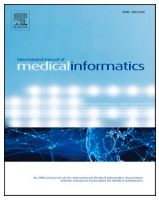
International Journal of Medical Informatics 197 (2025) 105846

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare☆

Massimo Salvi [a,*], Silvia Seoni [a], Andrea Campagner [b], Arkadiusz Gertych [c,d,e], U.Rajendra Acharya [f,g], Filippo Molinari [a], Federico Cabitza [b,h]

[a] Biolab, PoliToBIOMed Lab, Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy
[b] IRCCS Ospedale Galeazzi - Sant'Ambrogio, Milan, Italy
[c] Faculty of Biomedical Engineering, Silesian University of Technology, Zabrze, Poland
[d] Department of Surgery, Cedars-Sinai Medical Center, Los Angeles, CA, United States
[e] Department of Pathology and Laboratory Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, United States
[f] School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, Australia
[g] Centre for Health Research, University of Southern Queensland, Springfield, QLD 4300, Australia
[h] Department of Computer Science, Systems and Communication, University of Milano-Bicocca, Milan, Italy

ARTICLE INFO

Keywords:
Explainable AI
Uncertainty Quantification
Deep Learning
Healthcare
DL Model Interpretability
AI Trustworthiness

ABSTRACT

Background: The increasing use of Deep Learning (DL) in healthcare has highlighted the critical need for improved transparency and interpretability. While Explainable Artificial Intelligence (XAI) methods provide insights into model predictions, reliability cannot be guaranteed by simply relying on explanations.
Objectives: This position paper proposes the integration of Uncertainty Quantification (UQ) with XAI methods to improve model reliability and trustworthiness in healthcare applications.
Methods: We examine state-of-the-art XAI and UQ techniques, discuss implementation challenges, and suggest solutions to combine UQ with XAI methods. We propose a framework for estimating both aleatoric and epistemic uncertainty in the XAI context, providing illustrative examples of their potential application.
Results: Our analysis indicates that integrating UQ with XAI could significantly enhance the reliability of DL models in practice. This approach has the potential to reduce interpretation biases and over-reliance, leading to more cautious and conscious use of AI in healthcare.

1. Introduction

Deep Learning (DL) is a class of machine learning methods that gradually learns data representations at various levels of abstraction without explicitly defining the features to be used. DL models have achieved impressive performance in improving diagnosis and prognosis in healthcare domains [1,2], including oncology through histopathology interpretation [3], ophthalmology through OCT image analysis [4], and various other medical imaging modalities. However, their inherent complexity and lack of transparency often raise concerns about their interpretation of the model's decision, trustworthiness, and reliability [5].

Documented model failures have repeatedly shown how important interpretability and transparency are in healthcare AI applications. The risks of opaque decision-making processes are highlighted by research from Stanford [6] that revealed how AI models could inadvertently learn incorrect features – in their case, the algorithm was more likely to classify skin lesions as malignant when rulers were present in the images, simply because their training dataset contained more rulers in images of malignant lesions. This example demonstrates how a lack of transparency can mask potentially dangerous biases in model decision-making. Another instructive case [7] involved a pneumonia prediction model that showed strong internal performance but failed to generalize across hospitals, partly because it learned confounding variables, including whether X-rays were portable (common in severely ill inpatients). The study also demonstrated that the model could identify the originating hospital system with over 99.9 % accuracy by detecting subtle differences in image acquisition and processing, highlighting how

☆ This article is part of a special issue entitled: 'Machine Learning and Uncertainty Modeling' published in International Journal of Medical Informatics.

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

https://doi.org/10.1016/j.ijmedinf.2025.105846
Received 14 July 2024; Received in revised form 19 February 2025; Accepted 19 February 2025
Available online 21 February 2025

AI models may learn to rely on spurious correlations rather than clinically relevant features when making predictions.

In this scenario, Explainable Artificial Intelligence (XAI) emerges as a vital paradigm to tackle these challenges by providing explanations for the model's predictions [8,9,10,11]. XAI encompasses methods and techniques designed to make AI decisions more transparent and understandable to human users. While several approaches have been proposed to complement XAI, including interpretable feature engineering [12], attention mechanisms [13], and case-based reasoning [14], these methods often focus on different aspects of model interpretation without directly addressing the reliability of the explanations themselves. Moreover, the explanations alone may not be sufficient to guarantee the model's reliability, especially in healthcare applications [5]. This limitation has led to the increasing importance of Uncertainty Quantification (UQ), which complements XAI by offering quantitative measures of uncertainty associated with the model's predictions [15]. UQ enables the assessment of the model's confidence and identifies instances for which the model yields uncertain or unreliable predictions [16]. Despite the complementary nature of these approaches in enhancing DL model reliability, they have often been considered separately [17,18], resulting in a lack of understanding of their synergistic potential.

This position paper emphasizes the importance of integrating UQ with XAI methods to enhance the transparency and reliability of DL models. By quantifying the uncertainty of the explanations provided by XAI methods, a deeper understanding of the model's decision-making process and evaluating the trustworthiness of the explanations can be gained [19]. We postulate that UQ techniques should be incorporated within the XAI framework to expand how a system makes its output more understandable, transparent, and reliable [20,21]. By quantifying uncertainty and effectively communicating it to the user, the model avoids concealing uncertainty and instead aims to convey it transparently [22].

1.1. Objectives and Contributions

This position paper advocates for the integration of UQ and XAI proposing specific approaches where UQ techniques can be leveraged to serve XAI goals. Examples of such techniques include test time augmentation or Monte Carlo dropout during inference [23], or also techniques based on constructing uncertainty sets inspired by conformal prediction [19]. To advance this integration, we address three main objectives:

- Propose specific methodological approaches for combining UQ and XAI techniques, focusing on integrating uncertainty measures into existing explanation methods and developing unified frameworks that simultaneously address both aspects.
- Analyze the practical challenges in implementing combined UQ-XAI systems, including computational considerations, training requirements, clinical workflow integration, and regulatory compliance.
- Present concrete recommendations for future research directions and development of integrated UQ-XAI systems in healthcare, addressing validation methodologies and performance metrics.

By bridging the gap between XAI and UQ, we assert that it is possible to develop more comprehensive and reliable explainable AI systems that foster trust and pave the way for a broader adoption in critical domains like healthcare.

2. Background

2.1. Deep Learning in healthcare

From medical imaging to decision support systems, DL has transformed many facets of healthcare. Recent developments have achieved impressive results across several fields, highlighting the potential of these technologies to transform healthcare delivery and enhance patient outcomes.

DL models have demonstrated expert-level performance in a variety of medical imaging tasks. With exceptional results in brain tumor segmentation [24], mammography interpretation [25], and chest X-ray analysis [26], convolutional neural networks have shown particular success in radiology. These models have demonstrated the ability to recognize subtle patterns that may be difficult for human observers to consistently recognize, which could increase the accuracy of early detection and diagnosis.

These models have also been successfully used to predict drug response, disease progression and patient outcomes. Predicting intensive care unit mortality [27], identifying high-risk patients [28], and optimizing treatment plans [29] have all shown encouraging results in recent studies. DL models can now automatically analyze clinical notes and medical literature thanks to the successful application of natural language processing models to extract pertinent information from electronic health records [30].

The application of DL in pathology has advanced significantly, especially in the identification and categorization of cancer. Research has shown that it is possible to reduce inter-observer variability and improve diagnostic accuracy by identifying malignant tissue patterns [31] and grading tumors [32] with high accuracy. These developments strongly suggest that digital pathology aided by AI will play a significant role in clinical practice.

