Enhancing Trustworthiness in AI Models for High-Stakes Medical Imaging through Precise Uncertainty Conveyance

1. Introduction: The Critical Need for Trustworthy AI in High-Stakes Medicine

Artificial intelligence (AI) holds immense promise for revolutionizing medical image analysis, offering the potential to assist clinicians in identifying subtle details within complex images that might escape human perception [User Query]. By augmenting the diagnostic process with AI, healthcare professionals may experience increased efficiency and accuracy in their assessments [User Query]. AI's capability to process vast quantities of medical data and discern intricate patterns within seconds presents a significant advantage in the timely detection of diseases.[1] Furthermore, AI can automate routine tasks, thereby freeing up valuable time for clinicians to dedicate more attention to direct patient care and complex cases.[2] The capacity of AI to analyze large datasets and identify patterns offers substantial benefits in medical imaging, potentially leading to earlier and more precise diagnoses. This is particularly crucial as the sheer volume of medical images and the complexity of certain medical conditions can strain human cognitive abilities. AI's computational power can overcome these limitations, but only if its output is reliable and commands the trust of medical professionals.

Despite the potential benefits, the integration of AI in medical image analysis faces inherent challenges, most notably the ambiguity often present in medical images and the consequent need for reliable methods to quantify the uncertainty associated with AI predictions.[4] For instance, in a chest X-ray, distinguishing between pleural effusion, an abnormal fluid buildup in the lungs, and pulmonary infiltrates, which are accumulations of pus or blood, can be exceptionally difficult [User Query]. Medical image segmentation, the process of delineating areas of interest in an image, is frequently not a straightforward task, with even experienced human annotators sometimes exhibiting disagreement on the interpretation of image features.[4] Uncertainty is an intrinsic aspect of medical practice, stemming from factors such as incomplete knowledge about a patient's condition, the inherent limitations of individual physicians, and the imperfect predictive power of even the most advanced diagnostic tools.[6] The inherent ambiguity in medical images necessitates that AI systems not only generate predictions but also accurately reflect their uncertainty regarding those predictions. Given the high stakes involved in medical diagnoses, an AI that presents an incorrect diagnosis with high confidence could have severe ramifications for patients. Therefore, the ability of an AI to indicate when it is unsure is

as critical as its ability to make confident predictions.

Given the complexities and potential ambiguities in medical imaging, clinicians would ideally want to consider a range of possible diagnoses rather than relying on a single AI prediction.[5] While AI models often produce a probability score alongside each prediction to indicate the model's confidence, the reliability of these predicted probabilities has been called into question by numerous studies.[5] Deep learning models, in particular, have a tendency to generate predictions with unwarranted high confidence, which can be particularly problematic in healthcare settings.[10] Therefore, relying solely on a single AI prediction, especially when accompanied by potentially inaccurate confidence scores, poses a significant risk in high-stakes medical scenarios. A more responsible approach involves providing a set of the most likely diagnoses, coupled with a clear indication of the associated uncertainty. This acknowledges the inherent ambiguity and encourages a more thorough and considered evaluation by the clinician. A single, seemingly definitive prediction might inadvertently lead clinicians to overlook alternative possibilities, potentially resulting in delays or misdirection in treatment. Presenting a set of options recognizes the inherent uncertainty and promotes a more comprehensive diagnostic process.

2. The Challenge of Uncertainty in Medical Image Analysis

Clinicians often encounter significant difficulties when interpreting medical images, a process that demands careful analysis to differentiate between conditions that may exhibit similar visual characteristics [User Query]. For example, the subtle visual distinctions between pleural effusion and pulmonary infiltrates in a chest X-ray necessitate a high level of expertise and attention to detail [User Query]. Uncertainty in image interpretation can also arise from the indistinct nature of lesion boundaries, which may be affected by partial volume effects and unclear margins, especially in complex medical conditions like liver cirrhosis.[11] Furthermore, the variety of imaging modalities employed in medicine, such as computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound, each present their own unique challenges that contribute to diagnostic uncertainty. These challenges include motion artifacts in MRI, standardization issues in ultrasound, and the need for radiation dose optimization in CT.[11] The inherent complexity of medical images, coupled with variations in interpretation approaches among clinicians, the subjective nature of some diagnostic criteria, and the potential for human error, all contribute to the overall uncertainty in medical diagnosis.[10] The multifaceted nature of medical imaging, involving subtle visual cues, diverse modalities, and the potential for variability in human interpretation, underscores the complexity of achieving accurate diagnoses and the inherent uncertainty associated with this process. Unlike well-defined object

