prediction	Classification	TMED-2, CIFAR-10-Derived, Private Aortic Stenosis*
(Oh et al., 2024)	✓	✓	-	Acoustic Diffusion Method	Segmentation	Echonet-Dynamic, HMC-QU, CAMUS
(Zhao et al., 2022)	-	✓	-	Bayesian Learning	Segmentation	ACDC, MnM
(Zhao et al., 2024a)	-	✓	-	Bayesian & Ensemble Learning	Segmentation	ACDC
(Vaseli et al., 2023)	-	✓	-	Prototype Based Models	Classification	TMED-2
(Lu et al., 2023)	-	✓	-	Uncertainty Masks in Prototype Learning	Segmentation	TBAD
(Jahmunah et al., 2023)	-	✓	-	Dirichlet Distribution Classifier	Time-Series Classification	Physikalisch-Technische Bundesanstalt database
(Adams and Elhabian, 2023)	-	✓	-	Bayesian Learning & Variational inference	Shape Prediction	Private Left Atrium Dataset from UUtah*
(Kazemi Esfeh et al., 2022)	-	✓	-	Ensemble Learning	Regression	EchoNet-Dynamic
(Zhang et al., 2024c)	-	✓	-	Displacement and Variance Estimators	Segmentation	ACDC, CAMUS, Private 3D Echo*
(Barandas et al., 2024a)	-	✓	✓	Monte Carlo Dropout & Ensemble Learning	Classification	PhysioNet/CinC Challenge 2020
(Vranken et al., 2021)	-	-	✓	Monte Carlo Dropout & Variational Inference	Time-Series Classification	CPSC2018-Dynamic, UMCU-Triage*, UMCU-Diagnose*
			Neurology			
(Tanno et al., 2019)	✓	✓	-	Heteroscedastic Noise & Variational Dropout	Image Enhancement	WU-Minn HCP, Lifespan, Prisma †
			General Surgery			
(Yang et al., 2023a)	-	✓	✓	Evidential Learning	Segmentation	CholecSeg8K, Private Endoscopy Dataset*
			Ophthalmology			
(Hu et al., 2021a)	✓	-	✓	Calibration Error Estimation	Classification	EyePACS, FastMRI
(Ren et al., 2023)	-	✓	-	Evidential Deep Learning	Classification & Segmentation	REFUGE, ISPY I

Reference	Data	Train	Eval	UQ Method	Task	Datasets
(Leibig et al., 2016)	-	✓	-	Monte Carlo Dropout	Classification	Kaggle Diabetic Retinopathy Detection Dataset
(Wu et al., 2022)	-	✓	✓	Ensemble & Bayesian Learning	Surgery Time Estimation	Cataract-101
(Band et al., 2022)	-	✓	✓	Bayesian Learning	Classification	EyePACS, APTOS
				Orthopedics		
(Browning et al., 2021)	-	✓	-	Q-Learning	Detection	Private Knee MRI*
(Skärström et al., 2024)	-	✓	✓	Likelihood Scores	Fracture Analysis & Classification	SUPERB
(Teichmann et al., 2024)	-	✓	✓	Monte Carlo Dropout	Segmentation	Private Organs Dataset*
				Oncology		
(Li et al., 2022a)	✓	✓	-	Label Probability Distribution	Tumor Cellularity Scoring	BreastPathQ
(Aljuhani et al., 2022)	-	✓	-	Monte Carlo Dropout	Classification	TCGA
(Luo et al., 2021)	-	✓	-	Rectified Pyramid Consistency	Segmentation	Private Nasopharyngeal Carcinoma*
(Zou et al., 2022)	-	✓	-	Logic Theory	Segmentation	BraTS
(Hung et al., 2024)	-	✓	-	Evidential Deep Learning	Classification	PICAI
(Zepf et al., 2024)	-	✓	-	Laplacian Segmentation Network	Segmentation	ClinSkin, PAD-UFES-20, QUBIQ 2021 †
(Dong et al., 2024)	-	✓	-	Evidential Deep Learning	Image Grading	Private Prostatic Cancer Dataset*
(Ren et al., 2024a)	-	✓	-	Conformal Prediction	Classification	SkinCON
(Thiagarajan et al., 2022)	-	✓	-	Bayesian Learning	Classification	Breast Histopathology Kaggle
(Schott et al., 2024)	-	✓	-	Localized Gradients	Segmentation	LiTS, Bone Metastates
(Buddenkotte et al., 2023)	-	✓	-	Bayesian & Ensemble Learning	Segmentation	KiTS19, Private Ovarian Cancer CT Scans*
(Hu et al., 2020)	-	✓	-	Gaussian Process	Radiogenomics EGFR amplification	Private Self-Recorded Data*
(Abdar et al., 2021)	-	✓	-	Monte Carlo Dropout & Ensemble Learning	Classification	ISIC 2019, HAM10000
(Zhou and Zhu, 2023)	-	✓	-	Monte Carlo Dropout	Segmentation	BraTS 2018 & 2019
(Lennartz and Schultz, 2023b)	-	✓	-	Active Selection Sampling	Segmentation	DPL, FSM, UPL
(Sahlsten et al., 2024)	-	✓	✓	Bayesian Learning	Segmentation	HECKTOR, Private U-Texas Cancer Center*
(Peluso et al., 2024)	-	-	✓	Deep Abstaining Classifier	Clinical Text Classification	NCI SEER Report

