Stanford University
Human-Centered
Artificial Intelligence

Policy Brief
HAI Policy & Society
May 2026

Operationalizing Real-Time Monitoring of Clinical AI

Zhongnan Fang, Lina Y. Cheuy, Hye Sun Na, Akshay S. Chaudhari,* and David B. Larson*

AI tools are increasingly used in radiology, with the specialty accounting for approximately 76% of all FDA-authorized AI-enabled medical devices as of December 2025. A variety of tools can detect anomalies in X-rays or CT scans and provide diagnostic support. Yet many of these AI systems are deployed with limited mechanisms for monitoring and evaluating their performance, leaving clinicians to determine on their own which AI outputs are reliable. Without effective post-deployment oversight, these tools risk contributing to diagnostic errors and missed findings.

In our paper "Automated real-time assessment of intracranial hemorrhage detection AI using an ensemble monitoring model (EMM)," we introduce a new framework to enable real-time monitoring of AI radiology tool performance after deployment. Inspired by clinical consensus practices, the Ensemble Monitoring Model (EMM) measures agreement between a primary AI model and an ensemble of five independent submodels to estimate uncertainty without requiring access to black box model components. Using a large dataset focused on the detection of brain bleeds, we demonstrate that EMM can reduce radiologists' cognitive burden by effectively characterizing AI model uncertainty in real time at the point of care — when radiologists review both the images and the corresponding AI output — and guiding appropriate responses when cases are flagged for reduced accuracy.

Key Takeaways

Radiological AI tools account for the largest share of FDA-approved healthcare AI, yet clinical adoption remains slow and most deployed systems lack robust performance monitoring.

We introduce the Ensemble Monitoring Model (EMM) — a framework that assesses uncertainty in the predictions of radiology AI models trained to detect abnormalities (in this case, brain bleeds), thereby providing clinicians with actionable signals at the point of care and enabling real-time monitoring of AI tool performance after clinical adoption.

EMM addresses an urgent gap by offering a practical, customizable method for signaling when confidence is low in real time, diagnosing failure modes, and supporting retraining of clinical AI when needed.

Policymakers should treat continuous performance monitoring as a core component of responsible AI deployment in healthcare and consider requiring healthcare AI vendors to put in place post-deployment monitoring mechanisms.

* Equal Contribution

The growing reliance on AI in radiology and healthcare more broadly highlights that effective governance cannot stop at product approval. There is a critical need for total lifecycle management that ensures AI tools remain safe, accurate, and reliable after they are deployed in clinical settings. EMM enables AI models to be continually optimized and monitored after deployment. Policymakers should view methods like EMM as an important component of a broader regulatory strategy to ensure that AI in healthcare delivers measurable benefits without introducing new and unmanaged risks.

The growing reliance on AI in radiology and healthcare more broadly highlights that effective governance cannot stop at product approval.

Introduction

Despite an exponential increase in FDA-cleared radiological AI tools over the last decade, clinical adoption has been slow. These tools promise to enhance clinical efficiency — for example, by supporting radiology tasks that involve detecting anomalies in medical images and classifying or prioritizing different cases. Yet their adoption has also been accompanied by safety concerns, including a potential increase in misdiagnosis caused, for example, by cognitive pitfalls such as automation or confirmation bias. As a result, clinicians often have to meticulously verify each AI result.

Evidence shows that clinicians are strongly influenced by how certain an AI model claims to be about its predictions. When a system provides clear confidence information, physicians are more likely to incorporate the output into their decision-making. When no measure of certainty is available, clinicians are left to rely only on their own judgment and tend to trust the model far less.

Today, most monitoring of radiology AI systems still relies on retrospective, labor-intensive reviews of a small amount of manually labeled data, which provide only a partial view of real-world performance. To address this problem, researchers have developed a range of real-time monitoring techniques for estimating model confidence that use the same dataset the AI system was trained on to monitor it. Other methods approximate predictive reliability through the use of "deep ensembles," i.e., a collection of multiple smaller, independent models that stem from the same model architecture but are each trained from a different random starting point, causing them to learn in subtly different ways.

While these techniques can be effective in research settings, they share a major practical limitation: Nearly all of them require access to internal model components such as training datasets, model weights, or intermediate outputs. For commercial AI products, which are typically deployed as closed, black box

systems, this approach is largely unfeasible, leaving healthcare providers and policymakers without the means to oversee clinical adoption.

There is a need for real-time monitoring systems that can automatically characterize model confidence at the point of care without requiring access to internal model details. While measuring prediction uncertainty represents only one dimension of AI oversight — model performance can also be undermined by factors such as flawed input data, poor image quality, or improper image presentation — it remains a particularly important and substantive component of effective post-deployment evaluation.

There is a need for real-time monitoring systems that can automatically characterize model confidence at the point of care.

