Similarity-Based Explanations for Deep Interpretation of Capsule Endoscopy Images

Miguel Fontes[1,2(✉)], Danilo Leite[2], João Dallyson[3], and António Cunha[1,2]

[1] UTAD—University of Trás-os-Montes and Alto Douro, 5001-801 Vila Real, Portugal
[2] INESC TEC—Institute for Systems and Computer Engineering, Technology and Science, 4200-465 Porto, Portugal
miguel.f.fontes@inesctec.pt, {danilol,acunha}@utad.pt
[3] UFMA—Federal University of Maranhão, São Luís, Brazil
jdallyson@nca.ufma.br

Abstract. Artificial intelligence (AI) is playing a growing role today in several areas, especially in health, where understanding AI models and their predictions is extremely important for health professionals. In this context, Explainable AI (XAI) plays a crucial role in seeking to provide understandable explanations for these models.

This article analyzes two different XAI approaches applied to analyzing gastric endoscopy images. The first, more conventional approach uses Grad CAM, while the second, even less explored but with great potential, is based on "similarity-based explanations". This example-based XAI technique aims to provide representative examples to support the decisions of AI models.

In this study, we compare these two techniques applied to two different models: one based on the VGG16 architecture and the other based on ResNet50, designed to classify images from the KVASIR-capsule database. The results reveal that Grad-CAM provided intuitive explanations only for the VGG16 model, while the "similarity-based explanations" technique provided consistent explanations for both models. We conclude that exploring other XAI techniques can be a significant asset in improving the understanding of the various AI models.

Keywords: XAI · Example-based · Similarity-based explanations · endoscopy

1 Introduction

Deep learning has gained increasing importance in medical image processing, significantly influencing areas such as detection, recognition, segmentation and computer-aided diagnosis [1]. Deep learning models have demonstrated outstanding performance in tasks such as classification, lesion detection and segmentation [2]. However, the complexity of medical decisions requires a deep understanding and proper interpretation of the models in use.

Explainable Artificial Intelligence (XAI) refers to the development of Artificial Intelligence (AI) models and techniques that are able to offer transparent and interpretable

A. Cunha et al. (Eds.): MobiHealth 2023, LNICST 578, pp. 208–222, 2024.
https://doi.org/10.1007/978-3-031-60665-6_16

explanations for their decisions and predictions [3]. In medical imaging, XAI plays an essential role in providing information about the reasoning behind the model's predictions and thus earning the trust of doctors and patients [4]. Several XAI techniques are employed in medical imaging, including model interpretation, which involves understanding learned features, and interpretation of model results, which includes attribution-based methods such as saliency maps and class activation maps [4]. An exciting approach to XAI is example-based explanation, which consists of providing specific examples or instances to illustrate how the AI model arrived at a decision [5]. These example-based explanations can improve the interpretability of AI models, especially in medical diagnostic imaging, where the "black box" nature of deep learning models has hindered their adoption in clinical workflows [5].

1.1 Related Work

In medical image analysis using deep learning techniques, several recent studies have contributed to significant advances, incorporating XAI techniques to improve the interpretability of models.

One of these studies focused on developing an architecture called HAnet, applying deep convolutional neural networks (CNNs) to recognize ulcers in images of endoscopic capsules (WCE). In addition, the Class Activation Mapping (CAM) technique was used to visualize and interpret the network's activation regions, making CNN's decision-making process more transparent [6]. Another study proposed an explainable machine learning tool to support the interpretation of gastric images in vivo. Using convolutional neural networks and visual explanations, this research sought to provide understandable insights for healthcare professionals, highlighting the potential of these XAI approaches in medical contexts [7]. In addition, another study focused on the application of Convolutional Neural Networks (CNNs) in the segmentation of polyps in colonoscopy images, enhancing Fully Convolutional Networks (FCNs) architectures to achieve meaningful results, also using XAI techniques to interpret the results [8].

These investigations illustrate the growing importance of deep learning and XAI techniques, such as CAM, in improving the analysis of medical images and supporting clinical decision-making, making models more transparent and reliable for healthcare professionals.