recognition tasks in other domains, medical images often contain nuanced information and can be influenced by various factors related to image acquisition techniques and the specific physiological conditions of individual patients.

Given the inherent uncertainty in medical data, it becomes paramount for AI models designed to assist in medical image analysis to accurately represent their level of confidence in their predictions.[12] Uncertainty quantification (UQ) plays a crucial role in evaluating the reliability of AI predictions, providing a measure of the trustworthiness of the model's output.[9] In high-stakes medical scenarios, particularly when dealing with the diagnosis of life-threatening conditions, the reliability of AI-generated predictions is of utmost importance as clinical decisions directly impact patient outcomes.[12] Understanding the uncertainty associated with AI predictions empowers decision-makers to better gauge their confidence in the model's output and to account for potential variability in the input data.[12] The proper quantification of uncertainty in medical AI systems provides valuable information that can contribute to more accurate and reliable diagnoses.[14] In medical contexts, where errors can have severe consequences, it is essential for AI models to provide a clear indication of their uncertainty, allowing clinicians to make well-informed decisions and appropriately weigh the AI's contribution to the diagnostic process. A highly confident but ultimately incorrect AI diagnosis could lead to inappropriate or delayed treatment plans. Conversely, knowing when an AI model is uncertain enables clinicians to exercise greater caution and potentially seek additional information or expert consultation.

Traditional AI models often provide a probability score as an indicator of their confidence; however, the reliability of these scores has been increasingly questioned.[5] Extensive prior research has demonstrated that the predicted probabilities generated by standard deep learning models can be inaccurate and may not truly reflect the underlying uncertainty.[5] The softmax output, commonly used in classification networks to represent probabilities, is primarily designed to capture aleatoric uncertainty, which is the inherent noise and randomness within the data itself, but it often fails to adequately represent epistemic uncertainty, which reflects the model's own lack of knowledge or confidence in its parameters.[13] Furthermore, conventional convolutional neural networks (CNNs) with deterministic parameters lack the inherent capability to indicate the level of uncertainty associated with their predictions.[10] Therefore, the traditional probability scores provided by AI models often lack robust statistical validity and may not accurately convey the true level of uncertainty. This necessitates the exploration and adoption of more sophisticated and reliable uncertainty quantification techniques in medical AI applications.

3. Background: Understanding Conformal Classification and Test-Time Augmentation

Conformal Classification:

Conformal classification represents a paradigm shift in how AI models convey their predictions, moving beyond the provision of a single, potentially uncertain answer to offering a set of plausible diagnoses for a given medical image.[5] This set of possibilities is accompanied by a statistical guarantee, indicating a high probability that the correct diagnosis is indeed included within the generated set.[5] Instead of the AI asserting, "I am 90% sure it's condition X," conformal classification conveys, "I am 90% confident that the correct diagnosis is one of these few conditions." For classification tasks, the output of a conformal classifier is not a singular class label but rather a set of potential class labels that the input image might belong to.[18]

A significant advantage of conformal classification lies in its ability to provide prediction guarantees that are statistically more robust compared to the standard probability scores produced by many AI models.[16] Conformal prediction generates statistically valid prediction regions, which can take the form of prediction intervals for regression tasks or prediction sets for classification tasks.[18] This guarantee of coverage holds true regardless of the specific underlying machine learning model that is employed.[16] By offering a measure of confidence in the entire set of predictions rather than just a single prediction, conformal prediction enhances the interpretability and trustworthiness of the model's output, particularly in situations where the input data might be ambiguous or when the model encounters data that is significantly different from what it was trained on.[20]