However, these and similar advances face significant challenges. The complexity of healthcare data, which includes high dimensionality and heterogeneity, severely hampers model development and validation. Furthermore, the nature of making decisions in healthcare requires highly reliable and interpretable AI. These challenges have motivated the development of more sophisticated approaches to model explanation and uncertainty, which we discuss in the following sections.

2.2. Explainable AI (XAI)

XAI is a broad field encompassing a wide range of techniques, methods, algorithms, and functions aimed at providing additional data beyond the output of an AI system (whether it is an estimate, categorical advice, recommendation, or content) to enrich and make this output more understandable and, above all, "actionable" – that is, applicable to the decision and choice of the best course of action by the human user [33]. XAI enhances transparency by providing human-justifiable explanations for model decisions, increases trust by defining confidence levels, and improves the understanding of bias and fairness through comprehensive model analysis [34]. The scope of XAI methods can primarily focus on explaining individual data instances (*local* explainability) or on understanding the model as a whole (*global* explainability) [35].

XAI methods can be categorized based on the use, i.e., how the XAI model is developed. One important differentiation is between ante-hoc (*intrinsic*) explainability, where the model is designed to be inherently interpretable from the outset (i.e., basing their prediction on explicitly defined if-then rules), and post-hoc explainability, where explanations are generated after the model has been trained. The ante-hoc models are sometimes called intrinsically interpretable or transparent models, i.e., the decision tree model. Within the realm of post-hoc methods, a further distinction exists between model-agnostic techniques [35]. Model-agnostic techniques can be applied to any model regardless of its internal structure or architecture. These methods typically work by analyzing the relationship between inputs and outputs without needing to know the model's inner workings. For example, permutation feature importance, which measures the impact on model performance when a feature is randomly shuffled, can be applied to any predictive model [36]. In contrast, model-specific techniques are tailored to work with

particular model types or architectures. For instance, extracting decision rules from a trained decision tree is a model-specific technique that leverages the tree structure. Similarly, analyzing weight matrices in neural networks is specific to that class of models. While model-agnostic methods offer flexibility and broad applicability, model-specific techniques can often provide more detailed insights by exploiting the model's particular structure.

Finally, XAI methods can be categorized based on their focus: input-based methods use backpropagation to link predictions directly to input features, while parameter-based methods analyze the effect of input perturbations on model outputs. The former includes gradient-based techniques, while the latter encompasses various perturbation analysis approaches [37]. Fig. 1 summarizes a typical AI framework that includes XAI methods.

The landscape of the XAI method is diverse, with methods falling into several main categories [33]. One prominent group is the attribution methods, which aim to identify the input features that have the significant influence on a model's predictions. Examples include LIME [38], SHAP [39], and various saliency mapping techniques such as those based on network gradients [40], Layer-wise Relevance Backpropagation (LRP) [41], pattern attribution, and Randomized Input Sampling for Explanation (RISE) [42]. These methods offer intuitive explanations by highlighting the importance of different input components, particularly for image and text data [43].

Another category is ante-hoc explainable, or transparent models, designed from the ground up to be interpretable. Decision trees [44] and linear models [45] are a prime example, and their inherent transparency makes them a popular choice in domains like finance or medicine, where interpretability is paramount [46]. Rule-based systems and rule extraction techniques, both leveraging principles of symbolic AI, are gaining renewed interest in the XAI context [47]. While both approaches result in explainable models, they differ in their starting points: rule-based systems begin with explicit rules, whereas rule extraction techniques infer rules from existing models.

The third category is counterfactual and contrastive methods that aim to make model predictions more understandable by explaining how the predictions would change if the input were different [48,49,50]. Such methods offer clear explanations that closely match human causal-based and/or case-based reasoning [51], by identifying which features' perturbations would affect and change the models' predictions (in this sense, they are similar to attribution methods, which can also be used as the basis for producing counterfactual explanations [52], though their focus and the kinds of explanations they produce are different [53]).

Novel techniques, like attention mechanisms in neural architectures, are being developed to address the shortcomings of existing methods and provide more precise explanations [54]. Argumentation-based frameworks, which treat rules as arguments to be reasoned over, are another promising avenue for enhancing the interpretability of XAI systems [55,56]. Concept-based learning algorithms, which explain predictions in terms of human-understandable attributes or abstractions [57,58] represent another active frontier of research in this field.

Recent advances in Large Language Models (LLMs) and generative AI have introduced novel approaches to XAI in healthcare. LLMs can generate natural language explanations of model decisions [59,60]. A particularly promising development is Retrieval-Augmented Generation (RAG), which enhances explainability by combining LLMs' generative capabilities with domain-specific knowledge retrieval. RAG can ground model explanations in relevant medical literature and clinical guidelines [61], providing evidence-based context for model decisions.

While these developments have advanced XAI, several limitations in understanding AI still exist. First, the deterministic explanations offered by current methods frequently lack uncertainty measures, which could lead to overconfidence during interpretations. Second, counterfactual explanations can be challenging to validate in healthcare settings and feature attribution techniques might not be consistent across similar inputs. Finally, the trade-off between model interpretability and performance continues to pose challenges for clinical implementation, highlighting the need for more robust approaches.

2.3. Uncertainty quantification (UQ)

UQ is a critical aspect of DL that focuses on measuring the uncertainties associated with the model's predictions, providing a quantitative assessment of the model's confidence and reliability [15]. UQ can be defined as the process of determining the extent to which model's predictions may be uncertain or unreliable. The primary objectives of UQ are to quantify the uncertainty considering both aleatoric (data-related) and epistemic (model-related) uncertainties [16], as well as other potential forms of uncertainty that cannot be easily reconciled with these previous ones [62]. UQ identifies uncertain or unreliable model instances, aiding users in interpreting outputs with appropriate caution and enhancing transparency by effectively communicating uncertainties [63].

Several approaches exist to quantify aleatoric uncertainty. Test-time augmentation is one of the most popular methods, which perturbs images during inference to estimate uncertainty [64]. The model's predictions from these variations are then combined to assess how consistent the results are. Similar approaches include voting techniques, which use both data variations and context to improve prediction accuracy [65]. Another common method involves variance attenuation

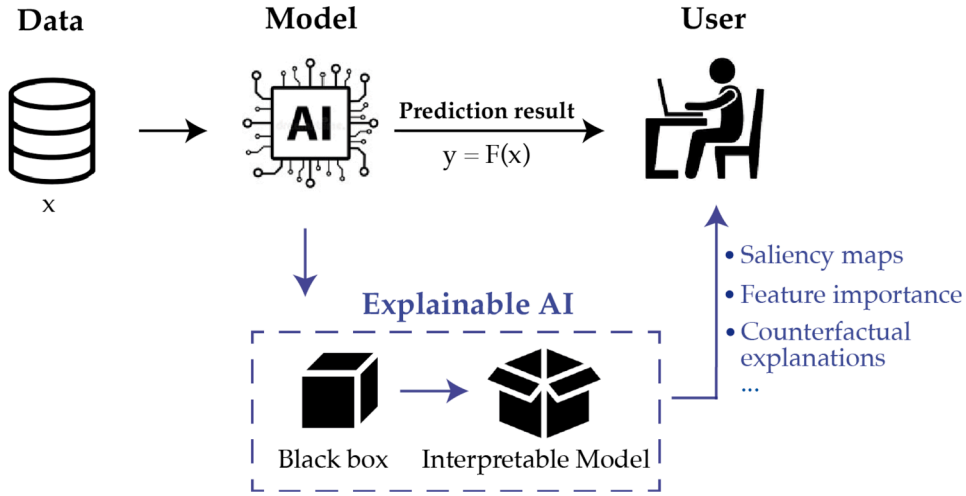


Fig. 1. Typical AI framework that includes XAI methods.

techniques, which measure uncertainty by incorporating data-specific variance into the model's training process [66,67,68,69].

