The Integration of Explainable AI Methods for the classification of Medical Image Data

Mina Nikolić
Computer Science Department
Faculty of Electronic Engineering
Niš, Serbia
mina.nikolic@elfak.ni.ac.rs & [0009-0009-9356-5935]

Dragan Janković
Computer Science Department
Faculty of Electronic Engineering
Niš, Serbia
dragan.jankovic@elfak.ni.ac.rs & [0000-0003-1198-0174]

Aleksandar Stanimirović
Computer Science Department
Faculty of Electronic Engineering
Niš, Serbia
aleksandar.stanimirovic@elfak.ni.ac.rs & [0000-0001-8772-4930]

Leonid Stoimenov
Computer Science Department
Faculty of Electronic Engineering
Niš, Serbia
leonid.stoimenov@elfak.ni.ac.rs & [0000-0003-1097-9833]

2024 11th International Conference on Electrical, Electronic and Computing Engineering (IcETRAN) | 979-8-3503-8699-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICETRAN62308.2024.10645095

***Abstract*— Deep convolutional neural network architectures have in recent years been widely used for enhancing various Computer vision tasks, such as Image classification, Semantic Segmentation and Object detection. With great advancements in terms of quality of the obtained results, the path was paved for using these kinds of neural networks in the medical domain. But, when working with sensitive matters involving human lives, there is a need to consider the interpretability and explainability of these models and not just the typical evaluation metrics for the given task. To do such a thing, tools such as LIME and PyTorch Grad-CAM can be used, among many others. The integration of Explainable AI (XAI) methods proposed in this paper aims to enable the paradigm of XAI to be used in medical image classification tasks with the standardized MedMNIST dataset. By doing such an integration, a deeper analysis regarding the quality of the model can be enabled. In that way, instances that were misclassified can be visually examined and used to paint a clearer picture of the complete model's decision-making process.**

Keywords—Explainable Artificial Intelligence (XAI), interpretability, explainability, Image classification, Deep Convolutional neural networks, LIME, Grad-CAM

I. Introduction

The rapid development of Artificial Intelligence algorithms in recent years has enabled the use of vastly more powerful and complex Deep Convolutional neural networks such as ResNet [1], VGGNet [2] and frameworks like MaskRCNN [3]. Typical Computer vision tasks such as Image classification, Semantic segmentation, Object detection and Deep feature factorization have been noticeably enhanced by using these neural network architectures.

Usually, the increase of the model's complexity negatively affects the interpretability and explainability of the obtained results, so these types of models are known to have a black-box approach. This can be especially problematic when working in the field of Medical Artificial Intelligence, where there is a need for the highest degree of both accuracy and transparency.

Defining transparency is not a straightforward task, but in this concrete manner, it will be discussed in terms of interpretability and explainability. The two terms are often used as synonyms in the field of Explainable AI, but in theory, they represent different metrics of XAI.

The term interpretability is defined as the "*ability to explain or to present in understandable terms to a human*" [4], whereas explainability is considered as "*the ability to explain or present the behavior of models in human-understandable terms*" [4-5].

In a way, interpretability is focused towards understanding the inner workings of a model in question, while explainability gives insight into the decisions made.

The authors of [4] also take the stance that interpretability is useful in the sense that it can be used to assess if other desiderata of a system are met.

It is also important to consider that not all systems require interpretability, and that the need for interpretability arises from incompleteness in the formalization of a problem [4].

Going forward, the term interpretability will be solely used thought this paper, as the definition is more appropriate for the specific use-case implemented.

For addressing the transparency issue of black-box models, various solutions have been proposed, but it is important to acknowledge that the XAI solutions greatly differ based on the task in question. Tackling the problem of interpreting the results of Image classification would be drastically different than trying to interpret Object detection techniques.

The specific image classification problem being addressed in this paper is regarded as a multi-class classification problem, as the output can be categorized into one of eight different classes.

The proposed integration of Explainable AI methods for the classification of medical image data is focused on enabling a deeper and more profound understanding of how Deep convolutional neural networks make predictions and to enable more trust-worthy and informed decision-making.

