Convolutional Neural Networks for the evaluation of cancer in Barrett’s esophagus: Explainable AI to lighten up the black-box

Luis A. de Souza Jr. a, b, Robert Mendel b, Sophia Strasser b, Alanna Ebigbo c, Andreas Probst c, Helmut Messmann c, João P. Papa d, f, Christoph Palm b, e

a Department of Computing, São Carlos Federal University - UFSCar, Brazil
b Regensburg Medical Image Computing (ReMIC), Ostbayrische Technische Hochschule Regensburg (OTH Regensburg), Germany
c Medizinische Klinik III, Universitätsklinikum Augsburg, Germany
d Department of Computing, São Paulo State University, UNESP, Brazil
e Regensburg Center of Health Sciences and Technology (RCHST), OTH Regensburg, Germany
f Corresponding author.
E-mail address: joao.papa@unesp.br (J.P. Papa).

https://doi.org/10.1016/j.compbiomed.2021.104578

Received 30 April 2021; Received in revised form 11 June 2021; Accepted 12 June 2021
Available online 17 June 2021

1. Introduction

Barrett’s esophagus (BE) is a condition in which the lower part of the esophagus’ mucosal cells change considerably, progressing to esophageal cancer (adenocarcinoma) in severe cases. Some risk factors may increase the number of BE-diagnosed patients, mainly in western populations [1–3]. The remission of esophageal cancer is directly related to early diagnosis, thus delivering the treatment with reduced morbidity and mortality rates, leading to complete remission after 10 years of treatment [2,4,5].

Optical coherence tomography, confocal laser endomicroscopy, and chromoendoscopy have been employed to BE and adenocarcinoma screening, enabling a more accurate manual evaluation of the esophageal histology through in vivo experiments [6]. However, BE may be misdiagnosed during endoscopy due to the inability to distinguish the columnar mucosa of the proximal stomach from the metaplastic epithelium in the distal esophagus. The Seattle biopsy protocol is highly recommended for BE lesion evaluation (i.e., dysplastic tissue), suggesting the extraction of four biopsies sampled at every 1 cm. Unfortunately, such a procedure is still susceptible to failures since the evaluated samples may not be large enough for proper screening [6–8].

The limitations related to interventional therapies (e.g., endoscopic resection and ablation techniques, which present a high potential for reducing the cancer risk in dysplasia-diagnosed patients) must be handled with monitoring methods and improved dysplasia state detection [9–11]. Computer-aided analysis of early cancerous tissue figures as an essential tool for intensive research in the past years. The prediction of peritoneal metastasis in gastric cancer patients [12] and the
automated diagnosis of breast cancer in mammograms using a Convolutional Neural Network (CNN) [13] are a few examples of recent works that make use of machine learning techniques in the context of medical imaging. The identification of early cancer in Barrett’s esophagus has also been a subject of concern in the recent years [14]. Recently conducted works evaluated the use of handcrafted features of endoscopic images based on texture and color [14]. In contrast, others assessed the well-known CNN to identify the disease automatically. Recent works proposed by Souza Jr. et al. [14–20], Mendel et al. [21], Ebigho et al. [22,23], de Groof et al. [24], van der Putten [25], Ma et al. [26], Ellis et al. [27] and Passos et al. [28] are some examples that make use of artificial intelligence (AI) techniques for automatic diagnosis.

The works conducted by Souza Jr. et al. [14–20], and Passos et al. [28] assessed the use of image representation techniques to describe and classify adenocarcinoma and Barrett’s esophagus regions. For such, the use Speeded-up Robust Features (SURF) [29] and Scale-Invariant Feature Transform (SIFT) [30], combined to Support Vector Machines (SVMs), Optimum Path-Forest (OPF) [31,32], and the infinite Restricted Boltzman Machine (iRBM) [33], were considered to provide the injured region prediction. Mendel et al. [21] and Ebigho et al. [22,23] introduced for the first time the use of deep learning techniques to classify expert-annotated esophagus samples presenting adenocarcinoma and Barrett’s esophagus, in a real-time analysis employing a ResNet-based DeepLabV3+ approach with transfer learning, while Souza et al. [19,20] extended such a work by introducing the use of Generative Adversarial Networks to the evaluation of early adenocarcinoma detection in Barrett’s esophagus context. In the study proposed by de Groof et al. [24], a hybrid ResNet-UNet architecture was proposed for a real-time detection of early neoplasia in BE-diagnosed patients, while the works proposed by van der Putten et al. [25,34] aimed at (i) achieving the same detection based on the combination of features (principal tissue-of-interest dimension encoded with the majority of useful information, such as contrast and homogeneity) and conventional machine learning techniques applied to in-vivo volumetric laser endomicroscopy samples, and (ii) detecting and localizing esophagus’ cancerous regions employing a multi-stage domain-specific pre-training technique to white-light endoscopy samples. Deep learning approaches applied to the evaluation and automatic identification of neoplasia regions in endoscopic images continue to figure an important research field.

