Computerised emergency work-up for mass-like brain MRI lesions: can explainable AI support radiologists?

Commentary on EURA-D-22-03305

Word count: 983

Intracranial intra-axial brain mass-like lesions have a wide differential diagnosis including malignant and non-malignant (tumefactive) disorders. Despite advances in the application of multiparametric MRI [1, 2], determining their cause can be challenging. Accurate diagnosis of tumorous versus tumefactive lesions supports the effective work-up of such patients, relying on the differentiation between the malignant and non-malignant lesions. Traditionally, this distinction has been reliant on visual interpretation by experienced (neuro)radiologists, a process that even in experienced hands remains subjective. An automated and objective approach that would help differentiate between non-malignant and malignant mass-like lesions would facilitate patient referral.

In the current issue of *European Radiology*, Shin and colleagues [3] tested a three-stage, deep learning-based approach to differentiate between malignant and non-malignant intra-axial mass-like lesions to determine an automated referral suggestion for patient presenting in an emergency-room setting. At each stage, deep neural networks make use of multi-parametric MRI scans of patients with a range of diseases to reach the final referral suggestion. In detail, the authors initially trained a U-NET-based lesion segmentation network on co-registered, post-contrast T1-weighted (T1CE) and Fluid-Attenuation-Recovery (FLAIR) sequences to create segmentation maps of contrast-enhancing lesions (CEL), non-enhancing FLAIR hyperintense lesions (NEL), the necrotic portion and the non-lesion background. These segmentation maps were afterwards combined with co-registered T1CE, FLAIR, pre-contrast T1-weighted (T1W1), diffusion weighted-imaging (DWI) images and apparent diffusion coefficient (ADC) maps. Following a hierarchical classification approach, two independent convolutional networks were trained to separate patient scans into malignant or non-malignant cases, and then provide a referral suggestion - either surgery or systemic work-up for patients with malignant lesions; medical treatment or conservative management for patients with non-malignant lesions. Using layer-wise-propagation [4], a post-training explainability technique that was applied on the tumour vs. non-tumour classification network, pixel-based heatmaps of relevance were generated to offer a visual representation of significant locations in the MR sequences that mostly affected the network decision.

Networks were trained on a dataset of 747 patients (mean age 53.59 years, 328 females) with approximately 75% malignant lesions (mean age 54.38 years, 263 females) and 25% non-malignant lesions (mean age 51.16 years, 65 females). Clinical validation was performed in an independent cohort of 130 patients (mean age 57.57 years, 67 females) as external test set with approximately 70% malignant lesions (mean age 59.65 years, 41 females) and 30% non-malignant lesions (mean age 52.53 years, 26 females) . The performance of the classification networks was compared to the performance of two expert neuroradiologists and four radiology residents. The average classification performance of the neural network in discriminating between malignant and non-malignant lesions and in making clinical referral suggestions on test set patients was on par with the performance of human readers. The primary convolutional network (accuracy of 87.7%) slightly outperformed both groups of readers (neuroradiologists accuracy 87.3% and radiology residents accuracy 87.5%) in

classifying patients with malignant versus non-malignant lesions, while the Area-Under-the-Curve value was 0.90 for the neural network and 0.85 on average for all human readers. The clinical referral suggestion network had an accuracy (72.3%) lying in between the performance of the radiology residents (70.2%) and neuroradiologists (77.3%). It is critical to observe the accuracy performance of radiology resident Reader 4 in the two tasks, who significantly underperformed. Mean human reader performance was decreased and brought closer to the performance of the automated method.

Thus far, the closest study to this work, proposing an automated diagnosis of brain lesion types has been the study by Rauscheker and colleagues [5]. By mimicking the way radiologists tackle such tasks and exploiting prior knowledge in the extraction of carefully selected features, they were able to set the ground for the employment of an automated diagnostic model. The approach of Rauscheker and colleagues depends on *a priori* selected segmentation features fed to a network, while the study of Shin et al. used all available raw MR images. The latter offers an end-to-end automated pipeline for the referral suggestion but opens this black box by the addition of an explainable component. With information being made available from each stage of the pipeline, i.e. the segmentation maps, classification decisions and relevance heatmaps, the user is able to understand the "reasoning" behind the computerised decision-making. An additional advantage of the approach by Shin et al. is the ability to rapidly extract neural-network-derived output, thus avoiding a time-consuming manual decision-making process. The individual network classification performance suggests that it is possible to establish a method that ~~will~~ might outperform the average expert human reader. This cannot be easily achieved though. Currently, AI-based models used in medicine require expert input and are far from being exclusively trusted. Thus, the ideal next step towards establishing such approaches should be a hybrid setup, where automated models are combined with decisions made by practitioners and assessed as supporting computerised tools rather than an autonomous replacement of the human readers.

This study does come with certain limitations. External testing was performed using scans from a single cohort, while many sub-categories of lesions present in the training set were absent from the test set, posing the question of how well the suggested methodology will be able to generalise to multi-site cohorts and specific lesion subtypes. For example, the Rauschecker study [5] encompassed a much wider range of pathology in the differential diagnosis, including abscesses, leukodystrophy and others. Further research will be needed to improve the performance of the AI model, starting from the addition of multi-site cohorts to add external validity to the study. Ideally, subgroups of disease categories which are not seen on the training cohort can be included, to simulate a real-world setup and further test the generalisation capabilities of the classification networks. To determine the true value of this study, additional work would be needed, evaluating the efficacy of the automated method, most importantly as a supporting tool for radiologists – the most likely clinical implementation scenario.

A realistic assessment of the current landscape shows us that the gap between automated diagnostic models and human readers is narrowing but still visible: neither the method of Shin and colleagues nor any other related studies prove that we can solely rely on computerised models to diagnose mass-like lesion types. What this work does prove, is that end-to-end models can be both interpretable and approach the performance of radiology residents and even fellow-level decisions. We can only be encouraged by these findings and seek further technical development that will offer the best possible computerised radiology assistance.

References

1. Hirschler L, Sollmann N, Schmitz-Abecassis B, et al. Advanced MR Techniques for Preoperative Glioma Characterization: Part 1. J Magn Reson Imaging. 2023;10.1002/jmri.28662. doi:10.1002/jmri.28662

2. Hangel G, Schmitz-Abecassis B, Sollmann N, et al. Advanced MR Techniques for Preoperative Glioma Characterization: Part 2. J Magn Reson Imaging. 2023;10.1002/jmri.28663. doi:10.1002/jmri.28663

3. Shin H, Park JI, Jun Y, Eo T, Lee Y, Kim JE, Lee DH, Moon HH, Park SI, Kim S, Hwang D, Kim HS. Deep Learning Referral Suggestion and Tumour Discrimination using Explainable Artificial Intelligence applied to Multiparametric MRI. European Radiology; ***this issue***

4. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS One. 2015;10(7):e0130140. doi:10.1371/journal.pone.0130140

5. Rauschecker AM, Rudie JD, Xie L, et al. Artificial Intelligence System Approaching Neuroradiologist-level Differential Diagnosis Accuracy at Brain MRI. Radiology. 2020;295(3):626-637. doi:10.1148/radiol.2020190283