By integrating Explainable AI methods in such a way, the process behind assessing the model's quality is more streamlined. In that way, a visual analysis regarding misclassified instances can provide a deeper understanding of the model's functionality and thus enable the necessary adjustments.

The image classification problem is analyzed by using the MedMNIST standardized dataset [6]. The concrete dataset from the group is referred to as BloodMNIST, intended for multi-class classification tasks.

ISBN 979-8-3503-8699-8/24/$31.00 ©2024 IEEE

It is also important to acknowledge that the concrete implementation proposed in the paper is intended to work with the MedMNIST dataset but can be modified to function with other datasets if needed.

Explainable AI tools chosen for the integration with the image classification task are LIME (Local Interpretable Model-agnostic Explanations) [7] and PyTorch Grad-CAM [8].

II. Related work

When discussing the interpretability of medical image analysis, it is important to acknowledge the multidisciplinary aspect of this field. Implementing the Explainable AI pipelines with existing medical image datasets is just the part of the problem. The other part concerns the skilled medical staff needed to make sense of the obtained interpretability results. But, for the purpose of this paper, only the first aspect will be discussed.

As stated by van der Velden et al [9], the stakes of the decision-making process in medical AI are often high, so, naturally, there is a lot of concern about the black box approach of Deep learning models.

On that note, the European Union has proposed some key requirements related to XAI techniques for Medical Imaging, summarized by [10] in the following categories: confidence and privacy, ethics, and responsibilities as well as bias and fairness.

The importance of XAI in the medical domain is also portrayed by the evident rise of published papers in the Explainable AI techniques for Medical AI field. The rise in recent years is exponential, as can be seen in [10].

Having all that in mind, there was a need for a clear taxonomy regarding the Explainable AI for medical imaging.

A taxonomy proposed in [9] distinguishes these techniques based on the following criteria: model-based versus post hoc, model-specific versus model-agnostic and global versus local.

The difference regarding the post hoc and model-based explanations is that the "*former trains a neural network and subsequently attempts to explain the behavior of the ensuing black box network, whereas the latter forces the neural network to be explainable*" [9].

On the other hand, model-specific explanation methods are limited to classes of models, while model-agnostic explanations are independent of the choice of a type of neural network, as also stated in [9].

Finally, the scope of explanation can be analyzed through the global and local paradigms. Global explanations are the ones regarding explanations for an entire model, whereas the local ones are based on a single output [9].

In that regard, Gradient-weighted class activation mapping (Grad-CAM) [11] is categorized as a post hoc, model-specific and local method. On the other hand, LIME belongs to post hoc, model-agnostic, and local techniques.

The reason for choosing these specific Explainable AI tools, among many others, is partially because they both employ the previously mentioned post hoc and local explanation paradigms. In that way, an analysis regarding the single data instance with a complex deep neural network model in the background can be done. Also, both tools are widely used, making the process of integration more seamless.

As stated in [9], the scientific research in the field of Explainable AI for medical data can be analyzed by the techniques used, but also by the anatomical location and modality.

The related work regarding the use of Grad-CAM with histology as the chosen modality [9] (the BloodMNIST dataset used in this paper is regarding the individual cells) can be seen in the work by Tang et al (classification of Alzheimer's disease pathologies) [12], Obikane et al (histopathological image segmentation) [13], He et al (lung adenocarcinoma classification) [14] and Teramoto et al (classification of benign and malignant cells from lung cytological images) [15].

Also, the work by Gupta et al (region of interest identification for cervical cancer) [16], GV et al (automatic classification of whole side pap smear images) [17], Korbar et al (interpreting whole-slide image analysis outcomes for colorectal polyps) [18], Kowsari et al (hierarchical medical image classification) [19], Ji et al (classification of pathological images) [20] and Chan et al (semantic segmentation of histological tissue type) [21] gives various insights into the problematic of interpreting medical image analysis.

Using LIME was far less present in the taxonomy given in [9], with only two papers, regarding just the X-ray [22] and endoscopy modalities [23].

III. The theory behind XAI tools – LIME and Grad-CAM

The core functionalities behind the chosen XAI tools need to be examined through the theoretical principles behind those implementations.

A. LIME (Local Interpretable Model-Agnostic Explanations)

Following the definition given by Ribeiro et al, LIME is a "*novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction*" [7].