Artificial intelligence and machine learning (ML), in general, have demonstrated remarkable performances in many tasks, especially in the medical image computing field. However, the translation of research AI systems into clinical practice depends not only on the performance of a system but also on the transparency of the decision for the physician.

Regarding the early-cancer detection in BE, the transparency is related not only to intellectual curiosity but also to risks and responsibilities intrinsic to the prediction output [35,36]. Unfortunately, the black-box nature of the deep learning techniques is still unresolved, not completely describable and presents a not trivial interpretation, leading to poorly understood decisions [37]. Especially, the question of the relationship of the suspicious region for the physician and the most important regions for the computer-based decision is of interest.

Current research suggests different methods and frameworks in the computational interpretation of CNN, making the explainable artificial intelligence (XAI) a hotspot field to be followed by the ML community. The visual explanation proposed by XAI algorithms tracks the work process of the learning techniques in visible ways, illustrating the learning process that supports its final outputs. The visual interpretation achieved by XAI techniques provides guiding posts for understanding not only correctness but errors behind the CNN learning process [38–40]. Many works have explored explainability in the medical field, such as the evaluation and analysis of sentiment with applications in medicine proposed by Zueco et al. [41], in which systematic methodologies were assessed to develop explainable Clinical Decision Support Systems. The evaluation of available interpretation models of cancer in chest radiography images has been proposed by Kallianos et al. [42], who attested the lack of effective and quantitative methods to cope with such a task. Lamy et al. [43] proposed a qualitative interpretation and detection of breast cancer in mammographic images using a case-based reasoning approach with visual outcomes. The classification of melanoma in hypertrophic cardiomyopathy-diagnosed images was conducted by Codella et al. [44], with interpretation based on an evidence-based classification using CNN features and kNN search and comparison of non-expert and automatic classifications (baseline and proposed method) using Area Under Curve similarity metrics (0.772 and 0.874, respectively).

Even with the progress related to the interpretation of deep learning decisions, there is still a long way to go in terms of interpretability, assessment, and criteria definition (in regard of notions of “interpretability”, “explainability” along with “reliability” and “trustworthiness”) [45–47].

This work aims at investigating the use of XAI techniques in the context of BE and early esophageal adenocarcinoma detection. In a quantitative fashion, our work clarifies which image regions are most important to discriminate these classes and compare them to experts’ delineations. In more details, we present the following main contributions:

- to introduce the use of XAI techniques at evaluating the classification rates in distinguishing between Barrett’s esophagus from adenocarcinoma;
- to propose a quantitative analysis of the CNN learning based on XAI techniques in the context of BE and adenocarcinoma evaluation;
- to assess whether there is an agreement in the visual interpretation of XAI techniques and visual interpretation provided by the experts in BE and adenocarcinoma image annotation;
- to investigate what is the XAI technique that provides the most accurate visual interpretation compared to the ground truth provided by different experts;
- to investigate if the agreement of XAI technique outputs and expert’s annotation is related to higher and more accurate classification results of early-cancer BE diagnosed patients.

The remainder of this work is organized as follows. Section 2 presents a brief theoretical background concerning XAI and the techniques used in the work, while Section 3 describes the methodology and the proposed method. Finally, Sections 4 and 5 state the experiments and results, as well as discussion and conclusions, respectively.

2. Explainable artificial intelligence

As long as autonomous machines and black-box algorithms make decisions entrusted to human knowledge, explaining themselves becomes necessary. Even considering the success in a wide range of tasks, including advertising, movie and book recommendations, medical assistance, and so on, there is in general mistrust about such black-box results. As the employment of black-box ML models has currently increased, to make important predictions in critical contexts, the demand for transparency has also raised from stakeholders in AI [48]. This may be justified by the fact that some output decisions are not clearly justifiable, legitimate, or with poor behavior details [47].

The explanations behind the model’s output decision are crucial in several areas that require high precision and figure experts requesting far more information from the model than just a binary prediction without extra support for such diagnosis [37]. Improvements in understanding ML systems can lead to a better definition of its parameters, helping to ensure impartiality in decision-making, i.e., to detect and correctly explain the bias for training sets and tasks. Such improvements make the entire model’s generalization more robust by highlighting potential adversarial and intrinsic problems that could harm prediction and evaluation. The explanation may be achieved by expressing which
features are meaningful in the output inference [47].

Especially in the medical domain, it is crucial that the interpretation of ML decisions is correlated to the human interpretation. Based on a previously trained ML model, the prediction and interpretation of new samples rely on their propagation through the model, with pixel impact visualizations (PIV) that may be based on layer, neuron, or prediction evaluations. This work highlights five different PIV techniques employed as tool for understanding the behavior behind CNN architecture decisions: saliency (SAL), guided backpropagation (GBP), integrated gradients (IGR), input $\times$ gradients (IXG) and DeepLIFT (DLF).

