Article published online: 2023-08-15



Thieme

# Closing the loop for AI-ready radiology

## Die Zukunft der Radiologie: Vertikale Integration und KI im Einklang

Authors

Moritz Fuchs[1], Camila Gonzalez[1], Yannik Frisch[1], Paul Hahn[1], Philipp Matthies[2], Maximilian Gruening[3], Daniel Pinto dos Santos[4, 5], Thomas Dratsch[4], Moon Kim[6, 7], Felix Nensa[6, 7], Manuel Trenz[3], Anirban Mukhopadhyay[1]

Affiliations

1 Informatics, TU Darmstadt, Germany
2 AI, Smart Reporting GmbH, München, Germany
3 Interorganisational Informationssystems, Georg-August-Universität Göttingen, Goettingen, Germany
4 Institute for Diagnostic and Interventional Radiology, Uniklinik Koln, Germany
5 Institute for Diagnostic and Interventional Radiology, Universitätsklinikum Frankfurt, Frankfurt am Main, Germany
6 Institute for Diagnostic and Interventional Radiology and Neuroradiology, Universitätsklinikum Essen, Germany
7 Institute for Artificial Intelligence in Medicine, Universitätsklinikum Essen, Germany

Key words

AI, lifelong learning, structured reports, ai visualization

**received** 27.02.2023
**accepted** 01.07.2023
**published online** 15.08.2023

Bibliography

Fortschr Röntgenstr 2024; 196: 154–162
**DOI** 10.1055/a-2124-1958
**ISSN** 1438-9029
© 2023. Thieme. All rights reserved.
Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Correspondence

Moritz Fuchs
Informatics, TU Darmstadt, Interactive Graphics Systems, Fraunhoferstr. 5, 64283 Darmstadt, Germany
Tel.: +49/61 51/15 56 55
moritz.fuchs@gris.tu-darmstadt.de

## ABSTRACT

**Background** In recent years, AI has made significant advancements in medical diagnosis and prognosis. However, the incorporation of AI into clinical practice is still challenging and under-appreciated. We aim to demonstrate a possible vertical integration approach to close the loop for AI-ready radiology.

**Method** This study highlights the importance of two-way communication for AI-assisted radiology. As a key part of the methodology, it demonstrates the integration of AI systems into clinical practice with structured reports and AI visualization, giving more insight into the AI system. By integrating cooperative lifelong learning into the AI system, we ensure the long-term effectiveness of the AI system, while keeping the radiologist in the loop.

**Results** We demonstrate the use of lifelong learning for AI systems by incorporating AI visualization and structured reports. We evaluate Memory Aware-Synapses and Rehearsal approach and find that both approaches work in practice. Furthermore, we see the advantage of lifelong learning algorithms that do not require the storing or maintaining of samples from previous datasets.

**Conclusion** In conclusion, incorporating AI into the clinical routine of radiology requires a two-way communication approach and seamless integration of the AI system, which we achieve with structured reports and visualization of the insight gained by the model. Closing the loop for radiology leads to successful integration, enabling lifelong learning for the AI system, which is crucial for sustainable long-term performance.

**Key Points:**
- The integration of AI systems into the clinical routine with structured reports and AI visualization.
- Two-way communication between AI and radiologists is necessary to enable AI that keeps the radiologist in the loop.
- Closing the loop enables lifelong learning, which is crucial for long-term, high-performing AI in radiology.

## ZUSAMMENFASSUNG

**Hintergrund** In den letzten Jahren hat die KI erhebliche Fortschritte bei der medizinischen Diagnose und Prognose erzielt. Jedoch bleibt die Integration von KI in die klinische Praxis eine Herausforderung und wird nicht ausreichend gewürdigt. Wir wollen einen möglichen vertikalen Integrationsansatz aufzeigen, um den Kreislauf für eine KI-kompatible Radiologie zu schließen.

**Methode** Diese Studie unterstreicht die Bedeutung der wechselseitigen Kommunikation für die KI-gestützte Radiologie. Darüber hinaus wird als wesentlicher Teil der Methodik die Integration des KI-Systems mit strukturierten Berichten und KI-Visualisierungen in die klinische Praxis demonstriert. Durch die Integration von lebenslangem Lernen stellen wir die langfristige Effektivität der KI sicher und halten gleichzeitig den Radiologen auf dem Laufenden.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

**Ergebnisse** Wir demonstrieren den Einsatz von lifelong learning für KI-Systeme durch die Einbeziehung von KI-Visualisierungen und strukturierten Befunden. Wir evaluieren Memory Aware-Synapses und Rehearsal-Methoden und zeigen in der Praxis, dass beide funktionieren. Wir sehen vor allem Vorteile von Algorithmen für lifelong learning, wie Memory Aware-Synapses, wenn sie keine Muster aus früheren Datensätzen speichern oder verwalten müssen.

**Schlussfolgerung** Die Einbindung von KI in die klinische Routine von Radiologen erfordert einen zweiseitigen Kommunikationsansatz und eine nahtlose Integration des KI-Systems mit strukturierten Berichten und KI-Visualisierungen, die Erkenntnisse des KI-Models repräsentieren. Die erfolgreiche Integration führt zu einem Kreislaufsystem mit Radiologen, das lebenslanges Lernen für KI-Systeme ermöglicht, was für die langfristige und nachhaltige Leistungsfähigkeit entscheidend ist.

