bialogics

White Paper | 2024

A Foundational Approach to Assess the Performance of AI in Diagnostic Imaging

Real-time AI performance monitoring can provide the needed transparency for the successful adoption of AI algorithms.

Real-time AI performance monitoring can provide the needed transparency for the successful adoption of AI algorithms

The future of radiology workflow necessitates the seamless integration of conventional data streams, DICOM-captured AI outputs, and structured data derived from Natural Language Processing engines. This integration is crucial for delivering the transparent monitoring and guidance for AI adoption.

The industry estimates that only 10% to 15% of the clinical radiology sector has deployed AI-enabled image analysis in practice, despite the influx of numerous AI algorithms coming to market. The slow uptake is attributed to various factors, most notably the absence of effective methods to gauge AI performance in practical settings.

This includes the challenge of accurately measuring algorithm accuracy and ROI. Multiple vendors, while claiming superiority in accuracy within controlled tests, are challenged to duplicate the performance in real-world settings. Moreover, these solutions tend to be costly with unproven returns on investment, aggravating the industry's struggle to validate vendors' claims using real-world data.

Ensuring the transparency in AI algorithm performance and accuracy through AI performance monitoring is crucial for fostering trust among clinicians, patients, and regulatory bodies. It assists the radiologists to better understand and have greater confidence in the AI-driven diagnostic or decision-making process, leading to more confident and informed healthcare decisions.

What is AI performance monitoring?

AI performance monitoring should provide clear metrics on the performance impact of all AI algorithms. It should include measures such as accuracy and any limitations or biases present in the AI's predictions based on the real-world data of the customers own patient demographics.

AI Performance monitoring compares AI outputs against the radiologists' diagnostic reports. It should take into account factors such as Concordance, Discordance, Sensitivity, and Specificity tracking, both in real time and over a longer period resulting in Drift Report predictive analysis. (The accuracy of AI can change over time if the environment changes, such as the introduction of new equipment, technologists, methodologies or the appearance of diseases such as Covid-19 which were previously unaccounted for in the algorithm.)

AI metrics must also include the impact of introducing AI into a busy radiology practice, and therefore perform its assessments in real-time for effective radiology use.

To effectively oversee and enhance AI systems, a specialized integration platform is crucial. This platform evaluates and measures AI model performance metrics, offering

bialogics
www.bialogics.com
info@bialogics.com

real-time dashboard analysis covering Concordance, Discordance, Sensitivity, Specificity, as well as Drift Reporting, enabling accurate predictions in radiology productivity analysis.

How does AI performance monitoring work?

AI algorithms used in image analysis and workflow assistance must seamlessly integrate into the radiology workflow, becoming an integral part of the data ecosystem across all radiology applications. This integration is crucial for effectively evaluating AI against the established benchmarks of radiology accuracy, productivity, and efficiency.

Establishing a standardized reporting approach starts with the real-time collection of the radiology workflow operational data. Standard HL7 and DICOM data associated with the radiology RIS/PACS systems are required, along with productivity metrics such as RVU look up tables and physician scheduling data. The radiology workflow operational data is then integrated with the clinical data, which is extracted from the diagnostic report and analyzed by machine learning (ML) engines leveraging Natural Language Processing (NLP).

To be AI data ready requires the convergence of vendor agnostic data into a single platform to extract the cross-data analysis of AI Performance Monitoring

The output of the AI algorithms can be collected using secondary capture and DICOM overlay techniques. This data is then compared to the previously collected operational and clinical data and the result of this comparison comprises AI performance monitoring. The integration and analysis of data from many diverse sources is critical to AI performance monitoring.