Our research group has maintained a significant focus on the analysis of capsule endoscopy images. One of the group's papers [9] demonstrated success in using a Visual Transformers model to analyze capsule endoscopy images, achieving exceptionally high accuracy and sensitivity rates. Another relevant work [10] explored the challenge of classifying abnormalities in small, unbalanced data sets, highlighting the potential of using transfer learning even with a limited number of samples. In addition, the group also developed an unsupervised method for estimating homography in capsule endoscopy frames [11], which has the potential to considerably improve the precise localization of the endoscopic capsule. The cumulative contributions of these studies reflect the group's commitment to promoting innovative solutions to the persistent challenges associated with capsule endoscopy image analysis, excelling in the advancement of effective diagnosis of conditions related to the gastrointestinal tract.

1.2 Objective

Given that most XAI techniques applied to medical images are based on visual explanations, such as Grad-CAM and LIME, this article aims to introduce an example-based XAI approach called "similarity-based explanations". In addition, it seeks to compare the explanations generated by this approach with the explanations of a more conventional technique, Grad-CAM. This research aims to provide valuable insights into the effectiveness and applicability of these techniques in interpreting results in medical images. Importantly, our focus extends beyond merely optimizing image classification accuracy, encompassing a broader understanding of the interpretative value of XAI in medical contexts.

2 Methodology

The methodology, in Fig. 1, adopted in this study follows a well-defined pipeline. Initially, the images contained in the database go through a pre-processing process. The data set is then divided into three parts: a training set, a validation set and a test set. The training and validation sets are used to train the models created to address the problem in question, while the test set is reserved exclusively for evaluating the model's performance.

Two XAI techniques are applied once the model's predictions have been obtained on the test set. These techniques are used to analyze and interpret the model predictions and provide a deeper understanding of the functioning and performance of the XAI techniques used. Finally, the results generated by both XAI techniques are analyzed, contributing to a deeper understanding of the problem under study and the methods employed.

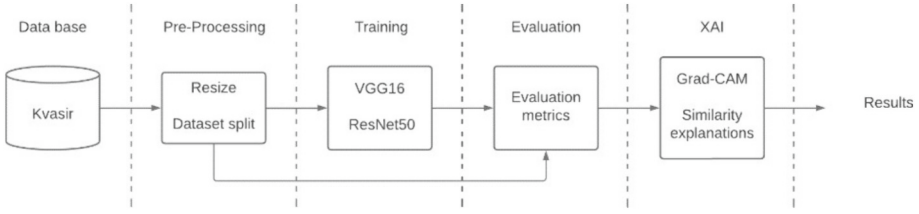

Fig. 1. Methodology Pipeline.

2.1 Database

The Kvasir-Capsule database is the only source of data for this study. It consists of 117 capsule endoscopy videos, which can be divided into 4,741,504 image frames. This database includes various information covering anatomical landmarks such as the Z-line, pylorus, and caecum and pathological findings, such as esophagitis, polyps and ulcerative colitis [12].

For this research, we used only part of the dataset, selecting 4,000 images from 8 specific medical classes within this database, which is represented in Fig. 2. This careful selection allowed us to focus on the classes relevant to the research.

This database is remarkable not only for the amount of data available but also for the quality of the medical annotations. A total of 47,238 image frames were meticulously annotated and reviewed by specialist doctors, providing a valuable labelled dataset.

In addition, the database includes a vast collection of 4,694,266 unlabeled frames, which represent a significant opportunity to explore advanced deep learning approaches.

This combination of labelled and unlabeled data provides a solid basis for investigating the interpretability of deep learning models in medical applications while seeing to achieve high classification performance.

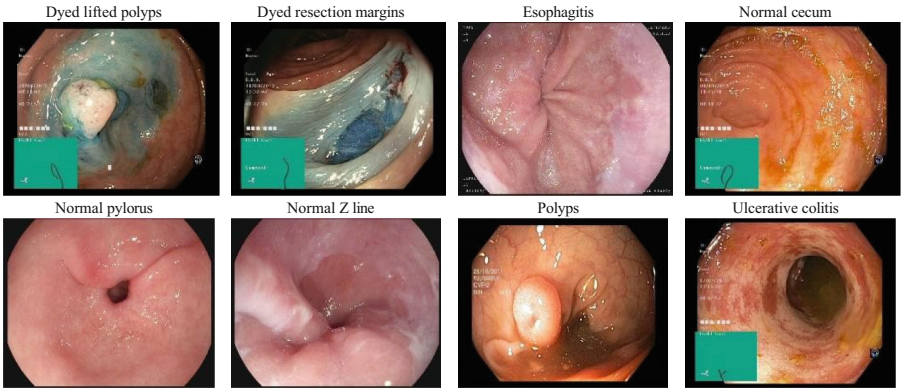

Fig. 2. Examples of images from the different classes.