**Kernaussagen:**
- Wir demonstrieren die Integration von KI-Systemen in klinische Routinen mit strukturierten Berichten und KI-Visualisierungen.
- Eine bi-direktionale Kommunikation zwischen KI und Radiologen ist notwendig, um KI im radiologischen Alltag zu ermöglichen.
- Der vorgestellte Kreislauf ermöglicht lebenslanges Lernen, was für eine langfristige, leistungsstarke KI in der Radiologie entscheidend ist.

**Zitierweise**
- Fuchs M, Gonzalez C, Frisch Y et al. Closing the loop for AI-ready radiology. Fortschr Röntgenstr 2024; 196: 154–162

## Introduction

The introduction of deep learning models in clinical radiology is evolving more gradually than anticipated due to issues involving effectiveness, regulatory concerns, and the difficulty in establishing sustainable business models [1, 2]. Furthermore, the development of artificial intelligence-driven radiology requires trust and interpretability of AI systems for radiologists and patients [3–5]. Deep learning and AI-driven radiology have the potential to aid radiotherapy planning [6] and tumor detection [7], among other use cases. These technologies promise to improve the accuracy and efficiency of many critical processes, leading to better patient outcomes.

For AI systems to become more relevant for radiology in practice, AI support needs to become seamless to reduce the time spent per case. Recent developments in AI try to elevate the tight time constraints that radiologists face when reviewing cases, which lead to missed findings and long turnaround times [8]. Considering this situation, we see the need for vertical integration of images of AI-driven decisions and machine-readable structured reports to support decision-making with AI systems.

Most current research bypasses complex technical integration into real-world applications by simplifying assumptions or working only with ***immaculate*** data that have been carefully cleaned and homogenous datasets selected for the research. In real-world applications, the data gathered in hospitals is heterogenous due to differences between hospital infrastructures, different devices, or other inconsistencies. As a result of such *domain shifts*, many AI systems suffer from **worse performance in practice as the AI system ages** [9, 10] or is deployed in new environments [11–13]. Common domain shifts occur between different institutions due to changes in populations or devices. The data also shifts over time due to updates to reconstruction algorithms and acquisition protocols. Neglecting technical integration problems while benchmarking assistive technology in isolation leads to ***silent failures*** of AI systems [9–13].

A leading approach to address this problematic setting is the concept of ***lifelong learning*** [10, 13, 14], also known as *continual learning*. We first developed an AI system to the best of our knowledge and continuously updated the system with new data. This concept has the advantage that it can handle data until it becomes inaccessible, e. g., due to GDPR [15], or adapt to newly available data distributions, e. g., in the case of COVID-19 [16]. An example of how these advantages can be leveraged to provide fast adaptability to unpredictable events, such as future pandemics, is by allowing individuals to altruistically share their relevant data without compromising their privacy. As the AI system matures, individuals may want to redact any shared data they had initially provided. This can be done without impacting the overall performance of the AI system.

In order to integrate lifelong learning into the clinical workflow, we must engage our AI systems in ***cooperative* lifelong learning** with radiologists. For successful collaboration, the AI system needs to receive information from radiologists through a machine-readable format instead of unstructured text. The radiologist, similarly, must receive human-readable insight from the AI system. A *machine-readable report* uses a structure that complies with guidelines and grasps relevant information without incurring additional costs. While the benefits of lifelong learning are evident [14], current medical device regulations hinder its applicability.

To collect structured reports efficiently, we support the radiologist with images of the insight gained by the AI system [17]. This insight needs to integrate seamlessly into the clinical workflow of a radiologist in order to foster trust and enhance the effectiveness of the cooperation. The integration of AI systems with structured reports, lifelong learning, and AI images enables realistic studies regarding AI that keeps the radiologist in the loop.

As an example of such a workflow, we demonstrate a possible use for diagnosing pulmonary embolism (PE) from CT scans. This article shows the integration of visualization techniques for AI systems with structured reports and lifelong learning. The system gives reliable insight into the model's predictions for radiologists

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.


▶ **Fig. 1** Four different images providing transparent insight for interpretable AI. The whole screen overlays the predicted class probability for the presence of a PE. The cutout version leaves a clear view window so as not to obstruct the view. The bar score displays the probability (here 51 %) from zero to one hundred percent (from left to right) above the corpus. The saliency maps show how slice-wise activation maps can be misleading by not correctly highlighting the PE.

▶ **Abb. 1** Vier verschiedene Visualisierungen bieten Einblicke für eine interpretierbarere KI. Der Gesamtbildschirm überlagert die vorhergesagte Klassenwahrscheinlichkeit für das Vorhandensein einer Lungenarterienembolie (LAE). Die Ausschnittversion lässt ein freies Sichtfenster, um die Sicht nicht zu behindern. Die Balkenanzeige über dem Korpus zeigt die Wahrscheinlichkeit (hier 51 %) von null bis hundert (von links nach rechts) Prozent an, das diese Sicht eine Embolie enthält. Die Salienzkarten zeigen, wie die scheibenweise Aktivierung irreführend sein kann, da sie die LAE nicht korrekt hervorhebt.

while collecting crucial information and increasing comprehensibility. Furthermore, we show how to integrate images in any existing PACS framework and enable structured reporting with AI support with the press of a single button.