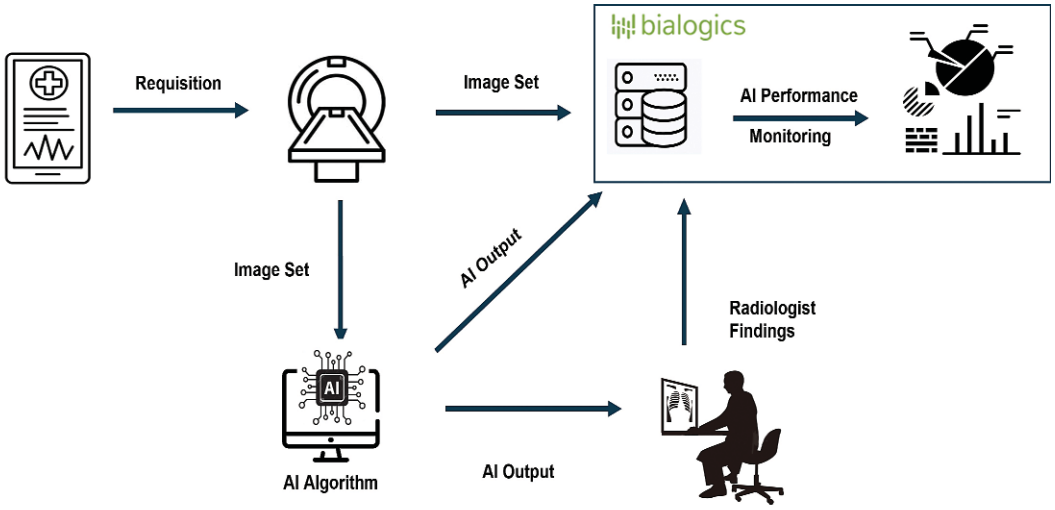


Fig 1: Real-Time AI Performance Monitoring Workflow

Continuously collecting and monitoring AI accuracy and effectiveness allows us to track AI effectiveness and accuracy over time, which is important for AI Drift Reporting. Preparing a radiology organization for data readiness involves establishing an integrated system that is vendor-agnostic and AI-flexible. This system should gather data from multiple RIS/PACS systems, various types of AI algorithms, and radiologists' clinical findings. Its real-time presentation capabilities can aid radiologists and enhance the overall radiology workflow.

bialogics
www.bialogics.com
info@bialogics.com

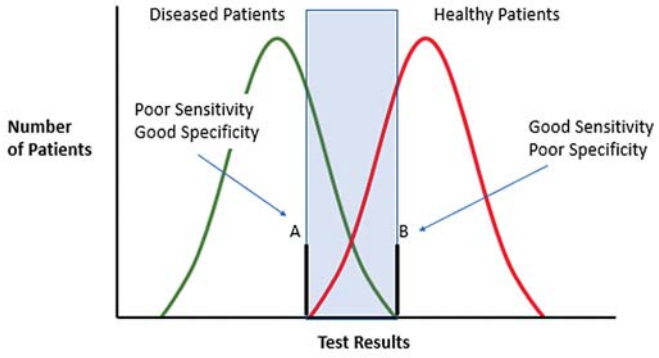
AI performance monitoring metrics

There are five primary operational analytics that need to be measured to provide real-time transparency into the effective use of AI in the radiology environment.

1. Sensitivity/Specificity

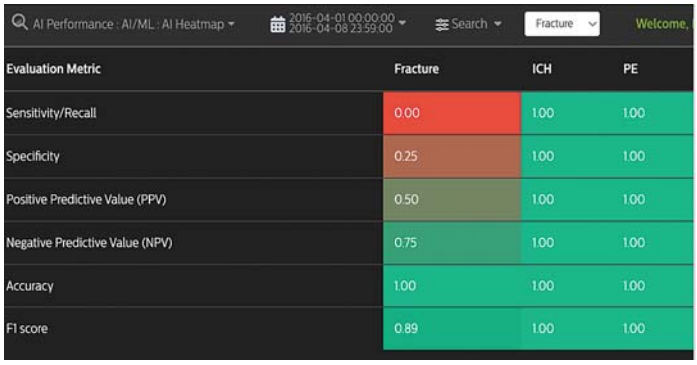
Two important concepts to evaluate the performance of a healthcare AI algorithm are sensitivity and specificity.