Fig. 1 illustrates the PIVs as heatmaps provided by each technique for an individual endoscopic instance.

2.1. Saliency

Saliency methods, firstly proposed by Simonyan et al. [49], perform algorithm explanations by assigning values that reflect the importance of input components in their contribution to the output prediction. These values could be part of probabilities, heatmaps, or super-pixels. For such a method, having a previously trained deep model, a given
class’s spatial support is calculated for a classified image using a single backpropagation pass through a classification step [49].

Given a fully trained CNN model and a class of interest, the saliency method provides a numerically generated image, which is representative of this specific class and which is based on the class scoring after a feedforward run through the model. This procedure is related to the model’s training, and the backpropagation optimizes the layer weights regarding the input image.

Therefore, for a given test image and a target class, the image’s pixels are ranked based on their respective influence on a score function related to the output of the classification model. Therefore, the class saliency map is calculated by means of the partial derivative of the class score regarding each test image pixel using backpropagation. Each pixel’s single class saliency value represents the derivative’s maximum magnitude across the color channels. Considering that the saliency maps are extracted using a deep model trained over image class labels, no additional annotation is required.

2.2. Guided backpropagation

The guided backpropagation technique [50] modifies the traditional backpropagation performed through the network to achieve inversion backward through a layer by zeroing negative signals from both the output or the input. To visualize the most activating image part for a specific high-level neuron (related to the highest value in the corresponding feature map), the guided backpropagation method performs a deconvolution backward pass, where negative values of either input or output are set to zero. The inversion-based method backpropagates the signal through the layers and still makes use of saliency maps for the activated signal visualization [50].

For deconvolution [51], the CNN data flow is inverted starting from a neuron activation in a higher layer down to the input image. Deconvolution and backpropagation mainly differ in the way the rectified linear unit (ReLU) is handled [50]. As an output, a reconstructed image shows the region that presents the strongest influence in activating a neuron [50].

Guided backpropagation proposes the combination of both deconvolution and backpropagation concepts, masking out only the values in which at least one is negative rather than masking out values related to negative entries of the top gradient (deconvolution) or bottom data (backpropagation). This means the corresponding neuron activation of the higher layers may be visualized, even though they present a decrease in their activation [50].

2.3. Integrated gradients and input \times gradients

While saliency and guided backpropagation focus on the score gradients regarding the input image, input \times gradients also takes the input value into account. The attributes are computed by means of a pixel-wise multiplication of the input value and the gradient for a specific pixel [52].

However, all these gradient-based methods share the same problem: if the gradient vanishes during the backpropagation task, the respective pixel’s impact diminishes. However, a pixel should present high impact if its existence makes a difference to the prediction outcome.

Therefore, the integrated gradients technique [53] not just compute the gradient once for each input pixel, but instead, for a fixed input image, a sequence of m intensity downsampled versions is generated applying a multiplication by \frac{1}{n} (R = 1, ..., m). This sequence simulates the stepwise vanishing of the signal at each pixel position. Then, the integrated gradients method sums up all the gradients related to the images of the sequence. Finally, and similar to input \times gradients, this aggregated gradient is multiplied by the pixel intensity over all image positions, respectively.

2.4. DeepLIFT

Deep Learning Important Features (DeepLIFT), proposed by Shrikumar [54], is a further development of integrated gradients. In such a method, the contributions of an input pixel to the output score are measured in relation to a reference image. This reference image should describe the unimportant background, which can be a constant zero image as a first choice. Given that reference, the output difference-from-reference value is explained in terms of the inputs difference-from-reference value. These contributions are split into positive and negative parts and backpropagated to all the neurons down to the input. Referring to a background baseline enables DeepLIFT to reveal dependencies among input pixels.

3. Methods and material

This section presents the methodology adopted to cope with the XAI interpretation and evaluation of Barrett’s esophagus and adenocarcinoma data.

3.1. Method

As mentioned earlier, one of the primary contributions of this work is to provide an interpretability evaluation of positive samples in cancer diagnosed images using XAI techniques and a comparison of segmentation outputs based on the experts’ annotations. To fulfill that purpose, we considered four different CNN architectures as models to be trained and validated for the XAI prediction interpretation, e.g., AlexNet, SqueezeNet, ResNet50, and VGG16, illustrated in Fig. 2. Such architectures were considered due to their extensive usage in the literature, but any other model could also be used. With several models, a more robust and cohesive interpretation of deep networks could be delivered, aiming to understand the critical regions of BE and adenocarcinoma context. Moreover, it is imperative to assess how different CNN architectures can deal with the problem addressed in this work and whether they express or not meaningful regions for early-stage adenocarcinoma prediction during the classification step.