## Materials and Methods

In order to achieve the complex vertical integration of our system, we first summarize possible methods to provide interpretable AI visualization and explain how a radiologist can use and interpret such images. Secondly, we outline how we built the structured reporting template. Lastly, we describe the integration of all concepts into the clinical workflow to enable lifelong learning and a sustainable high-quality AI system.

### AI Visualization

One of the predominant directions for interpretable AI is the usage of *saliency maps* [18, 19], which visualize the most relevant part of the image for the given prediction. However, recent research has shown some caveats when estimating saliency maps [19–21]. The most relevant concern is that a saliency map can highlight irrelevant regions, stick to edges, or be easily fooled by minimal changes [1, 18–20]. It has also been shown that the highlighting changes immensely with the neural network architecture [19] and only produces relevant insight for 2 D images. Furthermore, a user study showed that predicting how a given neural network's saliency map looks is impossible for humans [21]. These issues are detrimental to saliency map's comprehensibility for a radiologist and could have severe implications for their widespread adoption.

However, saliency maps are most valuable when localization is essential for the task, as in detecting PE [22]. In our framework, as is common among the winning RSNA PE detection models [23], we use a two-stage solution: we first train a SEResNeXt (50 layers) [24] on a slice-wise annotation level. Subsequently, we aggregate the 128 features with a two-layer LSTM [25] from the second to last layer for the whole volume prediction per patient.

We generate three different images to provide transparencies of the prediction probability. We select the best image that balances local information and comprehensibility for our radiologists. As a comparison, we also visualize the slice-wise saliency map, using the predicted probability for each slice of the first stage of our model.

▶ **Fig. 1** shows the four different approaches on the same frame – starting from the left, the whole screen and a cutout overlay, which displays the overlay only for positively predicted slices. The overlay uses alpha-blending [26], with the alpha values being proportional to the probability as intensity to indicate confidence in the prediction. However, our clinicians found the alpha values hard to judge. As the alpha value approaches higher values, the image only shows the overlay, making it hard to diagnose. The third example displays a bar, with alpha = 0.5, that indicates the positive prediction probability at the top of the torso, ranging from zero (left) to one (full bar till right). Lastly, we generate saliency maps with M3d-CAM [27]. We evaluate multiple methods for saliency visualization, and our radiologist found the implementation of Grad-cam++ [28] to consistently yield the best match. However, the example in ▶ **Fig. 1** shows how saliency maps might highlight non-relevant parts for detecting PE. This might distract clinicians, obstruct their view, and ruin their trust in the AI system, discouraging future usage. Furthermore, we often observed major changes in saliency maps by moving a single slice. The poor performance of saliency maps can be attributed to the sparse labelling of the CT volumes rather than giving slice-wise information, in our hospital's data annotations.

### Structured Reporting

Structured reporting is a method of describing medical images in a consistent way. A structured reporting scheme is fundamental for AI systems to engage in lifelong learning. One of the main advantages of integrating structured reports with lifelong learning

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.


▶ **Fig. 2** The integration of our bar score images for more transparent AI systems with the structured reporting template in the OHIF viewer. The viewer offers an intuitive and easy-to-fill design of the structured report for our use case of PE detection.

▶ **Abb. 2** Die Integration unserer Balkenanzeige Visualisierungen für transparentere KI-Systeme mit der strukturierten Befundvorlage im OHIF-Viewer. Der Viewer bietet ein intuitives und einfach auszufüllendes Design eines strukturierten Befunds für unseren Anwendungsfall der LAE-Erkennung.

into radiology is the ability to improve the interpretability of deep learning models.

In this section, we describe how we design our scheme and adapt it to better serve the integration of AI systems into the clinical workflow. First, we develop the structured reporting template for PE localization according to guidelines and our radiologists [29, 30]. The basic template covers all mandatory topics and questions to answer regarding PE. This includes a structure to classify the PE location according to different levels: central, lobar, segmental (seg.), and subsegmental (subseg.). Furthermore, we divide each side of the lung into three segments consisting of upper, middle, and lower for the left and right sides. Finally, a bifurcation can be labelled separately, resulting in 21 location labels. We support this with an easy-to-fill design in the *OHIF Viewer* [31] while leaving room to capture additional patient information. The different images are in the left bar. Our design allows for the communication of insight from the AI system to the clinician while being able to switch back to the original CT scan. By filling the report with more information in a structured way, the radiologist allows us to feed the AI system with new knowledge.

Additionally, our template includes fields for location and measurement of probability for all labels, which open up when a respective field is selected. The added transparency, provided by probabilities and visualization, increases the perceived reliability and trust in the AI system. However, the template can only cover some possible information in a structured way. Creating a comprehensive template that encompasses every potential finding in a chest CT scan is not possible since new unknown diseases are going to emerge (e. g., COVID-19). Consequently, it is crucial to determine an effective approach to address rare findings, such as atypical pathologies (e. g., Castleman's disease), foreign bodies (dental work, misplaced catheters), or anatomical variations (Azygos lobe, Horseshoe lung).