The authors also stated that "*explaining a prediction means presenting textual or visual artefacts that provide qualitative understanding of the relationship between instance's components and the model's prediction"* [7]. In the notion of working with image data, instance components are regarded to be image patches.

The explanations are calculated by following the provided mathematical formula found in [7]:

$$\xi(x) = \operatorname{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g). \quad (1)$$

Defining the notation present in (1), $g \in G$ represents an explanation and G is a class of potentially interpretable models. The term $\Omega(g)$ is defined as a measure of complexity, which is on the opposite side of the spectrum from interpretability.

The probability that x belongs to a class is defined by the function f(x) and $\pi_x(z)$ represents a proximity measure between an instance z to x. $\mathcal{L}(f, g, \pi_x)$ is a measure of how unfaithful g is in approximating f in the locality defined by π_x [7].

For the process of interpretability to be possible and valid, the first argument in (1) needs to be minimized, while the second argument must be low enough to still be interpretable by humans. It is important to acknowledge that interpretability also depends on the target audience, as stated in [7].

B. Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM is a technique for producing visual explanations for decisions from Convolutional Neural Network architectures. As stated by [11], Grad-CAM functions in a way that it "*uses the gradients of any target concept, flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept*".

Grad-CAM is implemented in a way that it is usable in a variety of CNN model families, including networks with fully connected layers, networks used for structured outputs, used in tasks with multi-modal inputs or even for reinforcement learning [11]. All the mentioned can be done without changing the model architecture, making Grad-CAM an incredibly versatile solution.

For image classification, the visualizations from Grad-CAM are used to help in the identification of dataset bias and give insights into failures of current CNNs.

The mathematical background of the core functionalities of Grad-CAM is described using the following equations [11]:

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2}$$

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k). \tag{3}$$

As stated in [11], for the class-discriminative localization map GRAD-CAM $L_{Grad-CAM}^c$ to be obtained, the gradient of the score for class c, y^c needs to be calculated. This needs to be done with the respect to feature maps A^k of a convolutional layer.

Also stated in [11], the weight α^c represents "*a partial linearization of the deep network downstream from A and captures the importance of feature map k for a target class c*".

The authors of [11] also argue that the ReLU (Rectified linear unit) is applied to the linear combination of maps because they are interested in only the features that have a positive influence on the class of interest. Without ReLU, localization maps can highlight more than just the desired class.

IV. Methodology

The integration of Explainable AI methods for the classification of Medical Image Data proposed in this paper is based on the explanations provided by both LIME and PyTorch Grad-CAM tools, as previously stated.

When integrating these Explainable AI methods with the provided data, a few key steps need to be considered.

Firstly, the Deep Convolutional Neural Network needs to be fine-tuned for working with the provided dataset.

In this case, the process of fine-tuning is done regarding the downstream task of multi-class classification. After that, the integration process can be done for both XAI tools, and the obtained results can be analyzed in detail.

A. Dataset

MedMNIST is a MNIST-like dataset collection of standardized biomedical images [6]. There are a total of 12 datasets for 2-dimensional pictures and 6 datasets for 3-dimensional pictures. The concrete dataset used is BloodMNIST, with Blood Cell Microscope as the data modality.

It is important to keep in mind that there are different versions of the MedMNIST dataset.

The version used in this paper is MedMNIST+ meaning that the size of images is larger compared to the standard dataset (224x224 pixels for the BloodMNIST dataset).

The dataset is regarding the task of multi-class classification of blood cells in eight different classes (basophil, eosinophil, erythroblast, immature granulocytes, lymphocyte, monocyte, neutrophil and platelet).

The total number of samples is 17092 of which 11959 are used for the purpose of training, 1712 for validation and 3421 for testing.

As stated in [6], the data is based on "*individual normal cells, captured from individuals without infection, hematologic or oncologic disease and free of any pharmacologic treatment at the moment of blood collection*".

B. Deep Convolution Neural Networks

The networks chosen for the purpose of medical image classification are *Resnet18* and *Resnet50* [1]. The process of fine-tuning was done thought five epochs, with the learning rate set to 0.001 (for the larger Resnet50, it was set to a value of 0.0001). The optimizer of choice was AdamW [24].