For quantifying epistemic uncertainty, a commonly used technique is Monte Carlo dropout (MCD) [23]. This approach uses dropout layers during inference; it performs multiple forward passes with different dropout masks to measure variance among predictions, quantifying epistemic uncertainty. More generally, model ensembles can be used in a similar way to estimate models' prediction variance. Similar results can be achieved using model ensembles, where multiple models work together to estimate prediction uncertainty. Another approach is Bayesian neural networks, which introduced probability distributions over the model's weight [70]. In this way, these networks allow quantifying epistemic uncertainty during inference time.

While most work on UQ focuses on disentangling between aleatoric and epistemic uncertainty, other popular UQ approaches do not make such a distinction [15]. Among these techniques, an increasingly popular technique is conformal prediction [71], which is a post-hoc UQ method that quantifies uncertainty using ideas inspired by statistical inference (in particular, confidence intervals) by creating a confidence set with an associated p-value that quantifies the uncertainty about the confidence set containing the correct prediction.

Interpreting and communicating uncertainty measures in DL models is still difficult, even with the variety of UQ techniques available [72]. The absence of universal and standardized metrics for assessing the accuracy and dependability of uncertainty estimates in DL models is one of the significant drawbacks [16]. The uncertainty estimation produced by various UQ techniques can differ and is often inconsistent, making the UQ techniques challenging to validate and standardize in clinical settings. Furthermore, uncertainty measures can be difficult to interpret and use by clinicians [73]. Likewise, it can be challenging for healthcare providers to evaluate the clinical relevance of uncertainty estimates because the majority of UQ approaches only address the prediction uncertainty without explaining its sources. These limitations highlight the potential advantage of combining UQ with XAI techniques to offer more practical and clinically significant insights.

3. Current challenges of XAI in Deep Learning models

Despite the advancements in XAI methods, several fundamental challenges remain when applying them to DL frameworks [33]. One challenge is that many XAI methods are overly quantitative and perceived as objective and precise by users. However, they should rather be associated with an honest measure of uncertainty to convey XAI limitations [20]. For example, saliency maps (e.g., generated by gradient-weighted class activation mapping (Grad-CAM), attention mechanisms) can be calculated for different DL model layers, each yielding different results and essentially complicating the interpretation of the model's inferences [74]. Another example is the use of probability scores outputted by the model along with the class labels as a simple XAI method. However, these scores can be misleading if the model is not properly calibrated [75]. While Grad-CAM, attention, or probability score maps can provide insights into a model's general functioning or highlight areas important in decision-making, they have limited practical utility as they often do not identify specific features or mechanisms that drive model's predictions [76]. This limitation is sometimes referred to as the "white-box paradox".

The varying nature of different healthcare tasks further complicates these challenges. In detection or segmentation tasks, the model's output natively provides spatial information about the model's operating, making the interpretation and validation of XAI methods more straightforward (and yet potentially even more misleading [77,78]. Classification tasks, however, present additional complexity as they require specific methods to understand which features influenced the model's decision. This inherent characteristic of classification problems makes them particularly relevant for studying the integration of XAI and UQ methods, as robust uncertainty quantification becomes crucial for validating feature importance and model explanations.

To date, only a few XAI methods can provide a quantifiable degree of uncertainty. One example is the conformal prediction, which, while originally proposed only for UQ, can also be used as XAI [19,79]. In this approach, the AI does not provide a single explanation, but rather a set of explanations for which there is 95 % confidence that at least one explanation associated with the optimal model is included [80,81]. The interpretation is that the larger the set of possible explanations generated by conformal prediction, the less "certain" the model becomes, indicating its uncertainty and suggesting that the user should acquire more information (such as additional domain-specific features, tests, or expert knowledge) to consider potential alternatives reliably. Another approach involves systems that provide examples from the training data similar to the analyzed input. This method, often called 'example-based explanations' or 'prototype selection', helps users understand model decisions by showing real instances that influenced the model's prediction. The number and diversity of examples provided can indicate the level of uncertainty: if it provides only a few cases, the uncertainty is low, and the support is strong; if it provides many diverse cases, the uncertainty is high, and the support is the best effort [82].

Concept-based XAI methods aim to explain model decisions in terms of high-level, human-understandable concepts rather than raw features, bridging the gap between low-level model inputs and the abstract concepts humans use to reason about problems. However, their development is still in its early stages, with most work focused on classification and regression problems [83]. Moreover, research examining the use of concept-based explanations in real-world settings as synthesizing meaningful concepts to human users presents significant challenges is scarce. While some concepts are universally understood, others are more subjective and can vary depending on the stakeholders involved, cultural context, and domain-specific knowledge related to the training data [84].

The trade-off between model performance and interpretability presents another challenge for XAI in DL [85,86]. In many cases, the most accurate and robust models are those that are most complex and opaque and, hence difficult to explain. Therefore, finding a balance between model performance and its interpretability is crucial for the successful application of XAI methods in real-world applications. For these reasons, there is a pressing need for an integrated approach that combines XAI and UQ techniques to address the challenges of interpretability and trust in DL models.

4. Integrating UQ in XAI methods: Potential solutions

Integrating UQ with XAI techniques enhances DL model transparency and trustworthiness through quantitative uncertainty measures. By quantifying the uncertainty associated with XAI explanations, users can better understand the model's decision-making and assess the credibility of the explanations [85]. One promising approach to integrate UQ with XAI is to estimate the uncertainty of the extracted features or saliency maps. This can provide additional insights into both the robustness of the model's predictions and the reliability of the salient features identified by the XAI methods. Another potential avenue is to incorporate uncertainty information directly into the XAI visualizations. This can help prevent over-reliance on the explanations and promote a more cautious and nuanced interpretation of the model's outputs [5].

In the following subsections, we explore specific examples of how UQ can be integrated with XAI methods, focusing on aleatoric and epistemic uncertainty estimation. We also present illustrative case studies to demonstrate the practical implementation and benefits of this integrated approach.

4.1. Aleatoric uncertainty in XAI methods

Estimating the aleatoric uncertainty of XAI methods, like saliency maps, provides valuable insights into prediction robustness and feature

reliability during inference. One potential application is to assess the reproducibility of XAI tools. Ideally, if the model's predictions are reproducible, it can be considered reliable, and the location of the most relevant features should remain consistent even under large perturbations of the input data.

This idea is particularly relevant to healthcare applications, where aleatoric uncertainty is represented by inherent noise in medical imaging, such as acquisition parameters, reconstruction methods, or image quality. Data augmentation techniques can estimate the uncertainty in feature attribution when using XAI methods in these situations. For example, we can observe how the explanation (e.g., saliency maps) changes across these perturbations by creating several versions of the input image with slight rotations or different lighting conditions. A measure of aleatoric uncertainty in the explanation itself is provided by this variation. Fig. 2 illustrates the concept of integrating aleatoric uncertainty in XAI frameworks. In this approach, the saliency maps obtained from multiple augmented samples are compared to estimate the uncertainty of the salient features.