Algorithm 1 summarizes the approach proposed in this study to quantitatively evaluate the XAI techniques. The output prediction of each CNN model (after performing training and testing tasks) is provided based on two different validation protocols, i.e., the leave-one-patient-out (LOPO-CV) or the 20-fold (lines 3–5). Further, for all true positive (TP) and false negative (FN) samples (inner loop in lines 6–12), the XAI heatmaps are calculated using five different XAI techniques: saliency, guided backpropagation, integrated gradients, input \times gradients, and DeepLIFT (lines 6–7). Considering that the XAI output is the pixel attribution of each evaluated sample, such attributes are normalized (line 8) for computing the OTSU threshold [55] and producing a segmentation mask to be compared with the respective experts’ annotation (lines 9–10). Three agreement measures are then employed over such manual and automatic segmentations: Cohen-Kappa (CK) [56], intersection-over-union (IoU) and pixel accuracy (PA) (percentage of segmented pixels inside the expert’s annotated area) (lines 11–12).

3.2. Datasets

Two high-definition white-light endoscopic datasets were used for performing an in-depth analysis of the proposed approach. The first dataset is composed of endoscopic examinations provided by the University Hospital Augsburg, Medizinische Klinik III, Germany. The dataset comprises a total of 76 endoscopic images captured from different BE-diagnosed patients, in which 42 present only BE and 34 present BE and early-stage adenocarcinoma. One physician manually annotated the cancerous biopsy-diagnosed images. Fig. 3 displays some images from
Fig. 2. Illustration of the selected models to perform the prediction interpretation: (a) AlexNet, (b) SqueezeNet, (c) ResNet50, and (d) VGG16.
the Augsburg dataset labeled as positive to adenocarcinoma.

The second dataset is composed of images from a benchmark dataset available at the “MICCAI 2015 EndoVis Challenge”, and called “MICCAI”. Such a dataset was published to encourage researchers to conduct studies for differentiating BE and early-cancerous images concerning how similar they look. Comprising 100 endoscopic images of the lower esophagus, the samples of such dataset were captured from 39 individuals. The MICCAI dataset presents 22 samples diagnosed as BE and 17 diagnosed as early-stage esophageal adenocarcinoma. Each patient figures a different amount of endoscopic images for this dataset, ranging from one to a maximum of eight. The dataset presents, in total, 50 images displaying cancerous tissue areas and 50 images showing BE disease. Five different experts have individually annotated suspicious regions in cancerous samples. Fig. 4 depicts some samples diagnosed as positive to adenocarcinoma from MICCAI dataset and their five respective experts’ annotations.

3.3. Experimental setup

This section presents the methodology used to conduct the model’s definition, classification, and interpretation steps for further evaluation of the results.

3.3.1. Deep model definition

To cope with the model generation step, four different CNN architectures were evaluated, i.e., AlexNet [57], SqueezeNet [58], ResNet50 [59] and VGG16 [60]. The main rationale behind such choices are: (i) to evaluate the accuracy, sensitivity and specificity of current and trending CNN architectures in the context of BE and adenocarcinoma diagnosis and; (ii) to understand the input discriminative parts learned by each CNN architecture and further assess their agreement with the human-expert annotation of early-cancerous regions in positive samples. All networks’ parameters were kept constant, with batch size of 4, and a learning rate of 0.001. The Adam optimizer [61] was used with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Adam optimization has been widely employed in several

Algorithm 1: Proposed method’s algorithm.

Result: Agreement Measures: CK, IoU and PA
1. Initialization: protocol definition (LOPO or 20-FOLD);
2. repeat
3. Definition of training and testing sets;
4. Train CNN architectures;
5. class ← Classify testing samples;
6. for $img$ as TP and FN samples of class do
7.  $attributes_{img} ← img$ interpretation using XAI techniques
   (saliency, gpb, igr, ixg, and dlf);
8.  $attributes_{nx} ←$ normalization of $attributes_{img}$;
9.  $threshold ←$ OTSU threshold of $attributes_{nx}$;
10. $attribute_{sk} ←$ binarization of $attributes_{nx}$ based on
     threshold;
11. agreement_measures ←
    segmentation_comparison($attribute_{sk}$, GroundTruth$_{img}$),
    with segmentation_comparison ∈
    (Cohen Kappa, Intersection over Union and Pixel Accuracy);
12. end
13. until protocol’s size;

Fig. 3. Some positive-to-adenocarcinoma images from the Augsburg dataset and their respective delineation.
classification tasks, due to its computationally efficient, little memory requirement, invariance to diagonal rescaling of the gradients, and well adaptation to problems that are large in terms of data and/or parameters. Also, such an optimization technique presents an intuitive interpretation and does not typically require intense tuning, being appropriate for non-stationary objectives and problems with very noisy and/or sparse gradients [19,20,62]. Therefore, we adopted standard values for all CNN’s model parameters, defined empirically, considering the main scope was to assess the connection between the computational model and human interpretations in the identification of early cancer in BE samples.