Having in mind that the task in hand is image classification, the fully connected layer of the neural networks needed to be changed to be compatible with the given dataset. Besides that, modifications regarding the addition of a Dropout layer were also implemented to reduce overfitting.

The complete implementation of the process of fine-tuning both Resnet18 and Resnet50 with the BloodMNIST dataset can be seen at [25], while the documentation used for this implementation is available at [26-27].

The obtained results regarding the classification metrics of both networks are displayed in Table 1. Having in mind that this is a multi-class classification problem, the numbers displayed in Table 1 are regarding the usage of average type "macro". Per class classification metrics are available while running the *experiment.py* script present in [25].

TABLE I. Classification metrics for Fine-tuned Resnet18 and Resnet50 models

Model	Classification metrics		
	Accuracy	***Precision***	***Recall***
Resnet18	0.98042	0.97147	0.98042
Resnet50	0.98287	0.98000	0.98287

C. The integration of XAI tools

Having previously discussed the notion of implementing the process of fine-tuning the chosen Deep Convolutional Neural Networks on the BloodMNIST dataset, the final step of integrating XAI methods can be addressed.

While working with LIME, a chosen test image needs to be preprocessed and adequately formatted to be used by the Lime Image Explainer. The specificity also lies in choosing the right parameters for an explanation to be as valid as possible.

Modifying the parameters regarding the built-in *explain instance* LIME function can provide far better results in terms of interpretability. Concretely, the *number of samples* parameter can be adjusted to consider a larger neighboring surface around the instance in question. By doing so, the gathered top labels for an explanation can provide much more insight, but the computation times are expected to be longer.

Implementing the Grad-CAM algorithm present in the PyTorch Grad-CAM library also requires specific image preparation, as well as choosing the appropriate target layer from the neural network architecture, that contains the most relevant information (the adequate layer varies based on the underlaying neural network architecture). The final convolutional layer is a good starting point when considering the adequacy of spatial information captured.

Finally, the power of doing such an integration lies in having multiple side-by-side results, or interpretations, with different underlying algorithms. In that way, the level of trust can be even more elevated and potential shortcomings can be addressed.

V. Results

The complete implementation of the process of integrating Explainable AI methods for the classification of medical image data can be seen at [28], while the documentation used is available at [29-31].

Traditional classification metrics such as accuracy, recall, precision, and F1-measure cannot be solely trusted without the detailed examination of the model's black box nature.

By analyzing the highlighted parts of the image, it can be seen why the model predicted such a value. So, in a way, the process of interpretability can be seen as a gateway for enabling the debugging capabilities regarding the model's output.

Firstly, a function for getting PyTorch probabilities was implemented, to have a sense of the expected class of an image. For discussing the results, an image from the test dataset was chosen, and the prediction was that the cell in question belongs to the eosinophil class (stated by using both the PyTorch predictions gathered from Resnet18 and Resnet50). The same test instance was used for both deep neural networks.

In Fig. 1. a side-by-side comparison of the results obtained by using Resnet18 with both LIME and PyTorch Grad-CAM can be seen. Going from left to right, the first image is the original, the second is regarding the explanations provided by LIME, while the third is for the PyTorch Grad-CAM implementation.

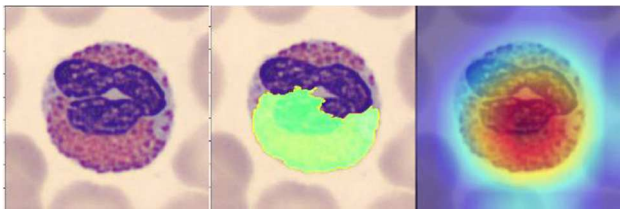

Fig. 1. Resnet18 with LIME and PyTorch Grad-CAM (eosinophil cell)

The results given by LIME can vary drastically depending on the number of features chosen for the explanation. In this example, the number of features was set to four. The green color is used for highlighting the part of the image that contributed the most towards a cell being classified in that way.

Grad-CAM uses a specific coloring scheme, highlighting in red the most crucial part of the image, meaning that the saturated red color indicates the part that contributed the most to an instance being of a specific class.