- Sensitivity indicates the proportion of actual positive cases correctly identified by the AI system. In radiology, sensitivity reflects the algorithm's ability to detect relevant findings or abnormalities in imaging studies. High sensitivity implies fewer false negatives, ensuring that the AI accurately identifies most of the true positive cases. Sensitivity is calculated as True Positives/(True Positives + False Negatives). It measures the ability of the test to correctly identify individuals who have the condition.



Sensitivity denotes the ratio of actual positive cases correctly identified, representing the Positive Predictive Value. Conversely, Specificity gauges the proportion of actual negative cases accurately identified, reflecting the Negative.

- Specificity measures the proportion of actual negative cases correctly identified by the AI system. In radiology, specificity signifies the algorithm's capacity to correctly rule out abnormalities when they are not present in the imaging studies. High specificity means fewer false positives, ensuring that the AI accurately identifies most of the true negative cases. Specificity is calculated as True Negatives/(True Negatives + False Positives). It measures the ability of the test to correctly identify individuals who do not have the condition.



Evaluation Metric	Fracture	ICH	PE
Sensitivity/Recall	0.00	1.00	1.00
Specificity	0.25	1.00	1.00
Positive Predictive Value (PPV)	0.50	1.00	1.00
Negative Predictive Value (NPV)	0.75	1.00	1.00
Accuracy	1.00	1.00	1.00
F1 score	0.89	1.00	1.00

These measurements play a vital role in evaluating the diagnostic accuracy of AI algorithms in radiology. A balance between sensitivity and specificity is essential. While high sensitivity helps in not missing actual positive cases, high specificity aids in

bialogics
www.bialogics.com
info@bialogics.com

reducing false positives, which can lead to unnecessary follow-up tests or interventions. Finding an optimal balance depends on the clinical context and the specific requirements of the diagnostic task.

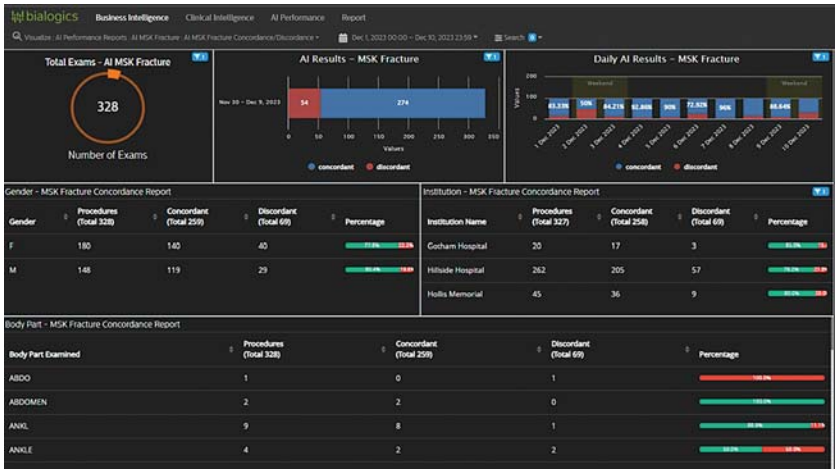
In practice, both Sensitivity and Specificity are calculated on the polarity of findings in both the AI algorithm and radiologist report and can be reported in real time, such as in the example below and filtered by many variables such as Body Part, AEtitle, Modality, Procedure and Radiologist, etc.

2. Concordance/Discordance

Concordance and Discordance reporting within radiology AI involves contrasting the diagnoses or assessments generated by the AI system against those determined by radiologists, often regarded as the ground truth. It's a straightforward assessment that entails comparing the AI's output with the findings of radiologists.

Concordance/Discordance indicates the accuracy between the AI output and the Radiologist ground truth, while also facilitating a real-time feedback loop for AI model enhancements.

- Concordance signifies agreement or similarity between the AI-generated diagnosis or assessment and the diagnosis by the radiologists or ground truth reporting. High concordance indicates that the AI system's findings align closely with those of the radiologist.
- Discordance signifies disagreement or divergence between the AI-generated diagnosis or assessment of the radiologists' diagnosis. Discordance might occur when the AI system identifies something that the human experts did not, or vice versa.



