MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231099

Explainable Artificial Intelligence for Deep-Learning Based Classification of Cystic Fibrosis Lung Changes in MRI

Friedemann G. RINGWALD[a,1], Anna MARTYNOVA[a], Julian MIERISCH[a,], Mark WIELPÜTZ[b] and Urs EISENMANN[a]
[a] *Institute of Medical Informatics, Heidelberg University Hospital, Germany*
[b] *Diagnostic and Interventional Radiology, Heidelberg University Hospital, Germany*
ORCiD ID: Friedemann G. Ringwald https://orcid.org/0000-0002-0572-4820

Abstract. Algorithms increasing the transparence and explain ability of neural networks are gaining more popularity. Applying them to custom neural network architectures and complex medical problems remains challenging. In this work, several algorithms such as integrated gradients and grad came were used to generate additional explainable outputs for the classification of lung perfusion changes and mucus plugging in cystic fibrosis patients on MRI. The algorithms are applied on top of an already existing deep learning-based classification pipeline. From six explain ability algorithms, four were implemented successfully and one yielded satisfactory results which might provide support to the radiologist. It was evident, that the areas relevant for the classification were highlighted, thus emphasizing the applicability of deep learning for classification of lung changes in CF patients. Using explainable concepts with deep learning could improve confidence of clinicians towards deep learning and introduction of more diagnostic decision support systems.

Keywords. Deep learning, explainable artificial intelligence, cystic fibrosis, magnetic resonance imaging

1. Introduction

For assessing lung changes of cystic fibrosis (CF) patients, magnetic resonance imaging (MRI) was introduced for interventional trials and routine imaging. MRI proved to be comparable to computed tomography (CT) imaging in many aspects, with the advantage of not using ionizing radiation and the possibility to rely on functional MRIs [1]. For systematic MRI evaluation a scoring system was developed. Based on six independent items, the morpho-functional score can be captured. With this score, reproducible results can be achieved to evaluate lungs of cystic fibrosis patients [2]. In order to speed up the scoring and decrease reader variability, a deep learning approach supporting radiologists is currently under development. Preliminary results have already been produced for two scoring items: lung perfusion and mucus plugging. While the underlying deep learning architecture still needs more training data, fine-

[1] Corresponding Author: Friedemann Ringwald, email: Friedemann.ringwald@med.uni-heidelberg.de

tuning and overall improvement, the results have shown the need for a component of explain ability. Since the central task is classification and not segmentation, the deep learning pipeline currently produces a numeric score for each lung half as prediction. Understanding this score is crucial for further improvement and increasing the trust of the radiologists in the deep learning approach. Receiving visual feedback on how the network selected the final decision, could strengthen this trust [3]. In this work, it is investigated, whether explainable artificial intelligence (XAI) algorithms can support the deep learning classifications on lung MRI of CF patients.

2. Methods

For the selection of fitting explainable artificial intelligence (XAI) algorithms, possible candidates were systematically reviewed [4]. Inclusion criteria were: 1) Suitability for medical purposes: While some algorithms are only suitable for language processing, in our use-case it is crucial that images can be used as input. 2) Existing implementation in PyTorch [5]. The classification pipeline on which the algorithms are applied is designed in PyTorch. To reduce the complexity and enable straight forward usage, the XAI implementation should also be implemented in PyTorch. 3) Applicability to a custom neural network architecture: Due to the architecture of the principal classification pipeline, which was customized to fit a multi-view, slice-based classification of MRIs [6], the XAI method is required to be adaptable to fit to the problem. 4) Post-Hoc algorithm: Post-hoc models can be applied to already trained models and are therefore prerequisite for this work [7].

After successful implementation and testing, each XAI algorithm is applied on a trained neural network and should provide an added value for the radiologists. Since it is not always tangible for a human reader, which visual features influence the classification outcome of a neural network for a patient, the results with the XAI should support the understanding.

To properly quantify the added value of the resulting visualizations, an analysis is conducted to retrieve a quantifiable result. In detail, all slices of one MRI examination are inspected and afterwards, a category was assigned to the MRI. The following four categories were available:

- Category 1: Maximum two slices of the MRI sequence had incorrect overlays.
- Category 2: Three or more slices but not all slices of the MRI sequence had incorrect overlays.
- Category 3: All slices had incorrect overlays or it was unclear which pixels had overlays.
- Category 4: Neutral, nothing is highlighted.

All algorithms were applied for the two classification problems lung perfusion defects and mucus plugging, which are part of the MRI-score. Mucus plugging is usually evaluated by the radiologists looking at the T2 BLADE sequence. Lung perfusion defects can only be made visible via a functional MRI with contrast agent injection and a subsequent subtraction of images over time to acquire optimal exposure and contrast agent distribution. This heavily affects the image quality and introduces blurring on the subtraction images.

3. Results

In total, six different algorithms passed the selection criteria and four could be implemented successfully. Those were: integrated gradients (IG), class activation maps (CAM), grad cam (GCam) and layer-wise relevance propagation (LRP). Calculation time for each algorithm was under three seconds for the entire stack of slices of each MR, depending on the lung size between 20 and 80 slices. The algorithms use the image information, together with the predicted score as a starting point for attribution calculations. The attribution results are presented as attribution maps. Green areas on an attribution map allude a higher and red areas a lower accuracy score. Meaning that regions that are in favor of the targeted class are painted green, while regions which indicate a misclassification are red. Altogether, the outputs for the CAMs looked best and were evaluated as most in line with the visual features the radiologists use for evaluation.