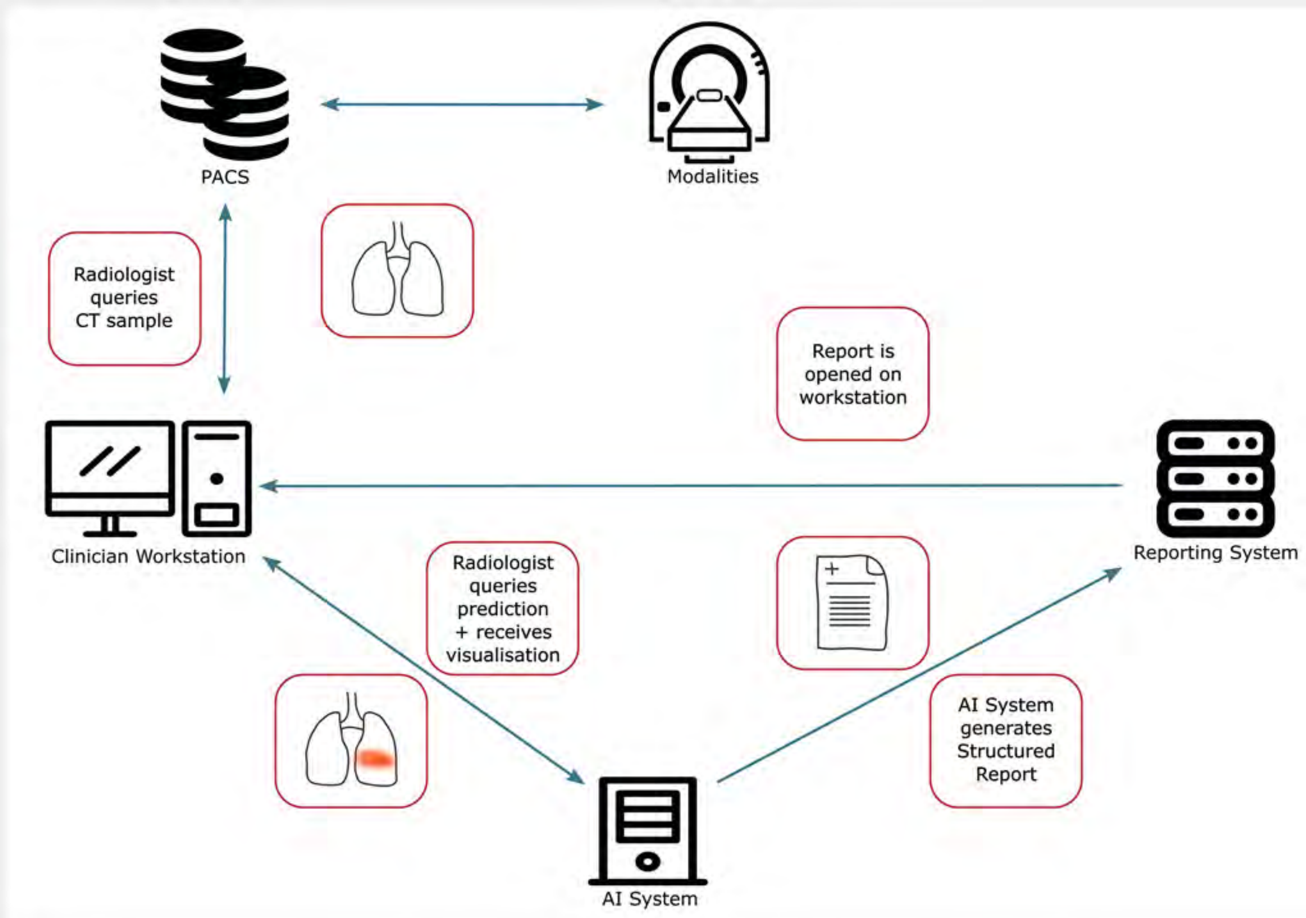
Therefore, we still offer the possibility of adding relevant observations in plain text. The unstructured text is essential for maintaining the daily clinical workflow, especially for sporadically occurring cases, as the information would otherwise be lost. However, text is initially hard to process for any AI system. Future AI systems could integrate this lost information with the help of advanced large language models. An example of a fictional patient being opened for reporting is shown in ▶ **Fig. 2**.

## Integration and Workflow for Lifelong Learning

To integrate the AI images into the radiology workflow, we run our model automatically on thorax CT examinations and generate possible predictions for PE location. We then extract the images and forward them as an additional modality. An overview of the integration is illustrated in ▶ **Fig. 3**.

Reporting on a workstation is quickly started by clicking on the selected study, which opens the reporting platform. The platform

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.




▶ **Fig. 3** Overview of the workflow that integrates the AI system into the reporting systems. When a new CT sample is generated, the AI system preprocesses the image to pre-fill the structured report and provides insight via the images. With a single press of a button, the reporting template with the images can be opened, and the report can be finalized. Finally, the information gained can be used to improve the AI model during later training episodes.