2.2 Pre-Processing

To use the images from the Kvasir-capsule database in this study, it was necessary to carry out appropriate pre-processing in line with the models employed. Since this study demonstrates two different XAI approaches, we opted to simplify the pre-processing rather than achieve high classification accuracy.

Pre-processing mainly consisted of resizing the images to make them compatible with the input format of the classification model. The images were resized to a dimension of 224 pixels high, 224 pixels wide and 3 color channels (RGB). This transformation allowed the images to be processed uniformly by the classification model, facilitating integration with XAI techniques.

It is important to note that this pre-processing was carried out in a minimalist manner since the main focus was evaluating the XAI approaches, and not optimizing the classification accuracy of the images.

At this stage, the database was also divided into training, validation and test datasets. 2240 images were allocated to the training set, 960 images to the validation set and 800 images to the test set. This division was carried out to ensure the robustness of the experiments and to allow proper evaluation of the XAI techniques on different datasets.

2.3 Training Models

To classify the images after the pre-processing stage, two models were developed.

Model selection

With regard to the models developed, the first adopts the VGG16 architecture, while the second is based on ResNet50.

The VGG16 architecture is a deep convolutional network (CNN) that is widely recognized and used in various research domains. This architecture consists of a total of 16 layers, of which 13 are convolutional layers, and 3 are fully connected layers. In the convolutional layers, 3×3 filters are used, making it possible to extract detailed and deep features from the images. In addition, the architecture incorporates pooling layers to reduce the dimensionality of the data, followed by fully connected layers to perform the final classification [13].

It is important to note that the VGG16 architecture provides approximately 138 million trainable parameters. This wealth of parameters allows the network to learn complex representations of images, making it capable of performing high-level classification tasks. However, this advantage comes with the need for an extensive set of training data and substantial computing resources [13].

In the context of this specific study, we chose to use only the convolutional layers of the VGG16 architecture combined with a custom classifier. The classifier consists of a "Global Average Pooling 2D" layer, followed by 5 dense layers and 2 dropout layers.

This configuration was chosen based on the needs of the task in question and the objective of extracting significant features from the input images.

The ResNet50 architecture, a variation of the ResNet (Residual Network) architecture, is a Convolutional Neural Network (CNN) notable for its depth and the implementation of residual connections [14]. ResNet50 comprises a total of 50 layers, including convolutional layers, pooling layers and dense layers. The distinctive feature of this network is the use of so-called "residual layers," in which the input of a block is added to the output of the previous block. This allows information to be transmitted directly between layers, minimizing information loss.

The residual connections approach is a fundamental innovation that solves the challenge of performance degradation in deep networks by allowing the network to learn deeper and more complex representations of image characteristics [14].

The ResNet50 architecture has been widely adopted in a number of computer vision applications in various fields, and its characteristics make it a popular choice for tasks that require deep neural networks with high performance.

In this study, we chose to use the ResNet50 architecture up to the last convolutional layer, connecting it to the same classifier employed in the VGG16-based model.

Optimizer and loss function

In both models developed in this study, the parameters related to the optimizer and loss function were kept consistent to ensure a fair and rigorous comparison. The configurations of these elements are detailed below:

Optimizer: We chose to use the "Stochastic Gradient Descent (SGD)" optimizer in both models. This choice was motivated by its effectiveness in supervised learning tasks and its ability to adjust the weights of the neural network iteratively, which is especially relevant for training complex models.

Loss function: Given that both tasks involved classifying data into several classes, we selected the "Categorical Crossentropy" loss function. This loss function is appropriate for multi-class classification tasks, helping the model to calculate the discrepancy between the predicted probabilities and the actual data labels.

Standardizing these elements in the configuration of both models ensured a solid basis for comparing results and a consistent approach throughout the study. These methodological choices were aimed at obtaining robust and reliable conclusions regarding the performance of the models in multi-class classification tasks.

Callbacks

In this study, three callback techniques were used during model training. These techniques were selected to optimize model performance, avoid overfitting and efficiently manage the available computing resources. The three Callback techniques used are described below:

ModelCheckpoint: We used the ModelCheckpoint Callback to save the model weights in temporary files during training. This procedure was triggered whenever the model reached its best performance in relation to the defined evaluation metrics. This allowed us to retain the weights corresponding to the model's best performance, guaranteeing the ability to restore it later if necessary.