In our recent work [87], we employed a similar approach and proposed a quantitative indicator called Spatial Uncertainty Estimation (SUE) to evaluate the repeatability of XAI tools during test-time augmentation. The study focused on the binary classification of coronary artery disease (CAD) using an ECG dataset [88] comprising 95,300 segments (80,000 healthy and 15,300 pathological segments). SUE measures the spatial consistency of the salient features across multiple augmented samples, providing a quantitative assessment of the reliability of the XAI methods. A higher SUE value indicated high repeatability of the XAI method, while a lower SUE value suggested potential instability or uncertainty in the identified salient features.

Using a CNN-BiLSTM architecture, we achieved excellent classification performance with 99.6 % accuracy, 99.8 % sensitivity, and 98.2 % specificity on the test set. To validate the SUE metric, we performed test-time augmentation using four different types of perturbations: Gaussian noise, power line interference, motion artifacts, and electrode motion artifacts. Our experiments showed a strong correlation between SUE and prediction correctness. Table 1 summarizes SUE values for correctly and misclassified samples under these perturbations, demonstrating that correct classifications generally have higher SUE values than misclassifications.

This approach not only evaluates the spatial location of the most relevant features but also provides quantitative information about their repeatability, enabling users to assess the reliability of the model's predictions. Moreover, this approach can be extended to medical image analysis tasks, where the spatial consistency of salient features is crucial for accurate diagnosis, prognosis and treatment planning.

Table 1
SUE values for correctly classified and misclassified ECG data under different perturbations in the test set (SUE: Spatial Uncertainty Estimation) from [87].

Type of perturbation	SUE on correctly classified instances	SUE on misclassified instances
Gaussian noise	0.982 ± 0.040	0.865 ± 0.277
Power line interference	0.967 ± 0.054	0.815 ± 0.290
Motion artifact	0.846 ± 0.106	0.463 ± 0.288
Electrode motion	0.834 ± 0.107	0.476 ± 0.294

4.2. Epistemic uncertainty in XAI methods

Quantifying epistemic uncertainty of extracted features offers insights into prediction reliability and explanation consistency. Reliable predictions should show minimal variation in the spatial location of features. Fig. 3 illustrates an example of estimating the epistemic uncertainty of XAI methods in image processing.

Epistemic uncertainty arises from a lack of knowledge about the model or the data. When a model encounters new, unseen data that differs from its training set, the epistemic uncertainty can increase, indicating that the model is less confident in its predictions. This can help identify areas where the model might be biased. For example, if a model trained primarily on data from one demographic group shows high epistemic uncertainty when applied to a different demographic, this suggests that the model may be biased toward the characteristics of the original group. In such cases, we propose using ensemble methods or Monte Carlo dropout not only to quantify prediction uncertainty but also to assess the stability of explanations. By generating multiple explanations through different model configurations, we can identify which features consistently contribute to the prediction and which ones show high variability, indicating areas of model uncertainty.

This approach is especially crucial in healthcare because understanding the lack of model's confidence can prevent potentially harmful decisions. For instance, in diagnostic imaging, regions of an image that show high variability in feature attribution across ensemble members may indicate areas where the model's reasoning is less reliable due to limited training data. For example, we considered the SARTAJ dataset [89], which comprises 3,264 brain MRI images categorized into four classes (no tumor, pituitary tumor, meningioma tumor, and glioma

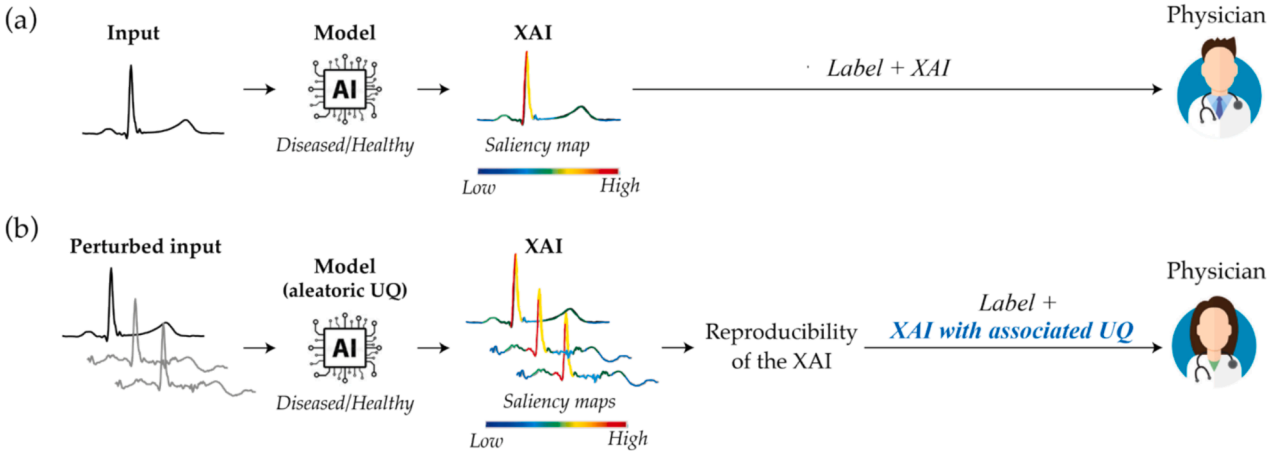


Fig. 2. Integrating aleatoric uncertainty into XAI frameworks. (a) In the traditional pipeline, the XAI method (saliency map) is presented to the user along with the model prediction. (b) By comparing saliency maps generated from multiple augmented samples, the consistency and spatial uncertainty of the most relevant features can be quantified. This provides insights into both the reliability of the prediction and the robustness of the explanation.

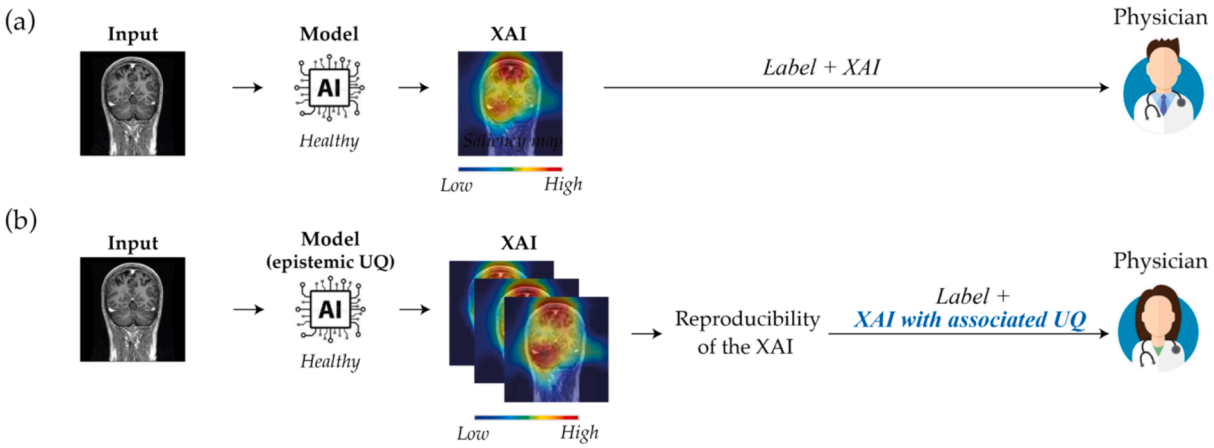


Fig. 3. Integrating epistemic uncertainty into XAI frameworks. (a) The standard approach consists of a single saliency map generated from model inference. (b) By performing multiple stochastic inferences, it is possible to evaluate the consistency of the XAI method's explanations, thereby assessing the reliability of the generated explanations.