▶ **Abb. 3** Überblick über den Arbeitsablauf, der die KI- und Krankenhaussysteme integriert. Wenn eine neue Thorax CT-Probe generiert wird, verarbeitet das KI-System das Bild vor, um den strukturierten Befund vorauszufüllen. Die Visualisierungen bieten Einblicke in die KI und können die Befundung beschleunigen. Mit einem Tastendruck wird die Befundvorlage mit den Visualisierungen geöffnet und der Befund kann ergänzt und verbessert werden. Schließlich können die gewonnenen Informationen dazu verwendet werden, das KI-System in späteren Trainingsepisoden zu verbessern.

consists of two parts. First, the study is displayed on the left half of the screen. The clinician can choose between the standard CT scan or the CT with the image of the prediction results in the left bar. Second, the template is opened on the right side, enabling structured reporting for the selected task according to appropriate guidelines. The AI system pre-fills the reporting template for the learned tasks. Additionally, it offers further information, such as predicted probability and images of the relevant slices for the prediction. The radiologist can use this information to accelerate the report's completion and avoid missing findings, e. g., in the case of multiple PEs.

After the radiologist completes the report, the new ground truth annotation can be fed back to the AI model for further improvement. This functionality opens up the possibility of adapting the AI system over time, which prevents the deterioration of model performance as time passes and data distribution changes [9, 10]. This loss in performance often goes unnoticed, as deep learning models report high confidence even for low-quality predictions. This is denoted as ***silent failure***. Allowing for model adaptation also makes it possible for the model to work in new hospital environments [9, 10, 16].

However, training models continuously introduces new risks that must be cautiously handled. For starters, care must be taken to avoid ***catastrophic forgetting*** [14]. In this phenomenon, the model's performance with respect to data from older distributions deteriorates significantly. The goal is to train a model that produces high-quality predictions for all data sources seen during the training process.

We address these two challenges, namely the inherent heterogeneity of data and catastrophic forgetting, in the following fashion.



This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

For the first problem, our structured reporting template provides a solution of structuring the output data for the AI system. While CT images vary from hospital to hospital as well as device to device, they are denoted in HU [32] and basic interoperability is given by the DICOM standard. However, this still might lead to a performance drop when the distribution of data shifts quite significantly, as seen for our red hospital in ▶ **Fig. 4**.

To achieve resiliency against catastrophic forgetting, we explore different popular continual learning methods. The first is *Rehearsal* [33], where we interleave samples (20 %) from the previous datasets into the present training. This approach produces good empirical results. However, GDPR guidelines [11] often do not allow the storage of patient studies, and even if they do, the studies could become unavailable later due to the *right to forget*. We compare these results to *Memory Aware Synapses (MAS)* [34]. In contrast to *Rehearsal*, MAS identifies the most important model parameters and prevents them from changing too much from their initial state. Therefore, MAS would be preferable under GDPR regulations.

## Results

In our analysis, we take the typical approach of starting with a state-of-the-art model pre-trained on a large, public, and heterogeneous dataset, namely the RSNA pulmonary embolism CT dataset [22] and the challenge-winning model [23]. We further collect cohorts from two German university clinics with our structured reporting template, generating annotations on a sample-wise level. The radiologist found the system easy-to-use and approved the transparency images. We examine the situation where we first fine-tune the model using the first clinic data, then the second clinic data. We aim to obtain a model that performs well across all test sets and training orders. While the RSNA dataset enables us to obtain a pre-trained model for the AI system, we adapt the final classifier to the structure of the reporting template.

Our results are displayed in ▶ **Fig. 4**. Each clinic training set consists of 694 samples, while the red clinic has 13.83 % PE-positive results and the blue clinic has 30.55 %. Each clinic's test dataset contains 86 examples, which we evaluate for PE detection and location in the form of localization labels consisting of central, lobar, seg., and subseg. embolisms. We first evaluate the latest dataset training based on the test set in red and then the previously trained clinic in blue.

The results show how the simple rehearsal approach to training the AI system leads to consistent performance with respect to PE detection (avg. accuracy of 75.85 %) for both datasets. The method causes a slight loss in ability to adapt to the data distribution of the latest clinic, e. g., rehearsal top row in red. Overall, the rehearsal method performs well for detecting and localizing pulmonary embolisms. However, it requires the storage of samples, which may become problematic with GDPR standards. On the other hand, MAS can keep up for the most part (avg. accuracy of 75.28 %) while reducing catastrophic forgetting and maintaining the ability to adapt to the new clinic's data.

