A review of XAI approaches for computer vision learning models in healthcare

Simon Dreyer Vetter
ec22175
ec22175@qmul.ac.uk

***Abstract* — Artificial intelligence (AI) has been widely adopted in medical imaging for tasks such as organ segmentation, lesion, and tumour detection. However, the lack of transparency in AI models has raised concerns among medical professionals regarding their use. Explainable AI (XAI) aims to improve the transparency and reliability of AI models by identifying the features that most strongly influence their predictions. In this review, we present and compare various XAI approaches for computer vision methods in healthcare.**

I. Introduction

XAI or Explainable Artificial Intelligence is a growing subfield of machine learning aiming to make intelligent systems more transparent and understandable. It helps to build trust in artificial intelligence (AI) systems that are notoriously abstract and hard to interpret. Additionally, using XAI can help identify potential biases or errors in the systems, allowing them to be corrected. In this review article, we will present different studies related to XAI techniques in the context of computer vision learning models in healthcare studies. Furthermore, we will highlight different approaches and potential shortcomings in the reviewed papers. Related works in this area often concern ethics and the need for explanations, but here we present current applications of XAI.

II. Problem Definition

Black box AI systems that make predictions without providing any justification are problematic for a number of reasons. The lack of transparency and the fact that they may conceal internal biases raises distrust and scepticism of the use of such technologies, notably in the medical field where wrong predictions may have fatal outcomes. In order for medical professionals to employ such sophisticated AI systems for i.e., medical education, research and clinical decision-making, it is crucial that the medical professionals have the capacity to comprehend and trace the machine decision-making process. The objective of this paper is to explore how XAI methods have been included alongside AI models in an attempt to develop a better understanding and interpretation of the system.

III. Key Works

Following the surge in the popularity of XAI, a plethora of different explainable systems have been developed and researched. Due to the scope of this project, only a few but largely used XAI systems will be mentioned here. Firstly, Riberio et al [5]. proposed Local Interpretable Model-Agnostic Explanations (LIME) which provide interpretable and trustworthy explainability of classifier predictions. Oversimplified, LIME does this by training a local surrogate model which approximates the predictions of the target system. The dataset that is used consists of perturbed versions of the input data and the corresponding predictions of the black box model. Furthermore, layer-wise relevance propagation (LRP) was introduced in 2015. This framework allows decomposing the prediction of a deep neural network computed over a sample, down to relevance scores for the single input dimensions of the sample such as subpixels of an image [3]. Lastly, a more modern method Grad-Cam was proposed to provide a better explainable result for mainly convolutional neural networks (CNN) model families. This method uses the gradient of the target class with respect to the activations of a convolutional layer to produce a weighted feature map which highlights the regions of the input image that are most important for the prediction [1].

IV. Evaluation Criteria

The articles reviewed in the following section will be evaluated based on their approaches to XAI and the interpretations drawn from their results. All datasets used in these studies consist of images that are open-sourced and well-documented.

V. Discussion

Genovese et al. took the approach to use XAI not only for validation but furthermore to arrive at a better and more accurate model when detecting acute lymphoblastic leukaemia [1]. The authors test an adaptive unsharpening technique to pre-process microscopy images of white blood cells to attempt to increase the performance of the model. They use several state-of-the-art CNN models such as VGG16 and VGG19, DenseNet, AlexNet, and ResNet. Grad-CAM is used to visualise which regions of the image contribute more to the result of the classification.

By comparing the results of the Grad-CAM method on pairs of corresponding images from the two datasets (applied the unsharpening technique and not), it was observed that the images from the unsharp dataset had higher Grad-CAM activity in regions centred on the white cells. This indicated that the proposed unsharpening method allowed the CNN to learn more features on the details of the cells, and not the background, which again led to increased accuracy of the classification. One potential flaw in this analysis is that it is based on a subjective interpretation of the Grad-CAM results. Human interpretability is hard to define in general [8]. Different users may have different capabilities for reading explanations which could lead to variations in the conclusions drawn from the analysis. Although an expert in the field might be able to determine what constitutes a "more centred" or "more intense" region of a white blood cell, this might not be the case for every user.

In a subsequent study, the authors used both Grad-CAM and Grad-CAM++ to improve the reliability of a neural network performing segmentation and quantification of COVID-19 infections from CT images [6]. Resultingly, they found that the Grad-CAM class activations were concentrated on areas containing COVID-specific lesions, suggesting that the model was making positive predictions based on the identification of damaged tissue caused by COVID. Additionally, the authors concluded that Grad-CAM++ outperformed Grad-CAM, producing class activations that were more confined and closer to lesion segments. One of the

limitations of Grad-CAM is its potential for poor localisation of regions of interest. Grad-CAM++ addresses this issue by using a weighted average of gradients at each feature map, rather than the plain average used in the original method. This is because the gradient alone may not adequately capture the varying levels of "importance" among the units in each feature map, leading to potentially inadequate localisation of regions of interest [1].