Reference	Data	Train	Eval	UQ Method	Task	Datasets
(Hamedani-KarAzmoudehFar et al., 2023)	-	-	✓	Monte Carlo Dropout & Ensemble Learning	Classification	WDBC
(Horii and Chikahara, 2023)	-	✓	✓	Bayesian Gaussian-Orocess-based UQ Framework	CATE Estimation	Synthetic, ACIC
(Deng et al., 2025)	-	✓	✓	Approximate Bayesian UQ	CATE Estimation	Simulated, MIMIC-IV
(Li et al., 2021b)	-	-	✓	Monte Carlo	CATE Estimation	CVSim, Cancer Growth
(Hess et al., 2024)	-	✓	✓	Bayesian Neural Controlled Differential Equation	CATE Estimation	Simulated, Pharmacokinetic-pharmacodynamic Tumor Growth Model
(Brouwer et al., 2022)	-	✓	✓	Uncertainty-Aware Latent Neural ODE	Individualized treatment effect Estimation	Synthetic, Cardiovascular System Modeling, Pharmacodynamics Model
(Huang et al., 2024b)	-	✓	✓	Evidential Regression Network	Time-to-Event Prediction	Synthetic, Simulated, METABRIC, GBSG, SUPPORT, MIMIC-IV
				Pulmonology		
(Li et al., 2023a)	-	✓	-	Dirichlet Distribution Classifier	Classification	ISIC18, Chest XRay8
(Yang et al., 2022b)	-	✓	-	Attention Masks for Uncertainty	Segmentation	LIDC-IDRI
				Specialized Applications		
(Ji et al., 2024)	-	✓	-	PCA Based Uncertainty Weighting	Pain Assessment	Biovid, Private Apon Dataset*
(Kang et al., 2021)	-	✓	-	Shannon Entropy	Sleep Pattern Assessment	Private Sleep Pattern Dataset*
(Lu et al., 2024)	-	✓	-	Hierarchical Interaction Network	Clinical Trial Approval Prediction	TOP clinical Trial Approval Prediction Benchmark
(Dusenberry et al., 2020)	-	✓	-	Bayesian & Ensemble Learning	Intensive Care Unit	MIMIC-III, eICU
(Oszkinat et al., 2023)	-	✓	-	Residual-Augmented Loss Function	Blood Alcohol Concentration Estimation	Alcohol Concentration Data
(Jeong et al., 2024)	-	✓	-	Pixel-Wise Uncertainty for Diffusion Model	Image Generation & Adversarial Attacks	Kvasir-SEG, ETIS-Larib Polyp DB
(Li et al., 2021c)	-	✓	✓	Gaussian Processes	Classification	CPRD
(Konuk et al., 2024)	-	✓	✓	Entropy & Confidence Based Uncertainty	Classification	Private OMLC-RS*
(Durso-Finley et al., 2023a)	-	-	✓	Bayesian Causal Models	Factual Error Correlation with Uncertainty	BRAVO, OPERA 1-2, DEFINE †