The experiments were conducted over 12,000 epochs, generating classification models based on two different protocol approaches, i.e., 20-fold cross-validation and LOPO-CV. In the 20-fold cross-validation approach, 80% of samples were randomly selected for training, and the remaining 20% were used for testing in each experimental fold. In the LOPO-CV approach, at each iteration, a different patient was taken out of the entire set for testing purposes, while the remaining samples were used as a training set. Using four CNN architectures and two different protocols, eight different models for each dataset are provided to be interpreted using the XAI techniques.

3.3.2. Explanable artificial intelligence evaluation

The XAI interpretation was conducted to assess the most discriminative region of each sample positive-to-cancer, regarding all five different XAI techniques applied to this study: SAL, GBP, IGR, IXG and, DLF.

It is clearly important to understand which regions influenced the class prediction of samples and if the number of pixels inside such regions matches the expert’s annotations of cancerous regions. This may give insight into the correlation of the agreement between humans and computational learnings in the definition of early-stage cancerous tissues for BE and adenocarcinoma samples.

The XAI techniques present as the output the evaluation of each pixel in the input samples classified as TP or FN for the target prediction, i.e., for each pixel, attributions are calculated with values correlated to its impact in the predicted class. A zero image was used as standard baseline to the techniques that are based on the first baseline assumption (integrated gradients and DeepLIFT). After the XAI heatmap calculation, a min-max normalization was performed for each image. However, the histograms of the pixel values for each image and XAI method are different. Therefore, and in a way to define the best binarization threshold for each sample, the OTSU threshold [55] was calculated to differentiate between meaningful and non-meaningful attributes. This binarized output can then be compared to the experts’ ground truth.

Furthermore, the comparison between the XAI segmented output and the ground truth of positive samples was performed employing three different techniques: CK, IoU, and PA. Along with the assessment of computational-and-human agreement for positive samples, the comparison between TP and FN predictions may be conducted. The hypothesis is that there is a low agreement of ground truth and meaningful pixels for each incorrectly predicted sample.

Last but not least, the correlation between the segmentation measures and the sensitivity rates of each CNN architecture was performed employing the Spearman’s Correlation Test [63]. With such a test, the agreement between computational and human segmentations in the correct classification of cancerous samples can be highlighted, sharply increasing the trustworthiness related to the CNN interpretation step.

4. Experimental results

This section presents the experiments used to evaluate the proposed methodology. The first round of experiments aimed at evaluating all the CNN architectures using the following Accuracy (A), Sensitivity (S), and Specificity (P) rates:

\[
S = \frac{TP}{TP + FN} \times 100, \tag{1}
\]

\[
P = \frac{TN}{TN + FP} \times 100, \tag{2}
\]

and

\[
A = \frac{TP + TN}{TP + TN + FP + FN} \times 100, \tag{3}
\]

with true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Additionally, a statistical evaluation using the signed-rank Wilcoxon test [64] was considered for comparison purposes between the segmentation measures over the XAI interpretation results.

The experiments were conducted on a 96 GB-memory computer equipped with an NVIDIA TitanX® Graphics Card of 12 GB and implementations were made using PyTorch and Captum.

4.1. Results on augsburg dataset

In this section, the Augsburg dataset results regarding the classification and XAI interpretation are presented.

---

Fig. 4. Some positive-to-adenocarcinoma images from the MICCAI dataset and their respective delineation.
Table 1
Mean classification rates and time-consuming for the training task considering both 20-fold and LOPO-CV validation protocols for the Augsburg dataset. The best results for each protocol are highlighted in bold, and the best overall result for each rate is marked with a ∗ symbol.

| Protocol   | Architecture | Accuracy (±) | Sensitivity (±) | Specificity (±) | Time (min)  |
|------------|--------------|--------------|----------------|----------------|-------------|
| 20-fold    | AlexNet      | 0.80 ± 0.04  | 0.72 ± 0.005   | 0.86 ± 0.003   | 111.09 ± 19.13 |
|            | SqueezeNet   | 0.73 ± 0.09  | 0.67 ± 0.005   | 0.80 ± 0.003   | 175.23 ± 18.36 |
|            | ResNet50     | 0.76 ± 0.07  | 0.75 ± 0.01    | 0.77 ± 0.007   | 221.39 ± 20.33  |
|            | VGG16        | 0.82 ± 0.13  | 0.81 ± 0.015   | 0.84 ± 0.002   | 137.14 ± 15.58  |
| LOPO-CV    | AlexNet      | 0.83 ± 0.04  | 0.81 ± 0.005   | 0.84 ± 0.002   | 139.20 ± 19.25  |
|            | SqueezeNet   | 0.79 ± 0.06  | 0.73 ± 0.01    | 0.83 ± 0.005   | 201.15 ± 21.37  |
|            | ResNet50     | 0.84 ± 0.12  | 0.87 ± 0.015   | 0.84 ± 0.002   | 244.66 ± 19.17  |
|            | VGG16        | 0.84 ± 0.12  | 0.83 ± 0.014   | 0.87 ± 0.007   | 152.21 ± 12.11  |