Additionally, the authors of this paper managed to identify some major problems in the model by using XAI. During initial trials of unmodified CT images, they observed that Grad-CAM highlighted areas outside of the lungs which indicated that the model's decisions were based on details outside of the lungs. This led to them pre-processing their original images, extracting only the lungs which were the regions of interest. However, this paper also suffers from the issue of a subjective interpretation of the Grad-Cam results.

In a similar study by Pitroda et al. [7], this problem of subjectivity is tackled by quantifying the complexity and faithfulness of the XAI techniques they implement, namely: LRP, LIME, Guided Backpropagation (GB), and Deep Taylor Decomposition (DTD). A customised CNN is developed to classify lung diseases using chest X-ray images. Likewise, with the previous paper, heatmaps are used to visualise important areas or features of an image relevant to the prediction. Entropy is used to assess the complexity or randomness of the explanations. If a heatmap produced by an XAI technique contains a lot of irrelevant information for the prediction, it may be considered more complex because it contains more information that is not directly related to the task at hand. On the other hand, if an explanation contains only relevant information and is therefore able to effectively convey the reasoning behind a prediction made by a machine learning model, it may be considered less complex. The pixel flipping metric is also used to assess the "faithfulness" of the model. This is done by removing features highlighted in the explanation (as most relevant) and testing if it leads to a strong decay in the model's predictions [8]. The results show that the DTD and LRP methods perform better than the GB method, with DTD slightly outperforming LRP.

In retrospect, using entropy in order to assess the complexity of an explanation may not be the most insightful method. It is important to note that the complexity of an explanation is not necessarily related to the amount of information it contains. An explanation that contains a small amount of highly relevant information may be more effective at conveying the reasoning behind a prediction than an explanation that contains a large amount of information, some of which may be less relevant. Furthermore, the pixel-flipping method is a reliable method to represent the local decision structure of the model however, the results of the method are not discussed in much detail. It would be useful to provide more information about the specific criteria used to evaluate the explanations and how the results were interpreted.

Lastly, in a study by Hamza et al. [4], various CNN architectures were utilised to predict Alzheimer's disease in patients at different stages using magnetic resonance imaging (MRI) data. The authors employed the LIME method to assess the explainability of their model generating heatmaps from the results. Upon examining these heatmaps, we can observe that the model's predictions were based on large regions outside of the brain in the MRI images, which should not have any influence on the prediction. This should raise a concern about the model's performance. Additionally, analysis of the accuracy plots for the validation and training datasets revealed that all the models tested in the study were severely overfitted, a finding that the authors did not acknowledge. The best-achieved accuracy for classifying different stages of Alzheimer's disease is 86.82% whereas training accuracy is 95.16%. In this case, the visual results from the LIME method work as a cue for the deficiencies in the model's performance, which unfortunately the authors fail to point out.

VI. Conclusion

Medical professionals often serve as interpreters for patients, conveying complex information in a way that is comprehensible to them. It is therefore essential that AI tools used by these professionals are themselves interpretable and easily understood. This paper has examined various approaches to improve the transparency of learning models' decisions, including the use of heatmaps to visualise key areas of images relevant to a prediction. While this method can be useful for medical experts, it can also be prone to subjective interpretation. One study addressed this limitation by calculating metrics of the complexity and faithfulness of the explanations provided. Notably, establishing a standardized framework for the explainability of AI systems that can be used by medical professionals remains a difficult challenge. Different tasks and datasets may require different approaches, and improvements to current techniques are still being developed to make them more flexible. This might be a necessary condition before we see more widespread adoption of XAI in medical practice

References

[1] A. Chattopadhay, A. Sarkar, P. Howlader and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839-847, doi: 10.1109/WACV.2018.00097.

[2] A. Genovese, M. S. Hosseini, V. Piuri, K. N. Plataniotis and F. Scotti, "Acute Lymphoblastic Leukemia Detection Based on Adaptive Unsharpening and Deep Learning," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 1205-1209, doi: 10.1109/ICASSP39728.2021.9414362.

[3] Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W (2015) On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE 10(7): e0130140. https://doi.org/10.1371/journal.pone.0130140

[4] H. A. Shad et al., "Exploring Alzheimer's Disease Prediction with XAI in various Neural Network Models," TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON), 2021, pp. 720-725, doi: 10.1109/TENCON54134.2021.9707468.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," inProc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2016, pp. 1135–1144.

[6] N. Darapaneni et al., "Explainable Diagnosis, Lesion Segmentation and Quantification of COVID-19 Infection from CT Images Using Convolutional Neural Networks," 2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2022, pp. 0171-0178, doi: 10.1109/IEMCON56893.2022.9946520.

[7] V. Pitroda, M. M. Fouda and Z. M. Fadlullah, "An Explainable AI Model for Interpretable Lung Disease Classification," 2021 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS), 2021, pp. 98-103, doi: 10.1109/IoTaIS53735.2021.9628573.

[8] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders and K. -R. Müller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," in Proceedings of the IEEE, vol. 109, no. 3, pp. 247-278, March 2021, doi: 10.1109/JPROC.2021.306048