Reduce Learning Rate on Plateau: We implemented the Reduce Learning Rate on Plateau Callback to monitor the model's performance over training epochs. This Callback automatically adjusts the learning rate if the model's performance does not show significant improvements after a specified number of consecutive epochs. We started training with a learning rate of 0.001, and the Callback configuration was set to reduce it by a factor of 0.01 after three epochs of no improvement in 'val_loss', with a minimum learning rate set at $1e{-}5$. This adaptive approach allows the model to adjust its learning rate according to fluctuations in performance, increasing the likelihood of effective convergence.

EarlyStopping: We integrated Callback EarlyStopping to stop training the model if it was no longer able to improve its performance. This was based on a defined stop metric. The inclusion of EarlyStopping helped to avoid excessively long training runs and unnecessary computational resources by ensuring that the model was only trained as long as it continued to improve at its specific task.

Epochs and batch size

When training the neural network, we adopted a batch size of 64 images per batch, and the training process was conducted over 100 epochs. These configurations were selected with the aim of balancing computational efficiency and the model's ability to converge on an optimal solution. It is important to note that, although the training was programmed for 100 epochs, we frequently observed the activation of the Early Stopping mechanism, which interrupted the training around epochs 30 to 35.

2.4 Evaluation

After training the model, we tested the model on the test dataset in order to obtain the model's predictions and used the following evaluation metrics:

Precision: Precision quantifies the proportion of correct predictions of a positive class in relation to the total number of positive predictions made by the model.

Recall: Recall measures the model's ability to effectively identify all samples belonging to a positive class, expressing the proportion of true positives (TP) in relation to the sum of true positives and false negatives (FN). Rev is relevant when you want to avoid losing positive samples.

F1-Score: The F1-Score is a metric that harmonizes Precision and Recall in a single measure. The F1-Score provides a balanced assessment of the model when there is a need to balance the ability to identify true positives and avoid false negatives and false positives.

Accuracy: Accuracy represents the proportion of correct predictions in relation to the total number of samples. It is calculated as the sum of TP and true negatives (TN) divided by the sum of TP, TN, false positives (FP) and FN. Accuracy provides an overview of the model's ability to classify all samples correctly but can be misleading in scenarios with class imbalance.

Macro and Weighted Average: The macro and weighted average provide aggregate summaries of the metrics (Precision, Recall and F1-Score) calculated for all classes. The macro average calculates a simple average, while the weighted average takes class imbalance into account, assigning weight based on the support of each class. These averages are useful for evaluating the overall performance of the model, taking all classes into account and adjusting for class imbalance where applicable.

These evaluation metrics are crucial for assessing the performance of machine learning models in classification tasks, providing valuable insights into the model's ability to make accurate predictions and correctly identify relevant samples.

2.5 Explainable Artificial Intelligence (XAI)

XAI is an area of research that seeks to make machine learning models understandable by revealing how models make decisions, highlight important features and reduce uncertainties, which is fundamental in critical applications. In this study, we used XAI techniques to understand how our model makes decisions and highlights features in images, improving our analysis by employing two different approaches.

Grad-CAM

Grad-CAM is an XAI technique that aims to reveal which regions of an image are most influential in the decision-making of a deep learning model during classification. Unlike other techniques that focus only on the last convolutional layer of a neural network, Grad-CAM analyzes activation gradients in intermediate layers [15].

It works by calculating gradients of the class of interest in relation to the activations of the intermediate layer of the neural network. These weighted gradients are then aggregated to create an activation map that highlights the most critical areas of the image for classification. Grad-CAM offers a valuable visual interpretation, allowing researchers and practitioners to better understand how the model makes decisions and to identify which image features are relevant to a given class [15].

This technique has been widely adopted in Machine Learning Explainability studies to improve the interpretability of deep learning models, making it a valuable tool for analyzing and understanding the behaviour of neural networks in computer vision and image classification tasks.

Similarity-based explanations

Similarity-based explanations are an approach that seeks to make the workings of machine learning models more understandable. This technique focuses on presenting examples of training data that are similar to the current data input, with the aim of explaining why the model made a certain prediction.