Table 2
CK, IoU, and PA mean values for the best XAI interpretation output of 20-fold and LOPO-CV validations over the Augsburg dataset. The best results for each protocol are highlighted in bold, and the best overall result for each measure is marked with a ∗ symbol.

| Protocol   | Architecture | Prediction | Best XAI technique | CK (±)   | IoU (±)   | PA (±)   |
|------------|--------------|------------|--------------------|---------|-----------|---------|
| 20-fold    | AlexNet      | TP         | SAL                | 0.30 ± 0.03 | 0.25 ± 0.03 | 0.59 ± 0.04 |
|            | SqueezeNet   | TP         | SAL                | 0.14 ± 0.02 | 0.09 ± 0.09 | 0.34 ± 0.10 |
|            | ResNet50     | TP         | SAL                | 0.33 ± 0.03 | 0.24 ± 0.06 | 0.51 ± 0.15 |
|            | VGG16        | TP         | SAL                | 0.33 ± 0.02 | 0.26 ± 0.01 | 0.59 ± 0.15 |
| LOPO-CV    | AlexNet      | TP         | SAL                | 0.18 ± 0.03 | 0.12 ± 0.01 | 0.28 ± 0.03 |
|            | SqueezeNet   | TP         | SAL                | 0.33 ± 0.06 | 0.25 ± 0.04 | 0.62 ± 0.07 |
|            | ResNet50     | TP         | SAL                | 0.32 ± 0.08 | 0.24 ± 0.08 | 0.57 ± 0.13 |
|            | VGG16        | TP         | SAL                | 0.36 ± 0.09 | 0.25 ± 0.06 | 0.55 ± 0.12 |

4.1. Classification
Table 1 presents the mean classification results related to the Augsburg dataset. Regarding the 20-fold cross-validation approach, one can draw the following conclusions: (i) VGG16 model presented the highest accuracy (83.37% ± 26.72%) and sensitivity (80.59% ± 27.56%) results; and (ii) AlexNet classifier provided the best specificity mean result (86.19% ± 22.59%). Concerning the LOPO-CV approach results, the following outcomes could be observed: (i) VGG16 model presented the highest accuracy (84.37% ± 22.93%) and specificity (86.62% ± 19.54%) results and; (ii) ResNet50 presented the best sensitivity rate (87.05% ± 8.84%) among all configurations. The LOPO-CV results present higher accuracy, sensitivity, and specificity outcomes in most experimental cases due to the training sets for such protocols comprising more samples in a “patient-based” delineation for the classification step.

4.1.2. XAI interpretation
Table 2 presents the best results related to the XAI interpretation of the Augsburg dataset when SAL, GBP, IGR, IXG and DLF were applied to the model interpretation task, and CK, IoU, and PA measures were applied to assess the computational and human segmentation agreement. In the evaluation of the 20-fold cross-validation approach, the saliency-based technique, when compared to the experts’ segmentations using all three agreement measures, provided the best results for every conducted experiment, outperforming the agreement observed in all remaining XAI techniques. In addition, the agreement between computational and human interpretations for TP samples was statistically higher for all the presented results (Table 2) compared to the FN ones. Concerning CK, IoU, and PA measures, the best results were obtained evaluating TP inputs of ResNet50 architecture, i.e., 0.332 ± 0.023, 0.258 ± 0.023, and 0.590 ± 0.151, respectively.

The saliency technique applied to the LOPO-CV approach achieved
the best results in all conducted experiments, outperforming both the results observed for the TP inputs using the other XAI techniques and the 20-fold CV outputs. For the Augsburg dataset, the obtained agreement between computational and human segmentations was statistically higher for TP samples than the FN ones. Concerning CK and IoU measures, the best mean results were obtained in the interpretation of TP samples using the VGG16 model, i.e., 0.357 ± 0.092 and 0.250 ± 0.056, while the best PA result was obtained for the interpretation of SqueezeNet positive-predicted samples, i.e., 0.622 ± 0.070.

The saliency method provided very satisfactory results for the BE positive-classified samples, showing the efficiency and importance of such a technique to interpret the sensitivity values during the classification process of the evaluated CNN models.

4.2. Results on MICCAI dataset

This section presents the results concerning the MICCAI dataset considering the classification and XAI interpretation experiments.

4.2.1. Classification

Table 3 presents the mean results related to the MICCAI dataset. Regarding the 20-fold cross-validation approach, one can draw the following conclusions: (i) VGG16 model presented the highest accuracy (83.73% ± 23.83%) and specificity (85.16% ± 38.74%) results; and (ii) ResNet50 classifier provided the best sensitivity mean value (88.77% ± 11.75%). Concerning the LOPO-CV approach results, the following outcomes could be observed: (i) ResNet50 architecture presented the highest accuracy (86.55% ± 11.63%) and sensitivity (88.51% ± 7.84%) results; and (ii) VGG16 classifier showed the best specificity rate (89.5% ± 10.21%) among all configurations. The LOPO-CV results present higher accuracy, specificity, and sensitivity outcomes in most experimental cases due to the training sets for such protocols comprising more samples in a “patient-based” classification goal.