However, a notable limitation of conformal classification is its potential to produce prediction sets that are impractically large, thereby diminishing its utility for clinicians in real-world scenarios.[5] In some instances, to maintain the statistical guarantee of including the correct diagnosis, a conformal classifier might output a substantial number of possible conditions, even if many of them are highly unlikely.[5] For example, when classifying an image of an animal into one of thousands of potential species, a conformal classifier might generate a prediction set containing hundreds of possibilities to ensure the true species is within that set [User Query]. Such extensive prediction sets can make it challenging for clinicians to efficiently narrow down to the most probable diagnosis and can hinder the streamlining of the treatment process [User Query]. While the guarantee provided by conformal classification is valuable, the practical application is compromised when the resulting sets of predictions are too broad to offer meaningful guidance. This necessitates the development of

methods to refine the prediction sets generated by conformal classifiers, making them smaller and more informative without compromising their statistical validity.

Test-Time Augmentation (TTA):

Test-time augmentation (TTA) is a technique employed in machine learning, particularly in computer vision, to enhance the reliability and accuracy of predictions made by trained models.[5] In essence, TTA involves presenting an AI model with multiple slightly modified versions of the same input image during the prediction phase.[5] These modifications, or augmentations, can include transformations such as slight rotations, cropping, horizontal or vertical flipping, and adjustments in zoom level.[5] The AI model then generates a prediction for each of these augmented versions of the original image.[5] Finally, these individual predictions are combined, or aggregated, to produce a more robust and accurate final prediction for the original, unaugmented image.[5]

TTA is a widely adopted strategy in computer vision aimed at improving the accuracy of models during the inference stage.[23] It has been shown to enhance the overall accuracy and robustness of predictions by making the model less sensitive to minor variations or noise in the input data.[25] By exposing the model to slightly different perspectives of the same image, TTA encourages it to focus on the more consistent and diagnostically relevant features, rather than being swayed by minor artifacts or variations in image presentation.[25] The process of aggregating predictions from multiple augmented views often leads to a more reliable and accurate final output compared to a prediction based on a single, unaugmented image.[25]

Several common augmentation techniques are typically used in TTA for image data. These include geometric transformations such as cropping, flipping (horizontally or vertically), zooming in or out, and rotating the image by various angles.[5] Additionally, photometric transformations, which involve adjusting the brightness, contrast, or color balance of the image, can also be employed.[23] It is important to note that these augmentations are designed to be label-preserving, meaning they alter the image in ways that do not change the underlying diagnosis or the class to which the image belongs.[22] The core idea behind TTA is that by presenting multiple slightly different versions of the same image to the AI, the impact of any single image's peculiarities can be reduced, leading to a more reliable and generalized prediction.

4. The Novel Research: Combining Conformal Classification and Test-Time Augmentation

Researchers at the Massachusetts Institute of Technology (MIT) have recently

developed an innovative method that combines the strengths of conformal classification and test-time augmentation to address the challenge of generating trustworthy AI predictions in high-stakes medical settings.[5] Their research introduces a simple yet effective enhancement to conformal classification that can significantly reduce the size of the prediction sets produced by these methods, achieving reductions of up to 30 percent, while simultaneously improving the overall reliability of the predictions.[5] The primary goal of this research was to overcome the practical limitation of impractically large prediction sets that can sometimes arise with conformal classification, particularly in critical applications such as medical image analysis.[25] The findings of this work are scheduled to be presented at the prestigious Conference on Computer Vision and Pattern Recognition (CVPR) in June.[22] The team of researchers behind this advancement includes Divya Shanmugam, Helen Lu, Swami Sankaranarayanan, and the senior author of the paper, Professor John Guttag.[22]

The core of the MIT researchers' method involves applying test-time augmentation (TTA) to the input medical images *before* subjecting them to the conformal classification process.[5] This initial step entails creating multiple augmented versions of each original medical image using a variety of label-preserving transformations, such as cropping different regions of the image, flipping it horizontally or vertically, and applying slight zoom adjustments.[5] Subsequently, the underlying AI model, which is typically a pre-trained computer vision model, is used to generate a prediction for each of these augmented images.[25] This results in a set of predictions for each original medical image, corresponding to its various augmented forms.