Reference	Data	Train	Eval	UQ Method	Task	Datasets
				Medical Imaging		
(Lee et al., 2023)	✓	-	-	Diffusion Probabilistic Modeling	Noise Reduction	Private MR Dataset*
(Khader et al., 2023)	✓	-	-	Diffusion Probabilistic Modeling	Data Generation	ADNI, Breast MRI
(Iuliano et al., 2024)	✓	✓	-	Diffusion Probabilistic Modeling	Data Generation	NLM Malaria
(Adib et al., 2023)	✓	✓	-	Diffusion Probabilistic Modeling	Data Generation	MIT-BIH Arrhythmia
(Angelopoulos et al., 2022)	✓	✓	-	Pixel-wise Uncertainty Intervals	Segmentation	BSCCM, TEMCA2, FastMRI
(Das et al., 2024)	✓	✓	-	Confidence Guided Pseudo-Label Optimizer	Segmentation	VinDr-CXR, TBX11K, B-Fract
(Scalco et al., 2024)	-	✓	-	Bayesian & Ensemble Learning	Segmentation	Kidney Tumor Segmentation Challenge 2021
(Judge et al., 2023b)	-	✓	-	Gaussian Probability Distributions	Segmentation	CAMUS, JSRT, Private US Dataset*
(Lennartz and Schultz, 2023a)	-	✓	-	Distance Regularization	Segmentation	Calgary-Campinas-359, ACDC, M&MS
(Larrazabal et al., 2023)	-	✓	-	KL Divergence	Segmentation	Atrial Segmentation Challenge, WMH
(Xiang et al., 2022a)	-	✓	-	Unsupervised Learning	Segmentation	DeepOrgan, 2018 Atria Segmentation Challenge
(Judge et al., 2022b)	-	✓	-	Contrastive Learning	Segmentation	CAMUS, HMC-QU, Shenzen †
(Xie et al., 2022c)	-	✓	-	Uncertainty Attention Module	Segmentation	CAMUS, TN-SCUI, HC18
(Fu et al., 2025)	-	✓	-	Uncertainty Weighted Class Activation Maps	Segmentation	Private Fetal Brain Dataset*
(Abdar et al., 2023)	-	✓	-	Monte Carlo Dropout	Classification	Retinal OCT, Lung CT, Pneumonia Chest X-Ray
(Upadhyay et al., 2021)	-	✓	-	Uncertainty-Guided GAN	Classification	IXI, Private PET to CT Dataset*
(Sharma et al., 2024)	-	✓	-	Entropy Driven Distillation Learning	Classification	HAM10000, APTOS
(Qendro et al., 2021)	-	✓	-	Ensemble Learning	Classification Benchmarking	ECG5000, EEG, ISIC2018
(Abboud et al., 2024)	-	✓	-	Bayesian Learning	Classification & Segmentation	ISIC, ChestMNIST, LIDC-IDRI
(Samareh and Huang, 2019)	-	✓	-	Contemporaneous Longitudinal Index	Degenerative Disease Modeling	Private Alzheimer's Dataset*
(Ramesh et al., 2024)	-	✓	-	Multi-head Geometric Transformations	3D Pose Prediction	INTERGROWTH Fetal Brain Ultrasound

Reference	Data	Train	Eval	UQ Method	Task	Datasets
(Chen et al., 2024a)	-	✓	✓	Conformal Prediction	Classification & Segmentation	ISIC 2018, BloodMNIST, OrganCMNIST
(Gong et al., 2023a)	-	✓	✓	Bayesian Learning & Knowledge Distillation	Image Denoising	Low-Dose CT Image
(Popordanoska et al., 2021)	-	✓	✓	Model Calibration	Segmentation	BraTS, ISLES
(Kushibar et al., 2022)	-	✓	✓	Ensemble Learning	Segmentation	BCDR, MnM
(Shi et al., 2024)	-	✓	✓	Dempster-Shafer Theory	Segmentation	MitoEM-(R,H), Kasthuri++, Lucchi++ †
(Gu et al., 2024b)	-	✓	✓	Ensemble Learning	Segmentation	RSNA Pneumonia, VinDr-CXR, Brain MRI †
(Yang et al., 2022b)	-	✓	✓	Uncertainty Attention Masks	Segmentation	LIDC-IDRI
(Zhang et al., 2023b)	-	✓	✓	Monte Carlo Dropout & Ensemble Learning	Segmentation	Heart-Disease, ISIC2019, PAMAP2

* Private datasets for specific medical applications, †Evaluation on more than 3 datasets.
Classification and segmentation refer to imaging based tasks unless specified otherwise.