Similarly, in Fig. 2. a side-by-side comparison of the results obtained by using Resnet50 with both tools is displayed.

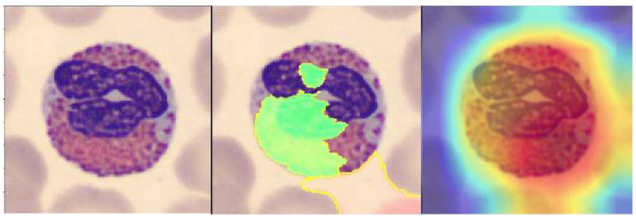

Fig. 2. Resnet50 with LIME and PyTorch Grad-CAM (eosinophil cell)

Interestingly, in the case of using Resnet50, the previously mentioned function for getting PyTorch probabilities and the output provided by LIME did not match. The top class predicted by getting the PyTorch probabilities was 1, and the one got from using LIME's explainer object was 6.

The mismatch could have occurred because of various factors, including overfitting as a possible cause. It is important to keep in mind that larger neural network architectures tend to overfit more easily, hence the classification metrics need to be monitored more carefully. In the case of such a mismatch, a detailed analysis needs to be implemented, for the cause to be far more certain.

Nevertheless, the visualization provided in Fig. 2. was implemented in a way to showcase the instance's belonging to class 1 (by adequately choosing the top class to be predicted).

In the case of using LIME, as previously stated, the positive notion is highlighted green, while the negative is red. For PyTorch Grad-CAM, the saturated red color highlights the pixels mostly contributing to an instance being classified in a specific class category.

By doing such a comparison including different model architectures, as well as tools, a more detailed instance analysis can be implemented.

For that to be possible, assistance from a medical practitioner would be highly valuable. In that way, the crucial parts of the image highlighted by the practitioner could be used to determine which model gave the more precise localization.

Exploring the interpretability regarding different instances from the BloodMNIST dataset can be done using the previously mentioned implementation, available at [28].

VI. Conclusion

The integration of Explainable AI tools for the classification of Medical Image data proposed in this paper intended to demonstrate the potential of demystifying predictions given by complex Deep Convolutional Neural networks such as Resnet in the field of Medical AI.

By incorporating a specific pipeline, instances from the MedMNIST dataset were analyzed with various Explainable AI tools, such as LIME and PyTorch Grad-CAM. The power of such analysis lies in the side-by side comparison of the obtained results. In that way, a more profound examination of the specific instances can be done, and potential biases and setbacks can be timely addressed.

The future work can be examined in various directions, including the use of other Deep neural network architectures. On the other hand, the rise of available Explainable AI tools is evident, so the potential lies in having a more profound integration pipeline.

To enable an even more intuitive use-case scenario, a visual interface in the form of a web-based application can be developed. By doing such a thing, a true no-code approach would be possible.

ACKNOWLEDGMENT

This work has been supported by the Serbian Ministry of Science, Technological Development and Innovation [grant number 451-03-65/2024-03/200102].

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[2] K. Simonyan, and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick. "Mask r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969. 2017.

[4] F. Doshi-Velez, and B. Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).

[5] M. Du, N. Liu, and X. Hu. "Techniques for interpretable machine learning." Communications of the ACM 63, no. 1 (2019): 68-77.

[6] J. Yang, R. Shi, and B. Ni. "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis." In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 191-195. IEEE, 2021.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144. 2016.

[8] J. Gildenblat and contributors. "PyTorch library for CAM methods." (2021).

[9] B. HM Van der Velden, H. J. Kuijf, K. GA Gilhuijs, and M. A. Viergever. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis." Medical Image Analysis 79 (2022): 102470.

[10] A. Chaddad, J. Peng, J. Xu, and A. Bouridane. "Survey of explainable AI techniques in healthcare." Sensors 23, no. 2 (2023): 634.

[11] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In Proceedings of the IEEE international conference on computer vision, pp. 618-626. 2017.

[12] Z. Tang, K. V. Chuang, C. DeCarli, LW Jin, L. Beckett, M. J. Keiser, and B. N. Dugger. "Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline." Nature communications 10, no. 1 (2019): 2173.