To effectively leverage the information from these multiple augmented views, the researchers developed a process to learn how to optimally combine the individual predictions to maximize the accuracy of the underlying AI model.[25] This learning process was conducted on a held-out portion of labeled image data, which is a subset of data that would normally be used in the conformal classification step.[26] By using this held-out data, the researchers could train a mechanism to automatically determine the most effective way to aggregate the predictions obtained from the different augmented versions of the images.[25] This learned aggregation strategy allows for a more nuanced combination of the augmented predictions compared to a simple averaging approach.

Once the TTA-transformed predictions were obtained and optimally aggregated, the researchers then applied the conformal classification method to these aggregated predictions.[26] The rationale behind this approach is that by first improving the accuracy and robustness of the underlying predictions through the application of TTA and a learned aggregation strategy, the subsequent conformal classification step can

then generate a smaller, more focused set of probable diagnoses while still maintaining the crucial guarantee of including the correct diagnosis within that set.[26] As Divya Shanmugam aptly stated, with fewer classes to consider in the prediction set, the results become naturally more informative.[28]

The experimental results of this research demonstrated a significant improvement in the efficiency of conformal classification. Compared to traditional conformal prediction methods, the TTA-augmented approach led to a reduction in the size of the prediction sets ranging from 10 to 30 percent across several standard image classification benchmark datasets.[26] Critically, this substantial reduction in the number of potential diagnoses included in the prediction sets was achieved without any compromise to the probabilistic guarantee that the correct diagnosis would be present within the set.[26] According to Divya Shanmugam, this improvement means that clinicians are not sacrificing accuracy for the sake of obtaining more informative and manageable prediction sets.[25] This innovative combination of TTA with conformal classification effectively addresses the core challenge of large prediction sets, making conformal prediction a more practical and valuable tool for high-stakes medical applications. By refining the quality of the initial predictions through TTA, the subsequent conformal classification step can achieve the same level of confidence with a more concise and relevant set of potential diagnoses.

5. Technical Deep Dive: How the Method Works

Conformal classification operates by first calculating nonconformity scores for a set of previously labeled data.[19] These scores serve as a measure of how dissimilar a new prediction is when compared to the historical data and their corresponding predictions.[20] For a new, unlabeled test data point, these computed nonconformity scores are then utilized to construct a prediction set, which is a set of possible labels or outcomes for that new data point.[19] In the context of classification, conformal classifiers specifically compute and output a p-value for each possible class. This p-value is determined by ranking the nonconformity score of the test object against the distribution of nonconformity scores obtained from the training data.[19] The p-value essentially indicates how well the new test instance conforms to each of the possible classes based on the patterns observed in the training data. If the calculated p-value for a particular class exceeds a pre-defined significance level, which is directly related to the desired level of confidence in the prediction, then that specific class is included in the final prediction set generated by the conformal classifier.[16] This approach allows for a statistically sound way to quantify the uncertainty associated with a prediction, as it is based on the observed consistency of predictions on historical data rather than solely on the internal confidence scores of the underlying

model.

Test-time augmentation (TTA) is implemented by generating multiple augmented versions of each individual image within the test dataset.[5] These augmentations typically involve applying a range of label-preserving transformations to the original image. Common types of augmentations used for images include geometric transformations, such as rotations by various degrees, horizontal and vertical flips, cropping to different portions of the image, and scaling or zooming.[23] Additionally, photometric transformations, which modify the pixel intensities of the image by adjusting brightness, contrast, and other color-related parameters, can also be employed.[23] Once these augmented versions of the test image are created, the pre-trained computer vision model is applied to each of them to obtain a set of predictions.[5] The predictions from all the augmented images corresponding to a single original test image are then combined, or aggregated, to arrive at a final, more robust prediction for that original image.[5] The traditional method for aggregating these predictions often involves calculating a simple average of the probability scores assigned to each class across all the augmented versions of the image.[34] This averaging process tends to smooth out the individual predictions and can lead to a more accurate and stable final prediction.