These reporting measures are useful in evaluating the performance and reliability of AI algorithms. Understanding concordance and discordance helps in assessing the AI system's strengths and weaknesses, highlighting areas where the AI might excel or fall short compared to human interpretations. Real time Concordance/Discordance monitoring can be evaluated in many ways, for example by age, gender, scanner, or by radiologist.

Monitoring and analyzing concordance and discordance aid in refining AI algorithms, improving their accuracy, and building confidence in their diagnostic capabilities. They also offer valuable insights into cases where the AI can complement or enhance the radiologists' findings, leading to more comprehensive and accurate diagnoses.

3. AI Confidence Scoring

AI confidence scoring in radiology refers to the algorithm's assessment or estimation of its own reliability or certainty regarding a particular diagnosis or decision made based on the imaging data.

When an AI system analyzes medical images, it doesn't just provide a diagnosis but often assigns a level of confidence or certainty to that diagnosis. However not all AI algorithms calculate this value. Confidence scoring involves the AI algorithm assigning a probability or confidence level to its output, indicating how confident it is about the accuracy of its assessment.

For instance, after examining an image, the AI might assign a high confidence score if it's very certain about its diagnosis or a lower score if it's less certain. This scoring helps clinicians understand the degree of trust they can place in the AI-generated diagnosis or recommendation.

Understanding the impact of AI on radiologist read times allows for workflow optimization, potentially reducing time spent on routine cases and focusing on complex studies.

The confidence score can be presented in various forms, such as a percentage, a grading scale, or a visual indicator. This information is valuable for radiologists as it allows them to factor in the AI's level of confidence when making their own assessments. It can also guide them in cases where the AI is uncertain, prompting further investigation or human intervention to ensure accurate diagnoses.

4. Measuring the impact of AI on radiology efficiency

Evaluating radiology efficiency and productivity goes back to basic business fundamentals, using pre- and post-AI evaluation times. Examples include Radiologist Turn Around Times (TAT), Read Times and Productivity measurements like Relative Value Units (RVU), all of which combine the messages from HL7, DICOM, look up tables and converting unstructured textual data to structured data with Natural Language Processing.

Understanding how AI impacts radiologist read times helps optimize workflow. If AI tools can expedite the interpretation process without compromising accuracy, they can potentially reduce the time radiologists spend on routine cases, allowing them to focus more on complex or critical studies.

Efficiency measurements assist in allocating resources effectively. Identifying which cases or modalities benefit the most from AI assistance helps in directing resources and AI tools to areas where they can make the most significant impact.

Effectiveness monitoring enables the comparison of radiologist read times before and after AI implementation for the assessment of AI's effectiveness. If there's a notable reduction in read times without compromising diagnostic accuracy, it showcases the efficiency gains achieved through AI integration.

Assessing AI efficiency by measuring radiologist read times offers multiple advantages, potentially resulting in cost savings through more efficient use. It's important to measure radiologist Read tTmes and/or Turnaround Times (TATs) when integrating AI into radiology workflow for gauging the comprehensive efficiency, impact, and advantages of AI adoption in clinical practice.

5. AI Drift Reporting and Predictive Trending

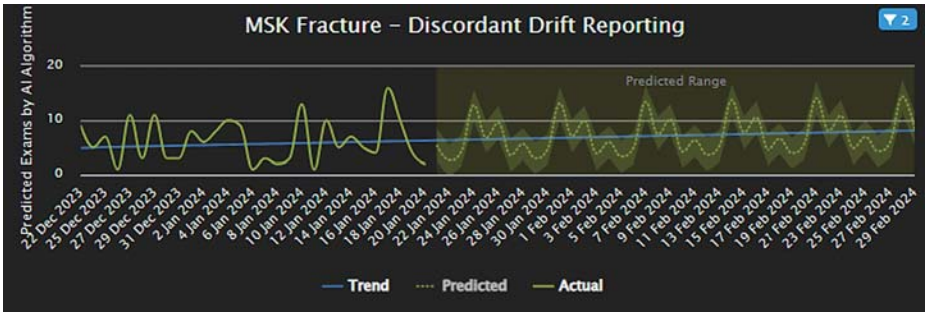
AI drift reporting and predictive trending are essential elements in monitoring and maintaining the performance and accuracy of AI algorithms over time.