## Discussion

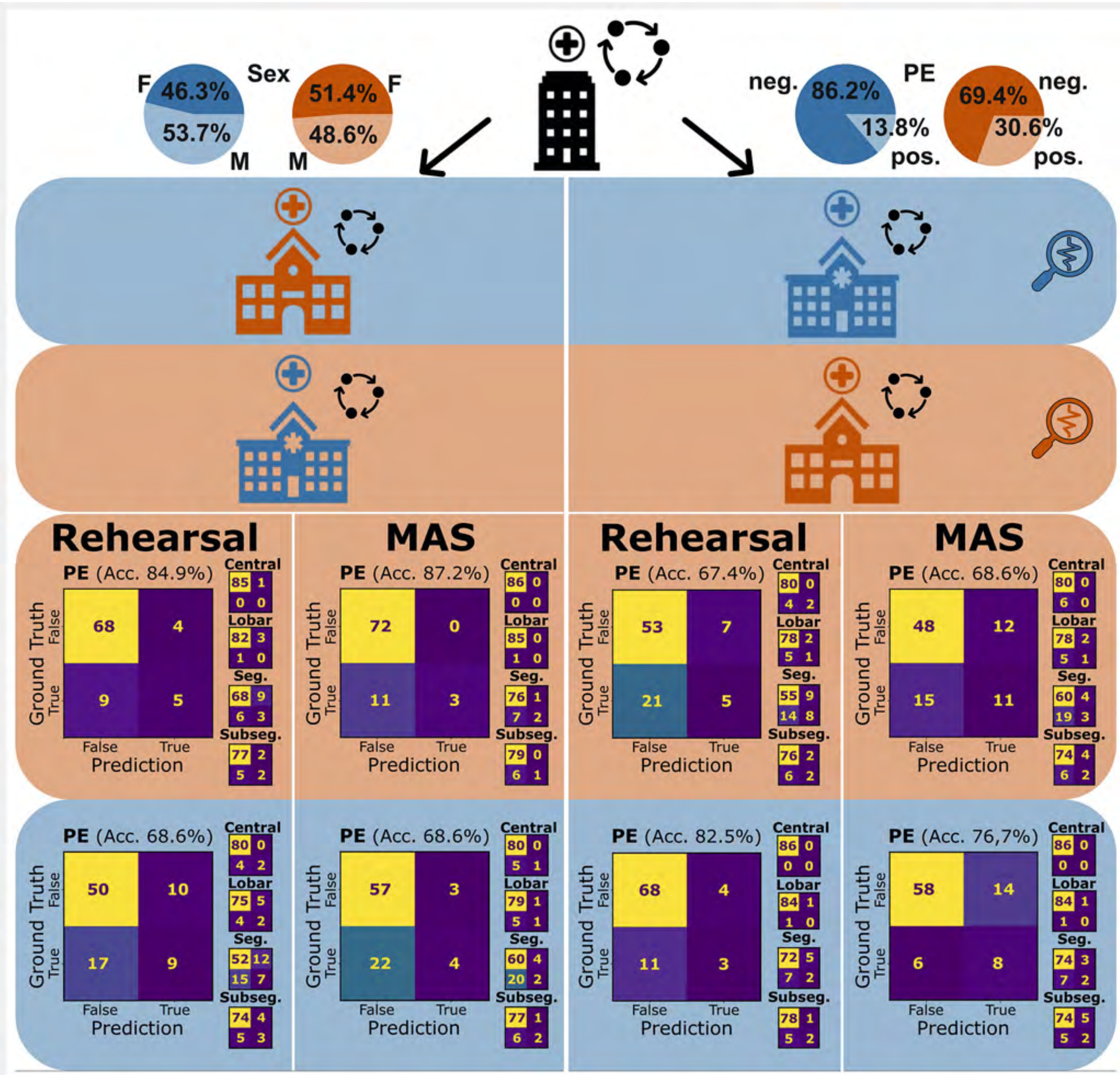
Reflecting on the results of our study, we find that the rehearsal approach achieves consistent performance on both datasets but lacks the ability to adapt to new data distributions. MAS instead performs similarly well in terms of reducing *catastrophic forgetting* and is more adaptable to new data. These findings are consistent with previous studies on continual learning [34]. Furthermore, rehearsal requires storage of samples and may not comply with GDPR standards, while MAS does not have these issues. Overall, both methods have their advantages and limitations.

Despite the urgent need for lifelong learning AI systems for radiology, the current legislative guidelines for medical devices do not yet provide an acceptable framework for quick and effective model updating. This requires the AI system provider to obtain approval whenever the system is deemed outdated [35, 36]. In practice, model updates – which incorporate new knowledge on acquisition practices and changes in the population – only become effective after a lengthy reverification cycle. Also, local fine-tuning steps where the model is adapted to specific characteristics of the data on-site become untenable.

We firmly believe that lifelong learning should be accompanied by monitoring of the model performance and annotations that are made. However, the current regulatory framework needs to include this significant opportunity for promoting the safe and effective use of AI. A potential solution to the current situation is a *pre-certification approach* [37, 38] that consists of a change control plan and predefined development and monitoring practices for the manufacturer to develop and update their devices safely and effectively rather than approving each individual update. However, we see two factors that must be improved upon for such an approach to succeed. Firstly, a closer collaboration is needed between regulators, device manufacturers, academic researchers, and other stakeholders to develop new strategies and guidelines for lifelong learning medical devices. The second important factor is increased transparency in AI systems, data acquisition, and the evaluation process. Transparency can be achieved by providing interpretable explanations and open access to the codebase, which would allow for building trust and understanding with experts and other stakeholders without regard to their background knowledge. As current plans by the FDA suggest [39], it would be risky to leave the design and evaluation of the development and monitoring practices solely to one party. When measurements and models can be updated at the same time, the risk for metric manipulation would be high to better suit marketing strategies or avoid direct comparisons with competitors. This highlights the importance of accountability and comparison of similar medical devices.

We show with our results that lifelong learning is beneficial – and indeed *needed* – for maintaining high predictive standards through the product lifecycle. Structured reporting allows the seamless integration of expert feedback into the learning loop. Giving radiologists insight into the AI system in the form of appropriate images provides a second layer of reliability to increase trust in our system.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.