In their research, the MIT team optimized the TTA process by utilizing a portion of the labeled image data that is typically reserved for the conformal classification step.[25] On this held-out dataset, they trained a mechanism to learn the most effective way to aggregate the predictions obtained from the various augmented versions of the images.[25] This learned aggregation policy is designed to maximize the accuracy of the initial AI model's predictions when applied to the augmented data.[25] By employing a separate dataset to learn the TTA policy and subsequently applying conformal classification to the predictions that have been transformed and aggregated using this learned policy, the researchers ensured that the crucial assumption of exchangeability, which is a fundamental requirement for the coverage guarantee provided by conformal prediction, is preserved.[22] This careful separation of the learning and prediction steps is essential for maintaining the statistical validity of the conformal prediction framework.

The researchers observed that the application of TTA led to an increase in the predicted probability of the true class, even in cases where the initial prediction for that class was relatively low.[22] This "promotion" of the correct class within the probability ranking is a key factor in the reduction of the prediction set size. When the true class is assigned a higher probability after the TTA process, the conformal classifier needs to include fewer other, potentially incorrect, classes in the prediction

set to maintain the desired level of confidence and guarantee the inclusion of the true diagnosis.[22] By making the underlying predictions more accurate and robust through TTA and the learned aggregation strategy, the overall uncertainty associated with the predictions is reduced. This reduction in uncertainty directly translates to the generation of smaller and more reliable prediction sets by the conformal classifier [User Query]. Furthermore, a significant advantage of this method is that it does not require any retraining of the original AI model, making it a computationally efficient and easily adaptable technique that can be implemented on top of existing machine learning systems.[26]

6. Potential Impact and Future Avenues

The development of smaller, more reliable prediction sets through the combination of conformal classification and test-time augmentation holds significant implications for clinicians working in high-stakes medical settings.[28] By providing a more focused and manageable set of potential diagnoses, this method can substantially improve the efficiency with which clinicians can arrive at the correct diagnosis.[28] This increased efficiency can lead to a more streamlined treatment process for patients, potentially resulting in improved health outcomes.[1] Moreover, the provision of a more concise set of possibilities, backed by a statistical confidence guarantee, has the potential to increase clinicians' trust in AI-assisted diagnostic systems.[12] The ability of uncertainty quantification to appropriately calibrate clinicians' trust in AI is crucial for the successful integration of these technologies into clinical workflows.[37] Highlighting medical images or specific regions within those images where the AI model exhibits uncertainty can further assist radiologists and other specialists in focusing their attention on the most critical aspects of a case.[37] Ultimately, more focused and reliable AI predictions, coupled with a clear indication of their uncertainty, can empower clinicians to make faster and more informed decisions, leading to tangible benefits in patient care.

Beyond medical imaging, the principles and techniques employed in this research have the potential for broad applicability across a wide range of classification tasks [User Query]. For instance, the method could be effectively used in applications such as identifying the species of an animal from an image captured in a wildlife park, where considering a set of likely possibilities with a confidence guarantee would be valuable [User Query]. The fundamental challenges of improving uncertainty quantification and providing reliable sets of potential outcomes are not unique to medical imaging and are relevant in any domain where classification tasks carry significant consequences or require careful consideration of multiple options.[33] Therefore, the benefits of this research could extend to numerous fields beyond

healthcare.