This explanation follows intuitive logic, where the model states something like, "This prediction is valid because similar examples in the training set also resulted in similar predictions". This approach is analogous to the way humans make decisions based on past experiences. The main benefit of similarity-based explanations is their high comprehensibility. They make it easier for users to understand the reasoning behind the model's predictions, which is particularly important in critical applications [16, 17].

However, it is important to note that this technique has its limitations, such as the reliance on high-quality training data and the need for similar instances in the training set. Furthermore, it may not be suitable for extremely complex models that are not based on direct similarities between instances.

In this context, we chose to use cosine distance as the main similarity metric. Cosine distance is a widely recognized metric applied in data analysis and machine learning. It measures the similarity between two feature vectors, taking into account their relative orientation in the feature space, rather than calculating a direct measure of the Euclidean distance between these vectors [16].

In simple terms, the cosine distance ranges from -1 to 1, where 1 indicates that the vectors have an identical direction (i.e. maximum similarity), 0 indicates that they are orthogonal (neutrality), and -1 indicates that they have opposite directions (i.e. maximum dissimilarity). This metric is particularly useful when dealing with vector data and is known for its interpretability, making it a suitable choice for explaining model predictions based on the intuitive logic that predictions are valid when similar examples have resulted in similar predictions in the training set.

Cosine distance offers an effective way of quantifying the similarity or dissimilarity between data instances, which contributes significantly to making our similarity-based explanations more understandable and accessible to users.

3 Results

3.1 Evaluation Metrics

The results of the evaluation metrics for the two models can be seen in Tables 1 and 2.
VGG16:

Table 1. Evaluation metrics of the VGG16 model.

	Precision	Recall	F1-Score	Number of images
Class 0	0.94	0.84	0.89	100
Class 1	0.88	0.96	0.92	100
Class 2	0.88	0.81	0.84	100
Class 3	0.92	0.98	0.95	100
Class 4	0.97	1.00	0.99	100
Class 5	0.83	0.89	0.86	100
Class 6	0.94	0.90	0.92	100
Class 7	0.97	0.95	0.96	100
Total accuracy	0.92			800
Macro avg	0.92	0.92	0.92	800
Weighted avg	0.92	0.92	0.92	800

ResNet50:

Table 2. Evaluation metrics of the ResNet50 model.

	Precision	Recall	F1-Score	Number of images
Class 0	0.88	0.82	0.85	100
Class 1	0.84	0.87	0.85	100
Class 2	0.86	0.80	0.83	100
Class 3	0.98	0.99	0.99	100
Class 4	0.98	0.95	0.96	100
Class 5	0.79	0.86	0.82	100
Class 6	0.92	0.98	0.95	100
Class 7	0.98	0.95	0.96	100
Total accuracy	0.90			800
Macro avg	0.90	0.90	0.90	800
Weighted avg	0.90	0.90	0.90	800

3.2 Explainable Artificial Intelligence (XAI)

The results of applying XAI techniques are as follows.

Grad-CAM

The results of the Grad-CAM technique show that it was applied to four images from the test set that were correctly classified by both models, as shown in Figs. 3 and 5. In addition, the technique was applied to four images from the test set that were incorrectly classified, as shown in Figs. 4 and 6.

VGG16:

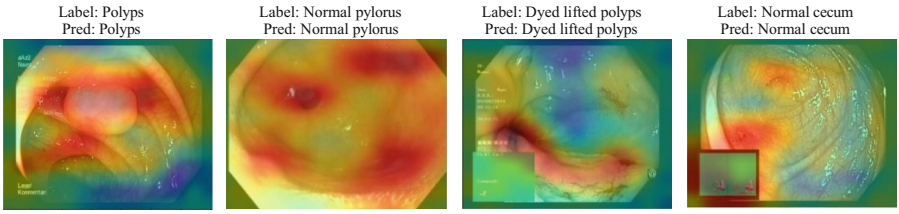

Fig. 3. Grad-CAM results on images correctly classified by the VGG16 model.

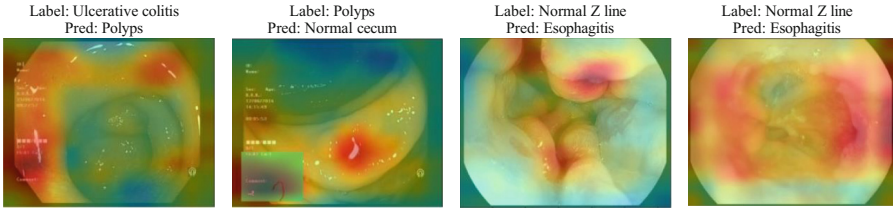

Fig. 4. Grad-CAM results on images incorrectly classified by the VGG16 model.