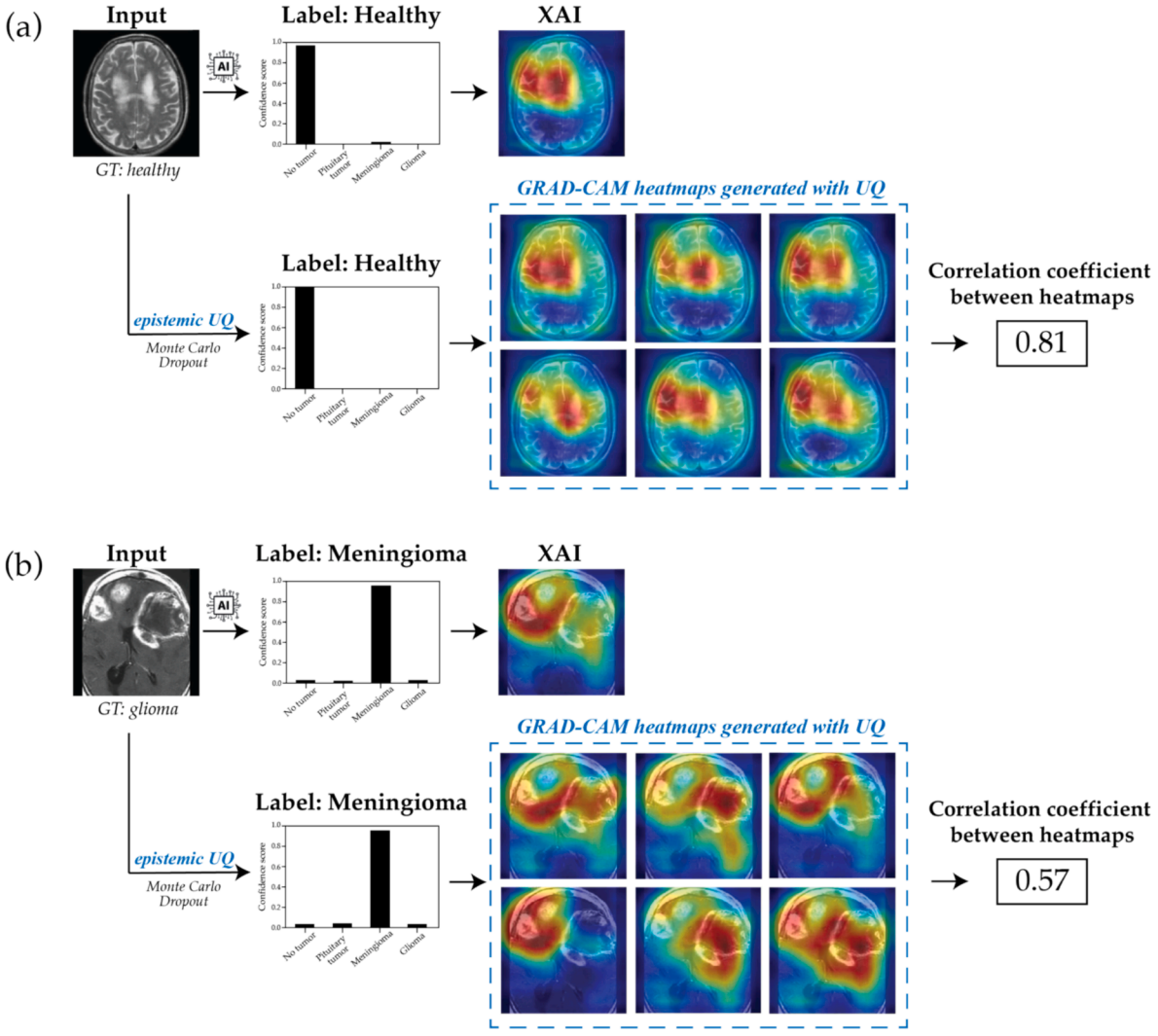


Fig. 4. Quantitative analysis of epistemic uncertainty in feature attribution using correlation between saliency maps. (a) Example of correct classification showing high repeatability in Grad-CAM heatmaps generated using MCD. A high average Spearman coefficient correlation between heatmaps indicates a low epistemic uncertainty and reliable feature attribution. (b) Example of misclassified sample demonstrating significant variability in Grad-CAM heatmaps, with lower Spearman correlation coefficients suggesting a high epistemic uncertainty in feature localization.

tumor). The dataset was divided into training (2,437 images, 74.7 %), validation (433 images, 13.3 %), and test sets (394 images, 12 %). We then trained a simple classification network (ResNet50) and performed inference on the test set. For the MC dropout implementation, we added dropout layers (rate = 0.5) after each residual block in the ResNet50 architecture and performed 20 stochastic forward passes during inference. This setup enabled us to generate multiple Grad-CAM heatmaps for each test image, providing a robust framework for uncertainty assessment. Incorporating uncertainty into the Grad-CAM heatmaps can lead to the quantification of the relevant regions that are certain versus those that are uncertain. Evaluating saliency can offer users more robust and reliable explanations, enhancing trust in the model. A cautious AI approach could involve the model refraining from predictions if the feature repeatability score is low, regardless of the confidence score.

To quantitatively assess the uncertainty in feature attribution, we computed the average Spearman correlation coefficient between saliency maps generated through MC dropout and those obtained from traditional inference. This correlation analysis measures spatial consistency in feature importance across different model configurations. High correlation values can indicate strong spatial reproducibility of salient regions, suggesting low epistemic uncertainty in feature localization. Conversely, low correlation values can suggest high variability in feature attribution, indicating regions where the model's reasoning may be less reliable. Our analysis revealed that correctly classified samples typically showed higher correlation coefficients (mean $\rho = 0.76 \pm 0.08$) compared to misclassified samples (mean $\rho = 0.54 \pm 0.12$), providing quantitative evidence that epistemic uncertainty in feature attribution can help identify potentially unreliable predictions.

Fig. 4 compares the Grad-CAM heatmaps between a correct and an incorrect prediction of the network, also showing the model's probability scores for the predicted class labels. It can be observed that the Grad-CAM heatmaps calculated using MCD are highly repeatable in the case of correct classification (Fig. 4a), while they tend to exhibit significant variability for misclassified samples (Fig. 4b). This aspect cannot be inferred from a single Grad-CAM heatmap alone. In some cases, the probability score provides contradictory results, with high confidence observed in both misclassified and correctly classified samples.

This example demonstrates the estimation of epistemic uncertainty using MCD, but it can be extended to other approaches for epistemic uncertainty estimation, such as Bayesian neural networks or ensemble models. Once the heatmaps are extracted, a correlation value can be calculated with respect to the average heatmap or the heatmaps of individual inferences to obtain quantitative information about the repeatability of these XAI tools [90].

5. Discussion

As a position paper, our work aims to bridge the theoretical gap between XAI and UQ methods, presenting a conceptual framework for their integration in healthcare applications. The integration of XAI and UQ methods holds the potential for enhancing the transparency, reliability, and trustworthiness of DL models. This integrated approach enables healthcare providers to better evaluate when to rely on AI predictions: while XAI methods, such as saliency maps and other pixel-attribution methods [77], highlight relevant diagnostic areas, UQ simultaneously provides confidence measures for these interpretations [20]. This dual information helps clinicians make more informed decisions about when to trust model predictions and when additional clinical investigation might be warranted, so as to increase appropriate reliance [91]. Such integration is particularly valuable in complex cases where model uncertainty signals may indicate the need for supplementary tests or expert consultation.