Research Outcomes

We designed an Ensemble Monitoring Model (EMM) that acts as a task-specific monitoring tool to estimate uncertainty of the internal black box of a deployed AI system. In our study, we use a large and diverse dataset of nearly 3,000 studies focused on the detection of intracranial hemorrhage (ICH), i.e., brain bleeds, to demonstrate the effectiveness of the EMM approach in characterizing the confidence of these AI systems. However, our findings can likely be generalized to other clinical applications as well.

Modeled on how clinicians reach group consensus, EMM pairs the primary model's diagnostic output with an independently generated confidence score derived from five submodels with diverse architectures (see Figure 1). This process produces a transparent agreement measure ranging from zero percent (none of the submodels agree with the primary model) to 100% (all five concur), indicating scenarios where the radiologist may have increased, similar, or decreased confidence in the primary model's output. This confidence score serves as an actionable signal, prompting the radiologist to proceed with certainty, take a closer look, or seek a second opinion.

A key question when evaluating EMM's real-world performance is whether warning clinicians about model uncertainty improves overall performance, even when some warnings are unnecessary. After all, false alarms (i.e., when the system flags a case as low-confidence despite the primary model's prediction being correct) could lead to increased cognitive burden and reduced overall trust in the AI system. Our analysis evaluated this trade-off by weighing accuracy gains from catching true errors against the downsides of false alerts.

We found that in cases where the primary model determines the patient to be ICH-positive, adding "level of uncertainty" alerts would consistently improve

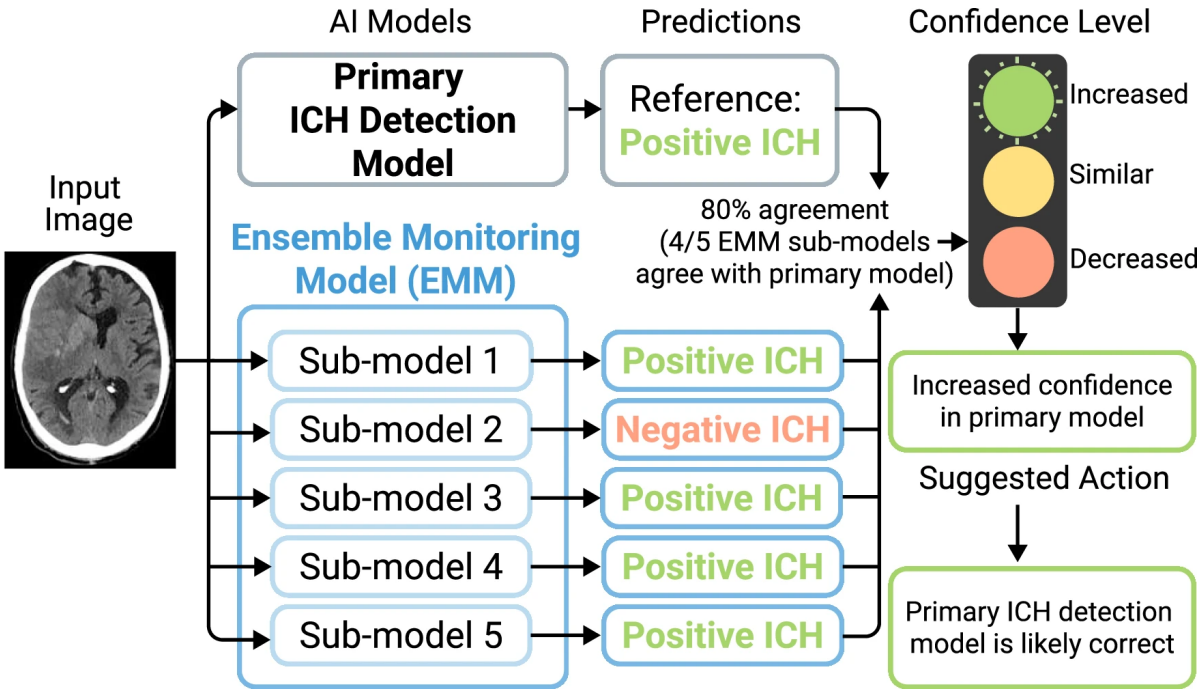

Figure 1: Each submodel within EMM is trained to perform the same task as the primary detection model. The independent submodel outputs are used to compute the level of agreement between the primary detection model and EMM, helping quantify confidence in the primary model's prediction and suggesting an appropriate subsequent action.

accuracy, helping radiologists either confirm the finding or detect and correct genuine mistakes, with benefits that substantially outweigh false-alarm rates across all prevalence levels. For cases predicted as ICH-negative, however, the value of alerts depends on disease prevalence. At higher prevalence levels, alerts provide modest gains, but at low prevalence, where baseline accuracy is already high, excessive false alarms often outweigh possible improvements.