ResNet50:

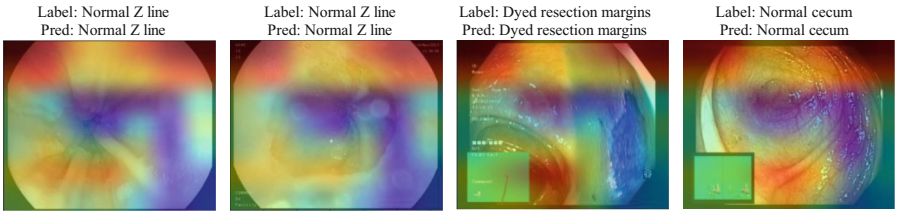

Fig. 5. Grad-CAM results on images correctly classified by the ResNet50 model.

Similarity-based explanations

In the results of the similarity-based explanation technique, it was possible to see four images from the test set for each model. For each of these images, the five most similar images found in the training set were displayed. Figure 7 shows the technique applied to the VGG16 model, while Fig. 8 shows the technique applied to the ResNet50 model.

VGG16:

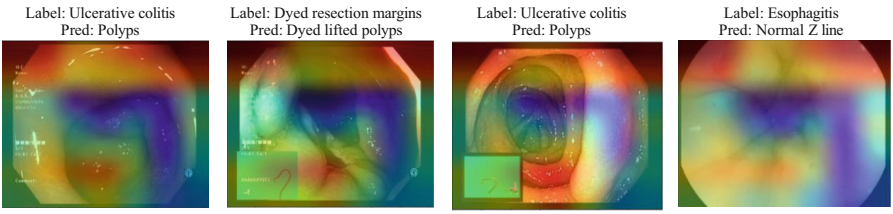

Fig. 6. Grad-CAM results on images incorrectly classified by the ResNet50 model.

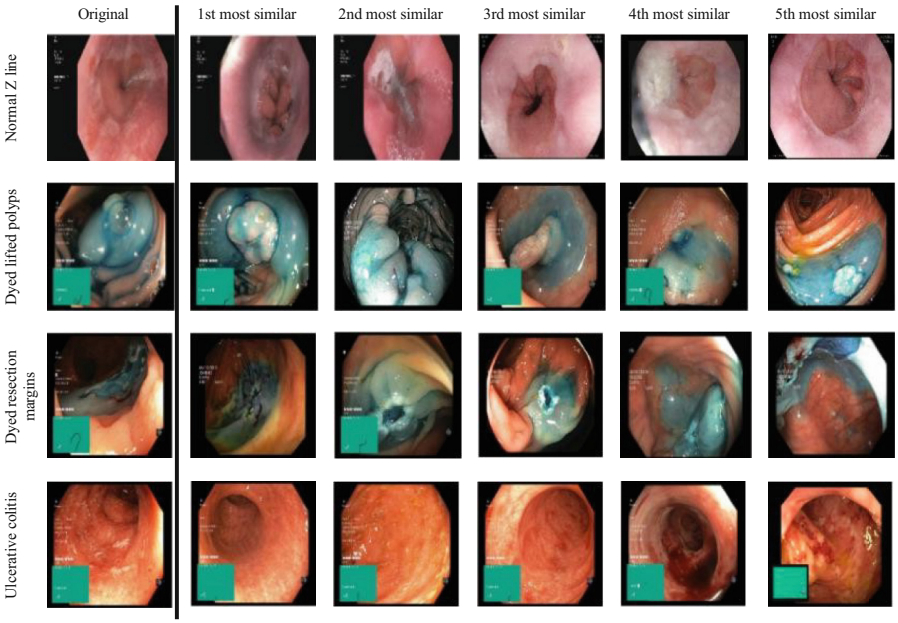

Fig. 7. Results of the similarity-based explanations technique of the VGG16 model.