One of the key advantages of this integrated approach is its ability to mitigate potential biases in user interpretation. When XAI methods are used in isolation, there is a risk that users may perceive the explanations as objective and precise, leading to overconfidence in the model's decisions. For instance, saliency maps generated using different techniques or at different layers of a DL model can yield varying results, making it challenging to assess the trustworthiness of the explanations [92,93]. However, by incorporating uncertainty measures, users are encouraged to adopt a more cautious and nuanced interpretation of the model's outputs, acknowledging the limitations and potential variability of the explanations [5]. This holistic approach to explainability and UQ can foster a deeper trust in the model's outputs and facilitate the adoption of DL models in healthcare applications [85]. Fig. 5 depicts our vision of this integrated approach.

The examples provided in this paper, such as estimating the uncertainty of saliency maps using aleatoric and epistemic uncertainty, demonstrate the practical benefits of integrating UQ into DL prediction pipelines. By evaluating the spatial consistency and repeatability of salient features across multiple perturbations or model inferences, users can assess the reliability of the model's predictions and the robustness of the explanations. Moreover, the integration of XAI and UQ methods can be extended beyond saliency maps to other XAI techniques, such as occlusion maps, activation maximization, RISE, and LRP interpretation of learned features.

We envision that clinical practice will be significantly impacted by the use of XAI combined with UQ techniques. This combination can empower clinicians providing a better understanding of the model's predictions. Likewise, by providing uncertainty-aware explanations, XAI-supported decision-making can be more trustworthy. Such explanations can be helpful in diagnostic imaging (radiology and digital pathology) to flag image regions as the potential source of uncertainty and prompt a more careful evaluation. As a result, diagnostic errors can decrease and the effectiveness and reliability of clinical decision-making can increase. Remarkably, the combination of XAI and UQ complies with new legal standards for AI in healthcare, especially those pertaining to reliability and transparency. Our method offers a framework for satisfying regulatory bodies' growing demands for explainable and reliable AI systems while maintaining high-performance standards.

While our study shows promise and benefits of combining XAI and UQ, several limitations should be acknowledged. First, our preliminary analyses focused on saliency maps, which represent one of many possible XAI methods, and future research should investigate the integration of UQ with other XAI techniques, such as LIME, SHAP, and counterfactual explanations. Furthermore, when integrating UQ with specific XAI methods, it's important to recognize that uncertainty can only be quantified for the aspects the XAI method actually explains. For instance, with saliency maps, while UQ can assess uncertainty in feature localization, low localization uncertainty does not guarantee correct predictions, as the model might consistently identify but misinterpret relevant features. Second, while we primarily addressed classification tasks due to their unique explainability challenges, the framework could be extended to other medical imaging tasks such as detection, segmentation, and image generation. Third, the proposed basic metrics (SUE and Spearman correlation) to quantify XAI uncertainty may not generalize across all applications and XAI and UQ combinations. Given the diversity of healthcare applications, developing more standardized and comprehensive evaluation metrics remains a significant challenge. Additionally, while we present separate pipelines for the electrocardiogram and magnetic resonance applications, modern healthcare AI systems increasingly rely on multimodal data integration [94]. Future research should explore how uncertainty-aware explanations can be extended to multiple data modalities, as this presents additional challenges for both UQ and explanation generation.

It is important to recognize that quantifying the uncertainty of XAI methods is essential to mitigate potential biases in user interpretation. These tools should be viewed as a collaborative rather than a competitive workflow, highlighting the complementary nature of XAI and UQ as "two sides of the same coin".

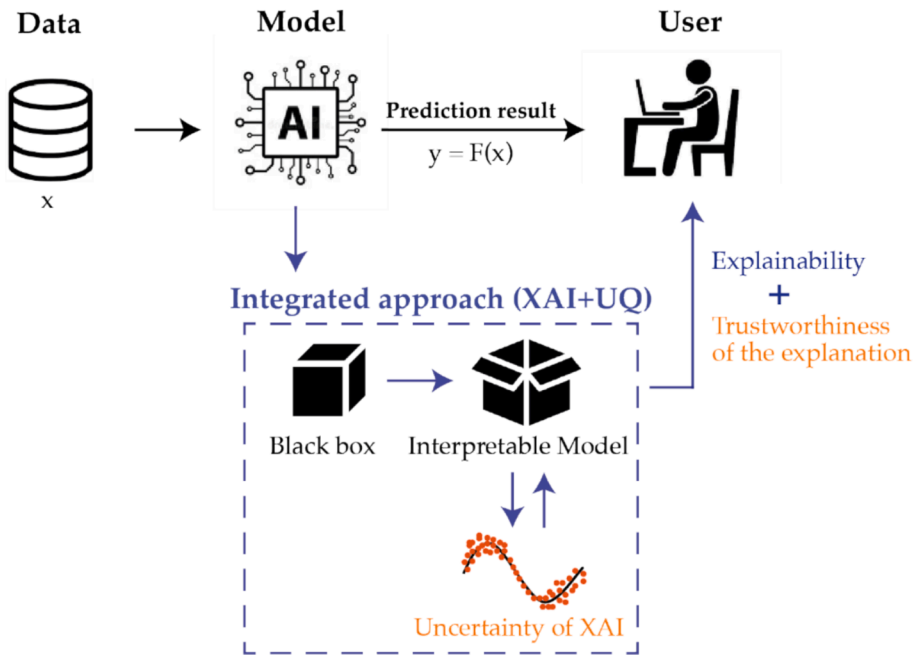


Fig. 5. Our vision of the integrated approach combines XAI and UQ methods.

Funding Acknowledgements

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data statement.

The implementation code for the aleatoric and epistemic uncertainty quantification is available upon reasonable request to the corresponding author.

Summary table.

What is already known:

- Deep Learning (DL) models in healthcare lack transparency and interpretability, raising concerns about reliability and trustworthiness
- Explainable AI (XAI) methods provide insights into model predictions but may not be sufficient to ensure reliability.

What this study adds to knowledge:

- Proposes integration of Uncertainty Quantification (UQ) with XAI to enhance transparency and reliability of DL models in healthcare
- Introduces methods to quantify the uncertainty of XAI explanations

CRediT authorship contribution statement

Massimo Salvi: Writing – original draft, Visualization, Methodology, Investigation, Conceptualization. **Silvia Seoni:** Writing – review & editing, Visualization, Software. **Andrea Campagner:** Writing – review & editing, Investigation. **Arkadiusz Gertych:** Writing – review & editing. **U.Rajendra Acharya:** Writing – review & editing. **Filippo Molinari:** Writing – review & editing. **Federico Cabitza:** Writing – review & editing, Supervision, Investigation.