Looking ahead, the researchers have identified several promising avenues for future work. One key direction involves validating the effectiveness of their approach in the context of AI models that are designed to classify text data, rather than images.[39] This would broaden the applicability of their method to other important domains where uncertainty quantification is critical. Additionally, the researchers are exploring strategies to reduce the computational resources required for the test-time augmentation process, aiming to further enhance the practicality and efficiency of their method [User Query]. Future research could also delve into identifying the optimal types and combinations of augmentations for specific medical imaging tasks and different AI model architectures.[41] Investigating how this approach impacts different types of uncertainty, such as aleatoric (data-related) versus epistemic (model-related) uncertainty, could provide deeper insights into its effectiveness.[13] Exploring the development of learned augmentation policies that are specific to different classes of medical conditions or types of diseases might also lead to further refinements in the generated prediction sets.[16] Finally, the researchers themselves have highlighted the interesting questions raised by their work regarding the optimal allocation of labeled data between the different post-training steps involved in their method, suggesting this as a significant area for future investigation.[26] These planned future research directions underscore the researchers' commitment to further developing and refining their method to enhance the trustworthiness of AI in healthcare and beyond.

7. Conclusion: Enhancing Trust in AI for Critical Decisions

The widespread adoption of deep learning models in clinical practice has been somewhat limited despite the numerous high-performing solutions reported in the literature, primarily due to concerns surrounding the reliability and interpretability of their predictions.[12] For AI assistants to be effectively deployed in high-stakes tasks such as medical image analysis, ensuring their trustworthiness is of paramount importance.[13] Building confidence among patients, clinicians, and the general public regarding the safety and reliability of AI is essential for its successful and widespread integration into healthcare systems.[42] Without this trust, even the most accurate AI systems will likely face resistance and underutilization in clinical settings.

The novel research presented by the MIT team makes a significant contribution towards addressing these concerns by effectively reducing the size of prediction sets generated by conformal classification, a key limitation that previously hindered its practical application.[28] Crucially, this reduction in the number of potential diagnoses is

achieved without compromising the fundamental guarantee that the correct diagnosis will be included within the set, thus maintaining the reliability of the predictions.[26] The technique itself is designed to be straightforward to implement, has proven effective in practice across various benchmark datasets, and importantly, does not require any computationally expensive retraining of the underlying AI model, making it readily adaptable to existing medical AI systems.[26] This research offers a practical and efficient approach to enhance the utility of conformal prediction, making it a more viable and user-friendly solution for uncertainty quantification in critical medical imaging applications.

By providing clinicians with smaller, more reliable sets of potential diagnoses, this work contributes significantly to making AI a more trustworthy and valuable tool in their daily practice.[25] This advancement has the potential to lead to more efficient diagnostic processes, ultimately improving patient outcomes and streamlining the delivery of medical care.[1] As AI continues to evolve and play an increasingly important role in healthcare, methods such as the one developed by these MIT researchers, which prioritize transparency, reliability, and the effective communication of uncertainty, will be crucial for realizing the full potential of AI to improve medical practice and enhance patient well-being.[6] This research represents a substantial step forward in making AI a more dependable and integrated component of medical workflows, ultimately contributing to the provision of better and safer healthcare for all patients.

Works cited

1. The importance of AI diagnostics and its impact on patients | Immunostep Biotech, accessed May 5, 2025, https://immunostep.com/2024/10/09/the-importance-of-ai-diagnostics-and-its-impact-on-patients/
2. How is AI used in healthcare? What you need to know and what's next, accessed May 5, 2025, https://exec.mit.edu/s/blog-post/how-is-ai-used-in-healthcare-what-you-need-to-know-and-what-s-next-20YU100000NQpBRMA1
3. The Benefits of the Latest AI Technologies for Patients and Clinicians, accessed May 5, 2025, https://postgraduateeducation.hms.harvard.edu/trends-medicine/benefits-latest-ai-technologies-patients-clinicians
4. New AI method captures uncertainty in medical images | MIT News, accessed May 5, 2025, https://news.mit.edu/2024/new-ai-method-captures-uncertainty-medical-images-0411
5. Making AI models more trustworthy for high-stakes settings | ScienceDaily,