ResNet50:

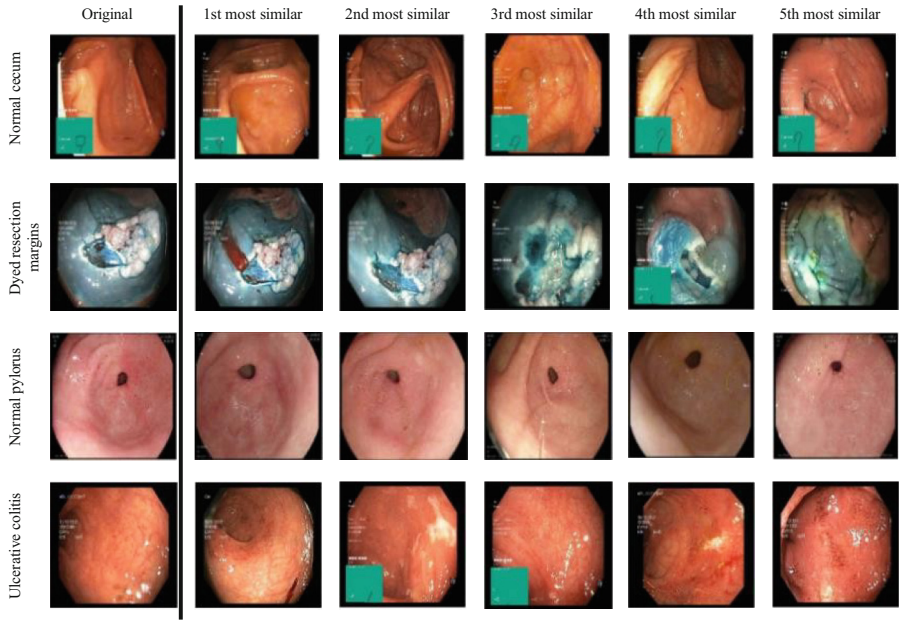

Fig. 8. Results of the similarity-based explanations technique of the ResNet50 model.

4 Discussion

4.1 Analysis of the Results of Evaluation Metrics and XAI Techniques

When analyzing the evaluation metrics presented in the tables for each model, it is clear that, although the main focus of this study was not training the models, both showed satisfactory performance in classifying the various classes present in the database. However, there were some notable difficulties in classifying images from the "dyed resection margins", "esophagitis", and "normal z-line" classes. These difficulties can be attributed to the existence of images in different classes that share similar characteristics, such as color and structure.

When applying the Grad-CAM technique to understand the predictions of each model, we observed a significant difference. The VGG16 model provided more intuitive explanations, focusing on features that are easily understood by humans. On the other hand, the explanations provided by ResNet50 seem to follow a less intuitive pattern for all the images in the database, not focusing on the specific structures of each class.

Due to this inconsistency in the explanations generated by the previous technique, we opted to implement the similarity-based explanations approach. This technique follows the logic of "This prediction is valid because similar examples in the training set also resulted in similar predictions." As a result, this approach provided the 5 most similar images from the training set for each image classified by the model. The results of this technique proved to be highly intuitive for both the VGG16 model and the ResNet50 model. It was able to present images from the training set that were remarkably similar

to the classified images, with identical structures and colors, making the explanations highly understandable and informative.

4.2 Limitations

The technique of similarity-based explanations for interpreting capsule endoscopy images, although innovative, has some significant limitations, mainly related to the quality and quantity of the data and the complexity of the deep learning models used. The effectiveness of the technique is strongly influenced by the diversity and representativeness of the data set. Limited or low-quality sets can lead to less accurate interpretations, while the presence of noise or artifacts in the images can compromise the generation of useful explanations. In addition, the complexity of the models used is another crucial factor. More complex models may offer greater precision, but their transparency and ease of interpretation are often reduced, a particularly challenging aspect in medical applications, where clarity of explanations is key. Therefore, there is an ongoing need for research to improve datasets and develop methods capable of providing clear and accurate explanations, even in highly complex models. The balance between accuracy, interpretability and practical applicability is a constant challenge in the field of XAI in medical imaging.

4.3 Perspectives and Future Work

After a thorough analysis and comparison of the results of this study, the exploration of various XAI techniques is of great importance, especially in the context of the rapid advancement of AI in various areas. It is essential to provide intuitive explanations for professionals in different sectors, given the growing adoption of complex AI models in various applications.

As part of future research, it would be highly beneficial to further explore example-based XAI techniques, such as the similarity-based explanation approach. These techniques have the potential to offer significant levels of explainability in a variety of applications, making them accessible to a wide audience. In addition, the development of new algorithms and approaches to further improve the effectiveness of these techniques can be a valuable contribution in several fields, most notably healthcare. Considering the combination of algorithms could also be a promising direction, as it could provide additional benefits in terms of the explainability and interpretability of machine learning models.