Declaration of competing interest

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

References

[1] M. Frasca, D. La Torre, G. Pravettoni, I. Cutica, Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review, Discover Artificial Intelligence 4 (2024) 15.
[2] R. Ha, C. Chin, J. Karcich, M.Z. Liu, P. Chang, S. Mutasa, E. Pascual Van Sant, R. T. Wynn, E. Connolly, S. Jambawalikar, Prior to initiation of chemotherapy, can we predict breast tumor response? Deep Learning Convolutional Neural Networks Approach Using a Breast MRI Tumor Dataset, J Digit Imaging 32 (2019) 693–701.
[3] M. Salvi, F. Molinari, N. Dogliani, M. Bosco, Automatic discrimination of neoplastic epithelium and stromal response in breast carcinoma, Comput Biol Med 110 (2019) 8–14.
[4] K. Karthik, M. Mahadevappa, Deep learning with adaptive convolutions for classification of retinal diseases via optical coherence tomography, Image vis Comput 146 (2024) 105044.
[5] S.N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J.H. Chen, X. Liu, Z. He, Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review, J. Am. Med. Inform. Assoc. 27 (2020) 1173–1185.
[6] A. Narla, B. Kuprel, K. Sarin, R. Novoa, J. Ko, Automated classification of skin lesions: from pixels to practice, J, Invest. Dermatol. 138 (2018) 2108–2110.
[7] J.R. Zech, M.A. Badgeley, M. Liu, A.B. Costa, J.J. Titano, E.K. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study, PLoS Med 15 (2018) e1002683.
[8] H.W. Loh, C.P. Ooi, S. Seoni, P.D. Barua, F. Molinari, U.R. Acharya, Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022), Comput Methods Programs Biomed 226 (2022) 107161.
[9] L. Chazette, K. Schneider, Explainability as a non-functional requirement: challenges and recommendations, Requir Eng 25 (2020) 493–514.
[10] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif Intell 267 (2019) 1–38.
[11] S. Bruckert, B. Finzel, U. Schmid, The next generation of medical decision support: A roadmap toward transparent expert companions, Front Artif Intell 3 (2020) 507973.
[12] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip Rev Data Min Knowl Discov 9 (2019) e1312.
[13] M. Rigotti, C. Miksovic, I. Giurgiu, T. Gschwind, P. Scotton, Attention-based interpretability with concept transformers, in: International Conference on Learning Representations, 2021.
[14] C. Molnar, G. Casalicchio, B. Bischl, Interpretable machine learning–a brief history, state-of-the-art and challenges, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2020, pp. 417–431.
[15] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods, Mach Learn 110 (2021) 457–506.
[16] S. Seoni, V. Jahmunah, M. Salvi, P.D. Barua, F. Molinari, U.R. Acharya, Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023), Comput Biol Med 107441 (2023).

[17] J. Amann, A. Blasimme, E. Vayena, D. Frey, V.I. Madai, P. Consortium, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC Med Inform Decis Mak 20 (2020) 1–9.
[18] M. Chua, D. Kim, J. Choi, N.G. Lee, V. Deshpande, J. Schwab, M.H. Lev, R. G. Gonzalez, M.S. Gee, S. Do, Tackling prediction uncertainty in machine learning for healthcare, Nat Biomed Eng 7 (2023) 711–718.
[19] C. Marx, Y. Park, H. Hasson, Y. Wang, S. Ermon, L. Huan, But are you sure? an uncertainty-aware perspective on explainable ai, in: In: International Conference on Artificial Intelligence and Statistics, 2023, pp. 7375–7391.
[20] D. Slack, A. Hilgard, S. Singh, H. Lakkaraju, Reliable post hoc explanations: Modeling uncertainty in explainability, Adv Neural Inf Process Syst 34 (2021) 9391–9404.
[21] Y.-H. Hung, C.-Y. Lee, BMB-LIME: LIME with modeling local nonlinearity and uncertainty in explainability, Knowl Based Syst 294 (2024) 111732.
[22] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Inf. Fusion 76 (2021) 243–297.
[23] F.C. Maruccio, W. Eppinga, M.-H. Laves, R.F. Navarro, M. Salvi, F. Molinari, P. Papaconstadopoulos, Clinical assessment of deep learning-based uncertainty maps in lung cancer segmentation, Phys Med Biol 69 (2024) 035007.
[24] R. Ranjbarzadeh, A. Caputo, E.B. Tirkolaee, S.J. Ghoushchi, M. Bendechache, Brain tumor segmentation of MRI images: A comprehensive review on the application of artificial intelligence tools, Comput Biol Med 152 (2023) 106405.
[25] A.W. Anderson, M.L. Marinovich, N. Houssami, K.P. Lowry, J.G. Elmore, D.S. M. Buist, S. Hofvind, C.I. Lee, Independent external validation of artificial intelligence algorithms for automated interpretation of screening mammography: a systematic review, J. Am. Coll. Radiol. 19 (2022) 259–273.
[26] I.-S. Tzeng, P.-C. Hsieh, W.-L. Su, T.-H. Hsieh, S.-C. Chang, Artificial Intelligence-assisted chest x-ray for the diagnosis of COVID-19: a systematic review and meta-analysis, Diagnostics 13 (2023) 584.
[27] C. Barboi, A. Tzavelis, L.N. Muhammad, Comparison of severity of illness scores and artificial intelligence models that are predictive of intensive care unit mortality: meta-analysis and review of the literature, JMIR Med Inform 10 (2022) e35293.
[28] N.-I. Yang, C.-H. Yeh, T.-H. Tsai, Y.-J. Chou, P.-W.-C. Hsu, C.-H. Li, Y.-H. Chan, L.-T. Kuo, C.-T. Mao, Y.-C. Shyu, Artificial intelligence-assisted identification of genetic factors predisposing high-risk individuals to asymptomatic heart failure, Cells 10 (2021) 2430.
[29] S.P. Pattyam, AI in Data Science for Healthcare: Advanced Techniques for Disease Prediction, Treatment Optimization, and Patient Management, Distributed Learning and Broad Applications in Scientific Research 5 (2019) 417–455.
[30] E. Hossain, R. Rana, N. Higgins, J. Soar, P.D. Barua, A.R. Pisani, K. Turner, Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review, Comput Biol Med 155 (2023) 106649.
[31] K. Bera, K.A. Schalper, D.L. Rimm, V. Velcheti, A. Madabhushi, Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology, Nat Rev Clin Oncol 16 (2019) 703–715.
[32] A. Madabhushi, M.D. Feldman, P. Leo, Deep-learning approaches for Gleason grading of prostate biopsies, Lancet Oncol 21 (2020) 187–189.
[33] W. Swartout, C. Paris, J. Moore, Explanations in knowledge systems: Design for explainable expert systems, IEEE Expert 6 (1991) 58–64.
[34] A. Gertych, O. Faust, AI Explainability and Bias Propagation in Medical Decision Support, Comput Methods Programs Biomed 108465 (2024).
[35] T. Speith, A review of taxonomies of explainable artificial intelligence (XAI) methods, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022: pp. 2239–2250.
[36] A. Altmann, L. Tološi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (2010) 1340–1347.
[37] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (xai): A survey, ArXiv Preprint ArXiv:2006.11371 (2020).
[38] M.T. Ribeiro, S. Singh, C. Guestrin, " Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: pp. 1135–1144.
[39] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv Neural Inf Process Syst 30 (2017).
[40] M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Towards better understanding of gradient-based attribution methods for deep neural networks, ArXiv Preprint ArXiv:1711.06104 (2017).
[41] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS One 10 (2015) e0130140.
[42] V. Petsiuk, A. Das, K. Saenko, Rise, Randomized input sampling for explanation of black-box models, ArXiv Preprint ArXiv:1806.07421 (2018).
[43] D. Garreau, U. Luxburg, Explaining the explainer: A first theoretical analysis of LIME, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020: pp. 1287–1296.
[44] L. Rokach, Decision forest: Twenty years of research, Inf. Fusion 27 (2016) 111–125.
[45] B. Ustun, C. Rudin, Methods and models for interpretable linear classification, ArXiv Preprint ArXiv:1405.4047 (2014).
[46] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat Mach Intell 1 (2019) 206–215.
[47] J. Fürnkranz, T. Kliegr, H. Paulheim, On cognitive preferences and the plausibility of rule-based models, Mach Learn 109 (2020) 853–898.
[48] I. Stepin, J.M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001.
[49] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, J. Jorge, Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications, Inf. Fusion 81 (2022) 59–83.
[50] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Min Knowl Discov (2022) 1–55.
[51] L. Celar, R.M.J. Byrne, How people reason with counterfactual and causal explanations for artificial intelligence decisions in familiar and unfamiliar domains, Mem Cognit 51 (2023) 1481–1496.
[52] A. Wijekoon, N. Wiratunga, I. Nkisi-Orji, C. Palihawadana, D. Corsar, K. Martin, How close is too close? The role of feature attributions in discovering counterfactual explanations, in: International Conference on Case-Based Reasoning, Springer, 2022: pp. 33–47.
[53] R. Kommiya Mothilal, D. Mahajan, C. Tan, A. Sharma, Towards unifying feature attribution and counterfactual explanations: Different means to the same end, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021: pp. 652–663.
[54] S.Ö. Arik, T. Pfister, Tabnet: Attentive interpretable tabular learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021: pp. 6679–6687.
[55] K. Čyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: a survey, ArXiv Preprint ArXiv:2105.11266 (2021).
[56] K. Baum, H. Hermanns, T. Speith, Towards a framework combining machine ethics and machine explainability, ArXiv Preprint ArXiv:1901.00590 (2019).
[57] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), in: International Conference on Machine Learning, PMLR, 2018: pp. 2668–2677.
[58] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J.K. Su, This looks like that: deep learning for interpretable image recognition, Adv Neural Inf Process Syst 32 (2019).
[59] J. Li, A. Dada, B. Puladi, J. Kleesiek, J. Egger, ChatGPT in healthcare: a taxonomy and systematic review, Comput Methods Programs Biomed 108013 (2024).
[60] P. Hager, F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren, G. Kaissis, Evaluation and mitigation of the limitations of large language models in clinical decision-making, Nat Med 30 (2024) 2613–2622.
[61] L.M. Amugongo, P. Mascheroni, S.G. Brooks, S. Doering, J. Seidel, Retrieval Augmented Generation for Large Language Models in Healthcare, A Systematic Review (2024).
[62] A. Campagner, L. Famiglini, A. Carobene, F. Cabitza, Everything is varied: The surprising impact of instantial variation on ML reliability, Appl Soft Comput 146 (2023) 110644.
[63] F. Cabitza, A. Campagner, G. Malgieri, C. Natali, D. Schneeberger, K. Stoeger, A. Holzinger, Quod erat demonstrandum?-Towards a typology of the concept of explanation for the design of explainable AI, Expert Syst Appl 213 (2023) 118888.
[64] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks, Neurocomputing 338 (2019) 34–45.
[65] Z. Swiderska-Chadaj, Z. Ma, N. Ing, T. Markiewicz, M. Lorent, S. Cierniak, A.E. Walts, B.S. Knudsen, A. Gertych, Contextual classification of tumor growth patterns in digital histology slides, in: Information Technology in Biomedicine, Springer, 2019: pp. 13–25.
[66] A. Amini, W. Schwarting, A. Soleimany, D. Rus, Deep evidential regression, Adv Neural Inf Process Syst 33 (2020) 14927–14937.
[67] M. Valdenegro-Toro, D.S. Mori, A deeper look into aleatoric and epistemic uncertainty disentanglement, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2022: pp. 1508–1516.
[68] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Adv Neural Inf Process Syst 30 (2017).
[69] W. Zhang, Z.M. Ma, S. Das, T.-W.-L. Weng, A. Megretski, L. Daniel, L.M. Nguyen, One step closer to unbiased aleatoric uncertainty estimation, in: In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 16857–16864.
[70] Y. Kwon, J.-H. Won, B.J. Kim, M.C. Paik, Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation, Comput Stat Data Anal 142 (2020) 106816.
[71] V. Vovk, A. Gammerman, G. Shafer, Algorithmic learning in a random world, Springer, 2005.
[72] E. Begoli, T. Bhattacharya, D. Kusnezov, The need for uncertainty quantification in machine-assisted medical decision making, Nat Mach Intell 1 (2019) 20–23.
[73] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: the system causability scale (SCS) comparing human and machine explanations, KI-Künstliche Intelligenz 34 (2020) 193–198.
[74] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, Adv Neural Inf Process Syst 31 (2018).
[75] A.S. Sambyal, U. Niyaz, N.C. Krishnan, D.R. Bathula, Understanding calibration of deep neural networks for medical image classification, Comput Methods Programs Biomed 242 (2023) 107816.
[76] F. Cabitza, A. Campagner, L. Ronzio, M. Cameli, G.E. Mandoli, M.C. Pastore, L. M. Sconfienza, D. Folgado, M. Barandas, H. Gamboa, Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis, Artif Intell Med 138 (2023) 102506.