accessed May 5, 2025, https://www.sciencedaily.com/releases/2025/05/250501164119.htm
6. The Potential of Artificial Intelligence Tools for Reducing Uncertainty in Medicine and Directions for Medical Education, accessed May 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11554287/
7. Uncertainty and Decisions in Medical Informatics - People | MIT CSAIL, accessed May 5, 2025, https://people.csail.mit.edu/psz/ftp/uncertainty.pdf
8. Navigating the Uncertainties of Medicine, accessed May 5, 2025, https://magazine.hms.harvard.edu/articles/navigating-uncertainties-medicine
9. Quantifying Uncertainty in Deep Learning of Radiologic Images - RSNA Journals, accessed May 5, 2025, https://pubs.rsna.org/doi/full/10.1148/radiol.222217
10. Handling the predictive uncertainty of convolutional neural network in medical image analysis: a review - Hirimutugoda, accessed May 5, 2025, https://jmai.amegroups.org/article/view/8191/html
11. Deep learning-based uncertainty quantification for quality assurance in hepatobiliary imaging-based techniques - PMC - PubMed Central, accessed May 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11970935/
12. A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods | Request PDF - ResearchGate, accessed May 5, 2025, https://www.researchgate.net/publication/381319369_A_review_of_uncertainty_quantification_in_medical_image_analysis_Probabilistic_and_non-probabilistic_methods
13. A Review of Uncertainty Estimation and its Application in Medical Imaging - arXiv, accessed May 5, 2025, https://arxiv.org/pdf/2302.08119
14. Handling of uncertainty in medical data using machine learning and probability theory techniques: a review of 30 years (1991–2020) - PubMed Central, accessed May 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC7982279/
15. Uncertainty Estimation in Medical Image Classification: Systematic Review, accessed May 5, 2025, https://medinform.jmir.org/2022/8/e36427/
16. Conformal Prediction for Machine Learning Classification -From the Ground Up, accessed May 5, 2025, https://towardsdatascience.com/conformal-prediction-for-machine-learning-classification-from-the-ground-up-a12fcf6860d0/
17. Intuitive explanation of conformal prediction - Cross Validated, accessed May 5, 2025, https://stats.stackexchange.com/questions/608844/intuitive-explanation-of-conformal-prediction
18. en.wikipedia.org, accessed May 5, 2025, https://en.wikipedia.org/wiki/Conformal_prediction#:~:text=Conformal%20prediction%20(CP)%20is%20a,assuming%20exchangeability%20of%20the%20data.
19. Conformal prediction - Wikipedia, accessed May 5, 2025, https://en.wikipedia.org/wiki/Conformal_prediction
20. What Does Conformal Prediction Add to Highly Accurate Models? - Cross Validated, accessed May 5, 2025, https://stats.stackexchange.com/questions/659071/what-does-conformal-predict