4.2.2. XAI interpretation

Table 4 presents the best results related to the XAI interpretation of MICCAI dataset using the five XAI evaluation techniques and the segmentation comparison measures for TP and FN classified inputs of each CNN architecture. Regarding the 20-fold cross-validation approach, the saliency technique provided the best agreement results for every conducted experiment once again, outperforming the results observed in all remaining XAI techniques. It is also important to highlight that, for the TP samples, the obtained agreement between computational and human segmentations was statistically higher for all the experiments when compared to the FN ones. Concerning CK and IoU measures, the best results were obtained in the interpretation of TP inputs using VGG16’s model (0.311 ± 0.039 and 0.293 ± 0.014), while the best PA measure was obtained in the interpretation of TP samples predicted by the ResNet50 model (0.582 ± 0.081).

For the LOPO-CV approach, one can observe that the saliency XAI technique provided, once again, the best results in almost every conducted experiment, outperforming not only the agreement observed in all remaining XAI techniques but also outperforming the 20-fold CV outputs. Again, it is important to highlight that the obtained agreement between computational and human segmentations was statistically higher for the TP samples compared to the FN results of the same measures. The best-obtained results for the segmentation measures in this protocol were: for CK and IoU obtained in the interpretation of TP inputs from AlexNet’s classification (0.324 ± 0.058 and 0.318 ± 0.025), while the best PA result was obtained in the interpretation of TP samples predicted with the SqueezeNet architecture (0.642 ± 0.132).

As one can also observe, for the MICCAI dataset, the saliency method provided very satisfactory results for the positive-classified observation of BE context. The higher agreement, when compared to the remaining techniques, gives us insights into how interesting the use of such technique is to understand the interesting regions related to the correct and wrong classification of cancerous samples.

4.3. Correlation test

The correlation test refers to the sensitivity value and the final interpretation presented by the use of XAI techniques. For such a task, the agreement measures and sensitivities of both TP and FN classified-and-interpreted samples were considered in the evaluation of each CNN model, taking its best protocol result into account. Table 5 presents the results of Spearman’s correlation test for the very best results achieved for each CNN architecture in the evaluation of both datasets, with bold values meaning the highest achieved correlation between sensitivity and agreement measure.

After interpreting TP and FN samples by applying the methodology, we could observe that saliency technique was clearly more related to the experts’ annotations than the others. Such a method provides the heatmap based on the calculated gradients of the target class, and for almost every experimental delineation, presented more attributes accorded to the experts’ annotation region. As a result, all CK, IoU, and PA measures were higher for saliency maps, suggesting that this technique may work better than the remaining ones when dealing with observation and description of similar tissues of different natures, as BE and early-cancer are presented in the endoscopic instances. Besides, one can observe Figs. 5 and 6, in which the best interpretation outputs of TP samples (from the best XAI technique) are presented for each CNN architecture. When analyzing such information, the attributes’ incidence inside the physicians’ delineations may be highlighted, even the techniques showing up that there were still relevant parts for the correct classification of positive samples outside of the agreement regions. Still, when comparing CK, IoU and PA measures observed for TP and FN samples in all experiments, FN-diagnosed segmentations presented lower values, corroborating the insights about the correlation among correct classification and region of interest agreement.

5. Discussion and conclusions

In this paper, we dealt with computer-assisted Barrett’s esophagus and adenocarcinoma identification, interpretation, and comparison by...
means of deeply-learnable features computed using AlexNet, SqueezeNet, ResNet50 and VGG16 networks. Such architectures were selected to make sure a robust evaluation of the approaches proposed in this paper. Well-known and widely employed models, ranging from simple (AlexNet) to sophisticated-and-deeper (ResNet50) architectures, were selected to conduct the first quantitative comparison of human and computation interpretation of early cancer definition in the esophagus region.

We could not observe any previous study proposing the interpretation of deep learning models in the BE and adenocarcinoma context to provide visual insights into the learning process. In this work, we fostered the research toward such tasks by introducing an interpretation of CNN classification based on XAI techniques, in both qualitative and quantitative assessment of it. Thus, we could extend some recently proposed works over similar database and protocol [14–18]. Besides, we highlight three works that proposed the use of some interpretation technique to light up the learning process behing the deep model generalization. The first one, proposed by Gu et al. [65], employs the use of extreme gradient boosting to predict breast cancer and case-based reasoning to explain the computational decisions. The second one, conducted by Moncada-Torres et al. [66], designed a system for interpreting breast cancer prediction based on several ML models and