▶ **Fig. 4** Comparison of Memory Aware Synapses (MAS) [34] and Rehearsal [33] approaches for lifelong learning. We first pre-trained our AI system to predict PEs using the public RSNA dataset. Then we deployed the AI system in a first clinic, followed by a second clinic and vice versa. We evaluated the latest dataset training using the test set in red and the previously trained clinic in blue. On average, Rehearsal achieves an accuracy (Acc.) for PE detection of 75.85 % and MAS 75.28 %. However, the performance between the two clinics varies. As we show at the top, this is not due to a bias regarding sex and is rather due to different PE occurrence rates and sizes.

▶ **Abb. 4** Der Vergleich von Memory Aware Synapses (MAS) [34] und Rehearsal [33] der beiden Ansätze in der Situation des lebenslangen Lernens. Wir haben unser KI-System zur LAE-Vorhersage auf dem öffentlichen RSNA-Datensatz vortrainiert. Als nächstes wird das KI-System in der ersten Klinik eingesetzt, gefolgt von der zweiten Klinik und umgekehrt. Wir evaluieren zunächst mit den Testdaten von der zuletzt trainierten Klinik in Rot. Anschließend wird die zuvor trainierte Klinik in Blau evaluiert. Im Mittel detektiert Rehearsal eine LAE mit einer Genauigkeit von 75,85 % und MAS mit 75,28 %. Die Leistung zwischen den Kliniken schwankt stark, ein Geschlechtsbias (oder Alter), wie in dem Tortendiagramm zu entnehmen, ist womöglich nicht vorhanden. Die Unterschiede sind am ehesten auf unterschiedliche Auftretensraten und Größen von Lungenembolien zurückzuführen.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

## Conclusion

We present an easy-to-integrate workflow for lifelong learning that leverages advances in structured reporting and interpretability. Our approach builds on the vertical integration of AI-ready radiology with a deep learning system, which requires two-way communication between both parties. We incorporate reliability measurements, namely prediction probabilities for labels and visualization to deliver dependable insight into the predictions of an AI system. Cooperating radiologists found our approach to be an easy-to-use system that facilitates lifelong learning. Furthermore, we discuss potential regulatory changes to improve the applicability of lifelong learning algorithms. Finally, we advocate for better integration of AI in radiology departments and closer collaboration between AI systems and clinicians.

### Funding

Bundesministerium für Gesundheit (EVA-KI [ZMVI1-2520DAT03A])

### Conflict of Interest

The authors of this manuscript declare relationships with the following companies: Phillip Matthies is an Employee of Smart Reporting GmbH.

### Acknowledgements

This work was supported by the Bundesministerium für Gesundheit (BMG) with the grant [ZMVI1–2520DAT03A]. Further, we would like to thank Zhangjie Yang, Mayur Arvind, Melda Eski and Jonathan Stieber for their involvement in earlier project stages.

### References

[1] FDA. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. 2021. Im Internet (Stand: 02.02.2023): https://www.fda.gov/media/145022/download

[2] Benjamin M, Aisen A, Benjamin E. Accelerating development and clinical deployment of diagnostic imaging artificial intelligence. Journal of the American College of Radiology 2021; 18: 1514–1516. doi:10.1016/j.jacr.2021.09.022

[3] Zhang J, Chao H, Dasegowda G et al. Overlooked Trustworthiness of Saliency Maps. In: Springer; 2022: 451–461

[4] LaRosa E, Danks D. Impacts on trust of healthcare AI. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. New Orleans LA USA: ACM; 2018: 210–215

[5] Hatherley JJ. Limits of trust in medical AI. Journal of medical ethics 2020; 46: 478–481. doi:10.1136/medethics-2019-105935

[6] Samarasinghe G, Jameson M, Vinod S et al. Deep learning for segmentation in radiation therapy planning: a review. Journal of Medical Imaging and Radiation Oncology 2021; 65: 578–595. doi:10.1111/1754-9485.13286

[7] Nazir M, Shakil S, Khurshid K. Role of deep learning in brain tumor detection and classification (2015 to 2020): A review. Computerized Medical Imaging and Graphics 2021; 91: 101940. doi:10.1016/j.compmedimag.2021.101940

[8] Zhou SK, Greenspan H, Davatzikos C et al. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE 2021; 109: 820–838. doi:10.1109/JPROC.2021.3054390

[9] Sanner A, Gonzalez C, Mukhopadhyay A. How reliable are out-of-distribution generalization methods for medical image segmentation? In: Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021, Bonn, Germany, September 28–October 1, 2021, Proceedings. Springer; 2022: 604–617

[10] Perkonigg M, Hofmanninger J, Herold Christian J et al. Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. Nat Commun 2021; 12: 5678. doi:10.1038/s41467-021-25858-z

[11] Fuchs M, Gonzalez C, Mukhopadhyay A. Practical uncertainty quantification for brain tumor segmentation. In: International Conference on Medical Imaging with Deep Learning. PMLR; 2022: 407–422

[12] Jospin LV, Laga H, Boussaid F et al. Hands-on Bayesian Neural Networks – a Tutorial for Deep Learning Users. IEEE Comput Intell Mag 2022; 17: 29–48. doi:10.1109/MCI.2022.3155327