ion-add-to-highly-accurate-models
21. Conformal Prediction, accessed May 5, 2025, https://www.stat.berkeley.edu/~ryantibs/statlearn-s23/lectures/conformal.pdf
22. dmshanmugam.github.io, accessed May 5, 2025, https://dmshanmugam.github.io/pdfs/CVPR_2025_TTA_CP.pdf
23. Test Time Augmentation (Experimental) - PyTorch Tabular, accessed May 5, 2025, https://pytorch-tabular.readthedocs.io/en/latest/tutorials/11-Test%20Time%20Augmentation/
24. How to Use Test-Time Augmentation to Make Better Predictions - Machine Learning Mastery, accessed May 5, 2025, https://machinelearningmastery.com/how-to-use-test-time-augmentation-to-improve-model-performance-for-image-classification/
25. MIT Combines Test-Time Augmentation and Conformal Classification to Enhance AI Trustworthiness and Reduce Uncertainty in Medical Imaging - Forward Pathway, accessed May 5, 2025, https://www.forwardpathway.us/mit-combines-test-time-augmentation-and-conformal-classification-to-enhance-ai-trustworthiness-and-reduce-uncertainty-in-medical-imaging
26. Making AI models more trustworthy for high-stakes settings | MIT News, accessed May 5, 2025, https://news.mit.edu/2025/making-ai-models-more-trustworthy-high-stakes-settings-0501
27. How MIT Researchers Made AI Models More Trustworthy For High-Stakes Medical Imaging, accessed May 5, 2025, https://quantumzeitgeist.com/how-mit-researchers-made-ai-models-more-trustworthy-for-high-stakes-medical-imaging/
28. Enhancing Trustworthiness of AI in Medical Imaging - The Munich Eye, accessed May 5, 2025, https://themunicheye.com/boosting-trust-ai-models-medical-imaging-20147
29. CVPR Poster Test-time augmentation improves efficiency in ..., accessed May 5, 2025, https://cvpr.thecvf.com/virtual/2025/poster/34685
30. CVPR 2025 Papers, accessed May 5, 2025, https://cvpr.thecvf.com/virtual/current/papers.html
31. CVPR 2025 Accepted Papers, accessed May 5, 2025, https://cvpr.thecvf.com/Conferences/2025/AcceptedPapers
32. Improving the efficiency of conformal predictors via test-time augmentation - OpenReview, accessed May 5, 2025, https://openreview.net/forum?id=ylNucFNbcZ
33. Conformal Prediction-based Machine Learning in Cheminformatics: Current Applications and New Challenges | Theoretical and Computational Chemistry | ChemRxiv | Cambridge Open Engage, accessed May 5, 2025, https://chemrxiv.org/engage/chemrxiv/article-details/679783486dde43c90894415d
34. Better Aggregation in Test-Time Augmentation - CVF Open Access, accessed May 5, 2025,

https://openaccess.thecvf.com/content/ICCV2021/papers/Shanmugam_Better_Aggregation_in_Test-Time_Augmentation_ICCV_2021_paper.pdf
35. Learning Loss for Test-Time Augmentation - NIPS papers, accessed May 5, 2025, https://proceedings.neurips.cc/paper/2020/file/2ba596643cbbbc20318224181fa46b28-Paper.pdf
36. Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis - PubMed, accessed May 5, 2025, https://pubmed.ncbi.nlm.nih.gov/38553168/
37. AI Predictive Uncertainty: A Step Forward | Radiology - RSNA Journals, accessed May 5, 2025, https://pubs.rsna.org/doi/full/10.1148/radiol.232144
38. Conformal Prediction in Classification - Ardigen, accessed May 5, 2025, https://ardigen.com/conformal-prediction-in-classification/
39. [2206.13607] Improved Text Classification via Test-Time Augmentation - arXiv, accessed May 5, 2025, https://arxiv.org/abs/2206.13607
40. [PDF] Improved Text Classification via Test-Time Augmentation, accessed May 5, 2025, https://www.semanticscholar.org/paper/Improved-Text-Classification-via-Test-Time-Lu-Shanmugam/3836dfda6067fef17cf34197009ef0596483d945
41. Test-Time Augmentation In Machine Learning. - YouTube, accessed May 5, 2025, https://www.youtube.com/watch?v=mEPeZjpA_eA
42. Safety challenges of AI in medicine in the era of large language models - arXiv, accessed May 5, 2025, https://arxiv.org/html/2409.18968v2
43. Resistance to Medical Artificial Intelligence | Journal of Consumer Research, accessed May 5, 2025, https://academic.oup.com/jcr/article/46/4/629/5485292?guestAccessKey=b9f4ae04-bba1-4fac-8e6c-61dbd2ea9bde
44. Fairness of artificial intelligence in healthcare: review and recommendations - PMC, accessed May 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10764412/
45. The challenge of uncertainty quantification of large language models in medicine - arXiv, accessed May 5, 2025, https://arxiv.org/html/2504.05278v1
46. How Artificial Intelligence Can Help Doctors with “Medical Poker” | MGH IHP, accessed May 5, 2025, https://www.mghihp.edu/news-and-more/stories/how-artificial-intelligence-can-help-doctors-medical-poker