[13] S. Obikane, and Y. Aoki. "Weakly supervised domain adaptation with point supervision in histopathological image segmentation." In Pattern Recognition: ACPR 2019 Workshops, Auckland, New Zealand, November 26, 2019, Proceedings 5, pp. 127-140. Springer Singapore, 2020.

[14] J. He, L. Shang, H. Ji, and XL. Zhang. "Deep learning features for lung adenocarcinoma classification with tissue pathology images." In Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part IV 24, pp. 742-751. Springer International Publishing, 2017.

[15] A. Teramoto, A. Yamada, Y. Kiriyama, T. Tsukamoto, K. Yan, L. Zhang, K. Imaizumi, K. Saito and H. Fujita, 2019. Automated classification of benign and malignant cells from lung cytological images using deep convolutional neural network. Informatics in Medicine Unlocked, 16, p.100205.

[16] M. Gupta, C. Das, A. Roy, P. Gupta, G. Radhakrishna Pillai, and K. Patole. "Region of interest identification for cervical cancer images." In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1293-1296. IEEE, 2020.

[17] K. Kiran GV, and G. Meghana Reddy. "Automatic classification of whole slide pap smear images using CNN with PCA based feature interpretation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0-0. 2019.

[18] B. Korbar, A. M. Olofson, A. P. Miraflor, C. M. Nicka, M. A. Suriawinata, L. Torresani, A. A. Suriawinata, and S. Hassanpour. "Looking under the hood: Deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps." In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 69-75. 2017.

[19] K. Kowsari, R. Sali, L. Ehsan, W. Adorno, A. Ali, S. Moore, B. Amadi, P. Kelly, S. Syed, and D. Brown. "Hmic: Hierarchical medical image classification, a deep learning approach." Information 11, no. 6 (2020): 318.

[20] J. Ji "Gradient-based interpretation on convolutional neural network for classification of pathological images." In 2019 International Conference on Information Technology and Computer Application (ITCA), pp. 83-86. IEEE, 2019.

[21] L. Chan, M. S. Hosseini, C. Rowsell, K. N. Plataniotis, and S. Damaskinos. "Histosegnet: Semantic segmentation of histological tissue type in whole slide images." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10662-10671. 2019.

[22] S. Rajaraman, S. Candemir, G. Thoma, and S. Antani. "Visualizing and explaining deep learning predictions for pneumonia detection in pediatric chest radiographs." In Medical Imaging 2019: Computer-Aided Diagnosis, vol. 10950, pp. 200-211. SPIE, 2019.

[23] A. Malhi, T. Kampik, H. Pannu, M. Madhikermi, and K. Främling. "Explaining machine learning-based classifications of in-vivo gastral images." 2019 Digital Image Computing: Techniques and Applications (DICTA) (2019): 1-7.

[24] I. Loshchilov, and F. Hutter. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101 (2017).

[25] The implementation of model fine-tuning, GitHub repository, https://github.com/minanikolic916/IntegrationXAIMed/tree/master, last accessed 2024/3/24

[26] Official documentation regarding the usage of MedMNIST with PyTorch https://github.com/MedMNIST/MedMNIST/blob/main/examples/getting_started.ipynb, last accessed 2024/3/24

[27] PyTorch transfer learning, https://www.learnpytorch.io/06_pytorch_transfer_learning/, last accessed 2024/3/24

[28] The implementation of the integration of Explainable AI Methods for the classification of Medical Image data, GitHub repository, https://github.com/minanikolic916/IntegrationXAIMed/blob/master/AIinMedicine%20(1).ipynb, last accessed 2024/3/24

[29] Official documentation regarding the usage of LIME with PyTorch, https://github.com/marcotcr/lime/blob/master/doc/notebooks/Tutorial%20-%20images%20-%20Pytorch.ipynb, last accessed 2024/3/24

[30] Introduction: Advanced Explainable AI for computer vision, https://jacobgil.github.io/pytorch-gradcam-book/introduction.html, last accessed 2024/3/24

[31] Tutorial: Concept activation maps, https://jacobgil.github.io/pytorch-gradcam-book/Pixel%20Attribution%20for%20embeddings.html, last accesed 2024/3/25