[13] Elskhawy A, Lisowska A, Keicher M et al. Continual Class Incremental Learning for CT Thoracic Segmentation. In: Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2. Springer; 2020: 106–116

[14] Mundt M, Lang S, Delfosse Q et al. CLEVA-compass: A continual learning evaluation assessment compass to promote research transparency and comparability. arXiv preprint arXiv:211003331. 2021

[15] EU. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). 2016

[16] Gonzalez C, Gotkowski K, Bucher A et al. Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24. Springer; 2021: 304–314

[17] Jussupow E, Spohrer K, Heinzl A et al. Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. Information Systems Research 2021; 32: 713–735. doi:10.1287/isre.2020.0980

[18] Arun N, Gaw N, Singh P et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artificial Intelligence 2021; 3: e200267. doi:10.1148/ryai.2021200267

[19] Adebayo J, Gilmer J, Muelly M et al. Sanity checks for saliency maps. Advances in neural information processing systems; 2018: 31

[20] Kim B, Seo J, Jeon S et al. Why are saliency maps noisy? cause of and solution to noisy saliency maps. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE; 2019: 4149–4157

[21] Alqaraawi A, Schuessler M, Weiß P et al. Evaluating saliency map explanations for convolutional neural networks: a user study. In: Proceedings of the 25th International Conference on Intelligent User Interfaces.2020: 275–285

[22] Colak E, Kitamura FC, Hobbs SB et al. The RSNA pulmonary embolism CT dataset. Radiology: Artificial Intelligence 2021; 3: e200254. doi:10.1148/ryai.2021200254

[23] XU G. RSNA STR Pulmonary Embolism Detection: 1st place solution with code. RSNA STR Pulmonary Embolism Detection: 1st place solution with code 2020. Im Internet (Stand: 01.02.2022): https://www.kaggle.com/competitions/rsna-str-pulmonary-embolism-detection/discussion/194145?focusReplyOnRender=true

[24] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition.2018: 7132–7141

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

[25] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation 1997; 9: 1735–1780. doi:10.1162/neco.1997.9.8.1735

[26] Szeliski R. Computer vision: algorithms and applications; Springer Nature; 2022

[27] Gotkowski K, Gonzalez C, Bucher A et al. M3d-CAM: A PyTorch library to generate 3D attention maps for medical deep learning. Bildverarbeitung für die Medizin 2021: Proceedings, German Workshop on Medical Image Computing, Regensburg, March 7-9, 2021; 2021: 217–222. doi:10.1007/978-3-658-33198-6_52

[28] Chattopadhay A, Sarkar A, Howlader P et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). IEEE; 2018: 839–847

[29] Sabel BO, Plum JL, Kneidinger N et al. Structured reporting of CT examinations in acute pulmonary embolism. Journal of Cardiovascular Computed Tomography 2017; 11: 188–195. doi:10.1016/j.jcct.2017.02.008

[30] Gilman MD, Kazerooni EA. Standardized Reporting of CT Pulmonary Angiography for Acute Pulmonary Embolism. Individual or Group PQI; 2015

[31] Ziegler E, Urban T, Brown D et al. Open Health Imaging Foundation Viewer: An Extensible Open-Source Framework for Building Web-Based Imaging Applications to Support Cancer Research. JCO Clinical Cancer Informatics 2020: 336–-345. doi:10.1200/CCI.19.00131

[32] Schneider U, Pedroni E, Lomax A. The calibration of CT Hounsfield units for radiotherapy treatment planning. Physics in Medicine & Biology 1996; 41: 111–124. doi:10.1088/0031-9155/41/1/009

[33] Rebuffi S-A, Kolesnikov A, Sperl G et al. iCaRL: Incremental Classifier and Representation Learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE; 2017: 5533–5542

[34] Aljundi R, Babiloni F, Elhoseiny M et al. Memory Aware Synapses: Learning What (not) to Forget. In: Ferrari V, Hebert M, Sminchisescu C, et al., Hrsg. Computer Vision – ECCV 2018. Cham: Springer International Publishing; 2018: 144–161

[35] Vokinger KN, Feuerriegel S, Kesselheim AS. Continual learning in medical devices: FDA's action plan and beyond. The Lancet Digital Health 2021; 3: e337–e338. doi:10.1016/S2589-7500(21)00076-5

[36] Vokinger KN, Gasser U. Regulating AI in medicine in the United States and Europe. Nature machine intelligence 2021; 3: 738–739. doi:10.1038/s42256-021-00386-z

[37] FDA Developing a software precertification program: A working model. Pre-Cert Working Model Version 10. 2019. Im Internet (Stand: 02.02.2023): https://www.fda.gov/media/119722/download

[38] Bartlett VL, Dhruva SS, Shah ND et al. Clinical studies sponsored by digital health companies participating in the FDA's Precertification Pilot Program: A cross-sectional analysis. Clinical Trials 2022; 19: 119–122. doi:10.1177/17407745211048493

[39] FDA. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions – Draft Guidance for Industry and Food and Drug Administration Staff 2023. Im Internet (Stand: 02.05.2023): https://www.fda.gov/media/166704/download


This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.