These findings highlight that EMM performance depends on carefully calibrated thresholds. If thresholds are set too loosely, unnecessary alerts can undermine trust and efficiency, underscoring the need to tailor monitoring strategies to specific clinical contexts and disease prevalence.

Furthermore, analysis of failure cases shows that when EMM is incorrect, the errors typically arise from scenarios that are similarly challenging for human radiologists, such as extremely subtle hemorrhages or imaging features that closely mimic hemorrhage patterns. Performance improves as additional training data is incorporated, although gains begin to level off after roughly 25% of the full dataset. Further experiments varying the number of submodels from one to five demonstrate that monitoring performance increases with ensemble size, with four to five submodels representing the most effective amount for future implementations.

Policy Discussion

Healthcare providers across the United States are at different stages of AI adoption and face varying barriers to implementing responsible AI practices. One of these barriers is finding a way to reliably monitor the performance of AI tools after they have been deployed in clinical settings. This is crucial to ensuring not only their safety and efficacy, but also their continued usefulness.

Performance monitoring cannot rest solely with individual institutions.

Recognizing the importance of continued, real-time monitoring, healthcare regulators have more recently shifted their regulatory approach for medical devices, including AI-enabled devices, toward total product lifecycle management — an approach that prioritizes continuous, ongoing oversight from design through post-market performance. However, there are currently still limited guidelines or best practices for real-time monitoring to communicate an AI model's confidence in its predictions.

EMM addresses this gap, offering a mechanism for healthcare providers who currently or in the future plan to use diagnostic AI tools to operationalize post-deployment monitoring in everyday practice. The EMM approach provides a practical, customizable method for identifying performance degradation in real time, suggesting possible failure modes, and supporting retraining when needed.

However, given the fragmented nature of AI adoption by healthcare providers, performance monitoring cannot rest solely with individual institutions. AI vendors must also take responsibility for monitoring their tools' performance. Policymakers should consider requiring vendors to enable performance monitoring as a condition of deployment rather than regulating each individual application after the fact. For vendors offering multiple AI applications, at least one core system or application should be designated to enforce trust, safety, and accountability standards across the entire product suite, creating a centralized governance anchor. However, since vendors are often hesitant to reveal specific details on the inner components of their models, an independent regulatory or certifying body may be needed to oversee post-deployment monitoring.

In the case of radiology AI tools, a professional body such as the American College of Radiology (ACR) could spearhead such oversight as part of its accreditation process — a quality certification program in which radiologists and peers evaluate whether a medical imaging facility meets professional standards for safe and effective imaging. The ACR has already begun developing an accreditation process designed specifically for AI tools in radiology — real-time monitoring should be a component of this process. Other professional groups could serve similar functions for AI tools applied in other medical fields.

Our EMM approach comes with some practical limitations. The current reliance on labeled, use-case-specific datasets to train EMM for each clinical application may constrain broader adoption across institutions with limited labeling capacity or computational resources. Future research should therefore systematically evaluate the degree of adaptation required to maintain EMM performance under resource-constrained conditions, including the relative benefits of retraining, recalibration, or incremental data augmentation. Subgroup analyses further revealed performance discrepancies across gender, age, and racial groups, likely driven by underrepresentation in training data or distributional mismatch between training and testing populations. Addressing these disparities will require targeted efforts to improve generalizability not only across institutions, but also across populations and disease types beyond intracranial hemorrhage.

As AI becomes increasingly embedded in radiology and other high-stakes clinical workflows, ensuring patient safety and clinical effectiveness will depend not only on rigorous pre-market evaluation, but on continuous, real-world oversight. EMM demonstrates that it is both feasible and valuable to implement practical, black box-compatible monitoring that provides clinicians and institutions with actionable signals based on confidence in the primary model's outputs at the point of care. Policymakers should therefore view real-time performance monitoring as a core component of responsible AI deployment in healthcare — one that complements existing approval pathways and accreditation frameworks to ensure that AI systems deliver sustained clinical benefit without introducing unmanaged risk.

Reference: The original article is available at Zhongnan Fang et al., "**Automated real-time assessment of intracranial hemorrhage detection AI using an ensembled monitoring model (EMM),**" *NPJ Digital Medicine* 8(1):608, October 2025, https://www.nature.com/articles/s41746-025-02007-0.

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu**.

Zhongnan Fang is a principal machine learning scientist at the radiology department's AI Development and Evaluation (AIDE) Lab at Stanford University.

Lina Y. Cheuy is a clinical AI technical writer at the AIDE Lab.

Hye Sun Na is the managing director of the AIDE Lab.

Akshay S. Chaudhari is an associate professor of radiology and biomedical data science at Stanford University and the co-director of the AIDE Lab.

David B. Larson is a professor of radiology at Stanford University and the principal investigator and founding director of the AIDE Lab.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: 353 Jane Stanford Way, Stanford CA 94305-5008

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**