5 Conclusion

In this research, we presented an example-based XAI approach, called “similarity-based explanations,” and compared it with the widely used Grad-CAM in medical image analysis. Our results show that the example-based XAI approach offers a valuable alternative perspective on the interpretation of deep learning models. The ability to explain decisions based on examples provides a more accessible and practical understanding for healthcare professionals, improving the transparency and reliability of predictions. This

research contributes to the evolution of the field of XAI in medical imaging, highlighting the importance of interpretability in clinical applications. As deep learning models gain ground in medicine, approaches such as "similarity-based explanations" will play a crucial role in confidence and informed decision-making in key clinical scenarios.

Acknowledgements. This work is financed by National Funds through the Portuguese funding agency,FCT Fundacäo para a Ciéncia e a Tecnologia, within project PTDC/EEIEEE/5557/2020. Co funded by the European Union (grant number 101095359) and supported by the UK Research and Innovation (grant number 10058099). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

References

1. Maier, A., Syben, C., Lasser, T., Riess, C.: A gentle introduction to deep learning in medical image processing (2019). https://doi.org/10.1016/j.zemedi.2018.12.003
2. Do, S., Song, K.D., Chung, J.W.: Basics of deep learning: a radiologist's guide to understanding published radiology articles on deep learning (2020). https://doi.org/10.3348/kjr.2019.0312
3. Barredo Arrieta, A., et al.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI (2020). https://doi.org/10.1016/j.inffus.2019.12.012
4. Huff, D.T., Weisman, A.J., Jeraj, R.: Interpretation and visualization techniques for deep learning models in medical imaging (2021). https://doi.org/10.1088/1361-6560/abcd17
5. Patrício, C., Neves, J.C., Teixeira, L.F.: Explainable deep learning methods in medical imaging diagnosis: a survey (2022). https://doi.org/10.48550/arxiv.2205.04766
6. Wang, S., Xing, Y., Zhang, L., Gao, H., Zhang, H.: Deep convolutional neural network for ulcer recognition in wireless capsule endoscopy: experimental feasibility and optimization (2019). https://doi.org/10.1155/2019/7546215
7. Malhi, A., Kampik, T., Pannu, H., Madhikermi, M., Framling, K.: Explaining machine learning-based classifications of in-vivo gastral images (2019). https://doi.org/10.1109/dicta47822.2019.8945986
8. Wickstrom, K., Kampffmeyer, M., Jenssen, R.: Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation (2018). https://doi.org/10.1109/mlsp.2018.8516998
9. Lima, D.L.S., Pessoa, A.C.P., De Paiva, A.C., da Silva Cunha, A.M.T., Júnior, G.B., De Almeida, J.D.S.: Classification of video capsule endoscopy images using visual transformers. In: 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 1–4 (2022). https://doi.org/10.1109/BHI56158.2022.9926791
10. Fonseca, F., Nunes, B., Salgado, M., Cunha, A.: Abnormality classification in small datasets of capsule endoscopy images. Proc. Comput. Sci. **196**, 469–476 (2022). https://doi.org/10.1016/j.procs.2021.12.038
11. Gomes, S., Valério, M.T., Salgado, M., Oliveira, H.P., Cunha, A.: Unsupervised neural network for homography estimation in capsule endoscopy frames. Proc. Comput. Sci. **164**, 602–609 (2019). https://doi.org/10.1016/j.procs.2019.12.226
12. Smedsrud, P.H., et al.: Kvasir-capsule, a video capsule endoscopy dataset (2021). https://doi.org/10.1038/s41597-021-00920-z

13. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition (2014). https://doi.org/10.48550/arxiv.1409.1556
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (2016). https://doi.org/10.1109/cvpr.2016.90
15. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization (2017). https://doi.org/10.1109/iccv.2017.74
16. Hanawa, K., Yokoi, S., Hara, S., Inui, K.: Evaluation of Similarity-Based Explanations (2020). https://doi.org/10.48550/arxiv.2006.04528
17. Charpiat, G., Girard, N., Felardos, L., Tarabalka, Y.: Input Similarity From the Neural Network Perspective (2021). https://doi.org/10.48550/arxiv.2102.05262