[77] R.Y.G. Paccotacya-Yanque, A. Bissoto, S. Avila, Are Explanations Helpful? A Comparative Analysis of Explainability Methods in Skin Lesion Classifiers, in: 2024 20th International Symposium on Medical Information Processing and Analysis (SIPAIM), IEEE, 2024: pp. 1–5.
[78] C. Natali, L. Famiglini, A. Campagner, G.A. La Maida, E. Gallazzi, F. Cabitza, Color shadows 2: Assessing the impact of xai on diagnostic decision-making, in: World Conference on Explainable Artificial Intelligence, Springer, 2023: pp. 618–629.
[79] P. Altmeyer, M. Farmanbar, A. van Deursen, C.C.S. Liem, Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024: pp. 10829–10837.
[80] A.N. Angelopoulos, S. Bates, Conformal prediction: A gentle introduction, Foundations and Trends®, Mach. Learn. 16 (2023) 494–591.
[81] G. Shafer, V. Vovk, A tutorial on conformal prediction, J. Mach. Learn. Res. 9 (2008).
[82] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: In: International Conference on Machine Learning, 2016, pp. 1050–1059.
[83] A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, Explainability in deep reinforcement learning, Knowl Based Syst 214 (2021) 106685.
[84] C. Meske, E. Bunde, J. Schneider, M. Gersch, Explainable artificial intelligence: objectives, stakeholders, and future research opportunities, Inf. Syst. Manag. 39 (2022) 53–63.
[85] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions, Inf. Fusion (2024) 102301.
[86] A.L. Alfeo, A.G. Zippo, V. Catrambone, M.G.C.A. Cimino, N. Toschi, G. Valenza, From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks, Comput Methods Programs Biomed 236 (2023) 107550.
[87] S. Seoni, F. Molinari, U.R. Acharya, O.S. Lih, P.D. Barua, S. García, M. Salvi, Application of spatial uncertainty predictor in CNN-BiLSTM model using coronary artery disease ECG signals, Inf Sci (n y) (2024) 120383.
[88] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J. E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (2000) e215–e220.
[89] Brain tumor classification (MRI). [Online]. Available: https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor- classification-mri, (n.d.).
[90] J. González-Abad, J. Baño-Medina, J.M. Gutiérrez, Using explainability to inform statistical downscaling based on deep learning beyond standard validation approaches, J Adv Model Earth Syst 15 (2023) e2023MS003641.
[91] F. Cabitza, A. Campagner, R. Angius, C. Natali, C. Reverberi, AI shall have no dominion: on how to measure technology dominance in AI-supported human decision-making, in: In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–20.
[92] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: In: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[93] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization, ArXiv Preprint ArXiv:1506.06579 (2015).
[94] M. Salvi, H.W. Loh, S. Seoni, P.D. Barua, S. García, F. Molinari, U.R. Acharya, Multi-modality approaches for medical support systems: A systematic review of the last decade, Inf. Fusion 102134 (2023).