AI drift reporting and predictive trending are vital for maintaining the performance and accuracy of AI algorithms, enabling continuous assessment against predefined standards.

AI Drift refers to the gradual change or degradation in the performance of an AI algorithm over time due to various factors such as changes in data patterns, evolving patient demographics, or shifts in equipment. Drift reporting involves the continuous assessment and analysis of AI algorithm performance against a set of predefined benchmarks or standards.

Any deviations or discrepancies in the algorithm's performance from the expected standards are identified in real-time, or more likely over a period of time, to see progressive changes in the number of discrepancies or changes to key metrics being used to measure the effectiveness of the algorithm.

Predictive trending involves forecasting future patterns or trends based on historical data and ongoing observations from AI algorithms.



Pattern Recognition, with Machine Learning (ML) data engines can be used to process large data sets analyzing historical data to identify trends, patterns, or recurring behaviors. By recognizing these patterns, the ML engines can predict potential future trends, allowing for proactive decision-making.

Both AI Drift reporting and predictive trending are necessary in ensuring the reliability, accuracy, and ongoing efficacy of AI algorithms in radiology by continuously assessing their performance and predicting future trends or changes. This ensures that the AI is still aligned with the evolving needs of clinical practice and patient care.

Summary

AI adoption guidelines are emerging to help potential users of AI-based solutions in radiology navigate the increasing number of commercial products. This encourages their adoption in real-world scenarios, by allowing their true potential to be assessed, as well as their weaknesses to be identified and addressed in a safe and effective way. As these incremental improvements are made, these tools will likely evolve to handle more varied data, become integrated into consolidated workflows, become more transparent, and ultimately more useful for increasing efficiency and improving patient care.

bialogics
www.bialogics.com
info@bialogics.com

About Bialogics:

Bialogics stands at the forefront of precision analysis in Diagnostic Imaging data and patient workflows. Their innovative approach establishes a real-time environment aimed at harnessing clinical and operational data to enhance business performance and clinical outcomes. Through the integration of multiple imaging data sources and deep learning ML, Bialogics has developed a practical solution for all size Diagnostic Imaging organizations This solution offers ongoing evaluation of Business and Clinical intelligence analytics as well as AI performance monitoring of AI algorithms in current use or for retrospective research, enabling real-time monitoring of AI metrics. The Bialogics platform is vendor agnostic and fully interoperable to all data sources used in Radiology

Contact: info@bialogics.com

References

Developing, Purchasing, Implementing and Monitoring AI Tools in Radiology: Practical Considerations. A Multi-Society Statement from the ACR, CAR, ESR, RANZCR and RSNA
https://pubs.rsna.org/doi/10.1148/ryai.230513

How Radiologists Can Expand the Utilization of AI
https://www.itnonline.com/article/how-radiologists-can-expand-utilization-ai]

Methods for Clinical Evaluation of Artificial Intelligence Algorithms for Medical Diagnosis
https://pubs.rsna.org/doi/full/10.1148/radiol.220182

Assessment of Radiology Artificial Intelligence Software: A Validation and Evaluation Framework
https://journals.sagepub.com/doi/full/10.1177/08465371221135760

(AI) applications in radiology: hindering and facilitating factors. Eur Radiol 30, 5525–5532 (2020).
https://doi.org/10.1007/s00330-020-06946-y

Foundations for Biomedical Science: Sensitivity and Specificity
https://oercollective.caul.edu.au/foundations-of-biomedical-science/chapter/10-4-sensitivity-and-specificity-are-inversely-related/

bialogics
www.bialogics.com
info@bialogics.com