This was also in line with the categorical analysis displayed in Table 1. Due to time constraints, the categorical analysis was only evaluated on the lung perfusion defect dataset and not for mucus plugging. In this analysis, all slices of one examination were visually inspected and assigned to one of the four categories mentioned above. More than every fifth MRI with CAM was sorted into the first category and almost half of all MRIs into the fourth category. For GCam, no MRIs reached category one, and the majority of MR examinations could be found in category two and three.

Table 1. Comparison of CAM and GCAM results for perfusion defects. The numbers correspond to the percentual values of MRIs corresponding to the different categories.

	Category 1	Category 2	Category 3	Category 4
CAM	20,14%	26,12%	8,20%	45,54%
GCam	-	44,03%	34,33%	21,64%

Some example slices are shown in Figure 1. In the top row, the raw slices without overlay are displayed. In the second row, the slices are shown with the attribution map overlay produced by the CAMs and which are considered good results (category A). In the third row, exemplary slices are shown, were no good overlay was created (category C and D). Slices were defined as unsatisfactory results if areas, not relevant for classification (e.g Figure 1, F, with highlights on the spinal cord) were marked red. In general, all XAI algorithms seemed to be working better on the MRIs used for determining mucus plugging severity.

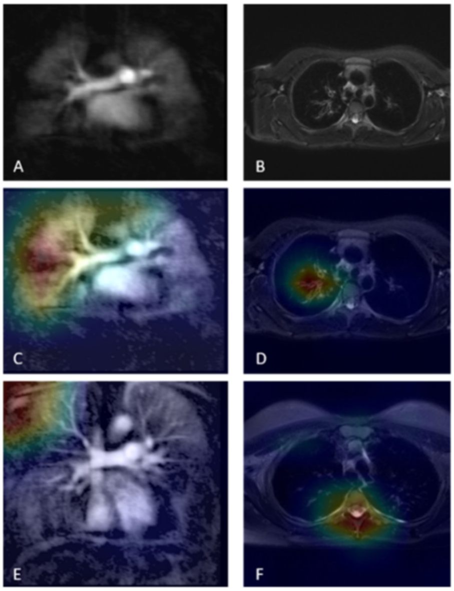

Figure 1. Selection of lung slices for the classification of lung perfusion defects (left) and mucus plugging (right). The top row (A, B) shows the raw MR slices. The second row (C, D) shows good attribution maps produced by CAMs. The last row (E, F) show incorrect attribution maps results.

4. Discussion

Solving medical image classification tasks with deep learning is producing increasingly better results. The availability of more training data and less computational restrictions due to high performant GPUs is speeding this process up even more. Nonetheless missing trust towards these black-box algorithms prevails. Properly understanding the training data and utilizing XAI methods are useful tools to help increase confidence. In this work, from six explainable artificial intelligence methods, CAMs produced the best visual overlay to help the radiologist in detecting lung perfusion defects and mucus plugging in CF patients. In terms of development of XAI methods, it was expected, that the GCAM would produce better results than CAM, since it is an advancement of CAM. Surprisingly, this was not the case and needs further investigation. Regarding the categorical analysis, many of the MRIs were classified into category four, corresponding to neutral. This was the case for many pathology free MRIs, which raises the question, how the overlay should look if no lung change is detected. Generally, the results for mucus plugging were better, which could be explained by the underlying image quality. The introduced blur by the subtraction imaging for lung perfusion might affect the work of the XAI algorithms. Furthermore, if the classification pipeline is producing incorrect outputs, the XAI will not produce satisfying outputs either. Testing CAM on other MRI sequences or using all planes (axial, coronal, sagittal) could improve the results.

5. Conclusions

In conclusion, the application of XAI algorithms proved to provide additional useful information for the radiologist when scoring lung MRIs of CF patients with lung perfusion defects or mucus plugging. It was evident, that the areas relevant for the classification were highlighted, thus emphasizing the applicability of deep learning for classification of lung changes in CF patients. Furthermore, in case the algorithm highlighted incorrect or irrelevant areas, the classification result should be treated with caution and questioned and revisited manually.

References

[1] Puderbach M, Eichinger M, Haeselbarth J, Ley S, Kopp-Schneider A, Tuengerthal S, Schmaehl A, Fink C, Plathow C, Wiebel M, Demirakca S. Assessment of morphological MRI for pulmonary changes in cystic fibrosis (CF) patients: comparison to thin-section CT and chest x-ray. Invest Radiol. 2007 Oct;42(10):715-24, doi:10.1097/RLI.0b013e318074fd81.

[2] Eichinger M, Optazaite DE, Kopp-Schneider A, Hintze C, Biederer J, Niemann A, Mall MA, Wielpütz MO, Kauczor HU, Puderbach M. Morphologic and functional scoring of cystic fibrosis lung disease using MRI. Eur J Radiol. 2012 Jun;81(6):1321-9, doi:10.1016/j.ejrad.2011.02.045.

[3] Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Trans Neural Netw Learn Syst. 2020 Oct;32(11):4793-813, doi:10.1109/TNNLS.2020.3027314.

[4] Van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. 2022 May;79:102470, doi:10.1016/j.media.2022.102470.

[5] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems 32 [Internet]. Curran Associates, Inc.; 2019. p. 8024–35. Available from: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[6] Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, Bereket M, Patel BN, Yeom KW, Shpanskaya K, Halabi S. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Med. 2018 Nov;15(11):e1002699, doi:10.1371/journal.pmed.1002699.

[7] He K, Zhang X, Ren S, Sun J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Trans Pattern Anal Mach Intell. 2015 Sep;37(9):1904-16, doi: 10.1109/TPAMI.2015.2389824.