### Table 4

| Protocol | Architecture | Prediction | Best XAI technique | CK       | IoU       | PA       |
|----------|--------------|------------|--------------------|----------|-----------|----------|
| 20-fold  | AlexNet      | TP         | SAL                | 0.31 ± 0.03 | 0.29 ± 0.01 | 0.51 ± 0.12 |
|          | SqueezeNet   | TP         | SAL                | 0.26 ± 0.02 | 0.20 ± 0.01 | 0.57 ± 0.10 |
|          | ResNet50     | TP         | SAL                | 0.20 ± 0.04 | 0.09 ± 0.04 | 0.19 ± 0.08 |
|          | VGG16        | TP         | SAL                | 0.31 ± 0.04 | 0.29 ± 0.01 | 0.51 ± 0.05 |

**LOPO-CV**

| Protocol | Architecture | Prediction | Best XAI technique | CK       | IoU       | PA       |
|----------|--------------|------------|--------------------|----------|-----------|----------|
|          | AlexNet      | FN         | SAL                | 0.11 ± 0.07 | 0.07 ± 0.01 | 0.19 ± 0.05 |
|          | SqueezeNet   | TP         | SAL                | +0.32 ± 0.06 | +0.32 ± 0.03 | 0.63 ± 0.06 |
|          | ResNet50     | FP         | SAL                | 0.28 ± 0.06 | 0.22 ± 0.05 | +0.64 ± 0.13 |
|          | VGG16        | FP         | GBP                | 0.32 ± 0.03 | 0.30 ± 0.02 | 0.56 ± 0.14 |

### Table 5

| Dataset   | CNN            | Best Protocol | Best XAI Technique | Spearman’s Correlation Test | Sensitivity × CK | Sensitivity × IoU | Sensitivity × PA |
|-----------|----------------|---------------|--------------------|----------------------------|------------------|------------------|------------------|
| Augsburg  | AlexNet        | LOPO CV       | SAL                | 0.53                       | 0.51             | 0.62             |
| SqueezeNet| LOPO CV        | SAL           | 0.48               | 0.52                       | 0.57             |
| ResNet50  | 20-fold CV     | SAL           | 0.62               | 0.63                       | 0.61             |
| VGG16     | LOPO CV        | SAL           | 0.44               | 0.55                       | 0.63             |
| MICCAI    | AlexNet        | LOPO CV       | SAL                | 0.51                       | 0.61             | 0.71             |
| SqueezeNet| LOPO CV        | SAL           | 0.41               | 0.66                       | 0.68             |
| ResNet50  | 20-fold CV     | SAL           | 0.61               | +0.67                      | +0.72            |
| VGG16     | LOPO CV        | GBP           | +0.63              | 0.64                       | 0.69             |
Shapley Additive exPlanation. The last one from Sabol et al. [67] showed improvements in the accountability in decision-making of colorectal cancer comparing CNN model outcomes from a novel XAI model that, besides the classification, presents: (i) a semantical explanation, (ii) a visualization of the training image most responsible for a given prediction, and (iii) a visualization of training images of other types of tissues to explain the decision. All the aforementioned works show promising interpretation of ML model decisions. Obviously, the XAI application to explain, in details, the ML decisions is crucial, but the proper quantitative evaluation is necessary to make such an interpretation robust enough, not only relying on the visual evaluation of experts’ and computational outcomes. Therefore, our method proposes an interpretation completely based on the quantitative correlation of manual and automatic explanations of the decisions, not only relying on visual insights of its outcomes.

The interpretation of deep learning outcomes seems to be a must-do task, once machine learning techniques have been widely applied to the medical field with promising results through the years. As long as the results are improved, the decisions behind the black-box generalization, learning process, and evaluation of samples must be understood, driving to the experts’ insights about how the problem was dealt, and further directions for closer observations of regions they did not observe previously. Along with the CNN model, XAI tools can interpret new-classified samples and further evaluate positive correct and incorrect classifications. Five different techniques were assessed for the interpretation task, where after the heatmap been calculated, the segmentation output was calculated based on the most discriminative features provided by them. This allows to test the hypothesis that models with high sensitivity results correspond to models with high agreement between high impact attributes and experts’ annotations. Hence, the agreement of human and computational annotations was performed using the CK, IoU, and PA measures to satisfy or deny such a hypothesis.

The Spearman’s correlation test was conducted to understand if the computational-segmented images really presented correlation to the sensitivity results for both TP and FN interpretations. As one can observe in Table 5, the achieved relationship for all experimental sets lay on positive moderate to positive strong correlation among sensitivity and agreement measures (moderate is the range within ±0.30 and ±0.49, and strong is the range within ±0.5 and ±1) [68].

In an in-depth analysis of the correlation results, one can observe that, with positive moderate and strong correlations, as long as the number of interpreted pixels inside the experts’ annotations increases for the correct classification, the sensitivity result also increases. Considering the highest correlation result (i.e., MICCAI dataset for PA agreement measure in ResNet50 interpretation), one can observe that such outcome suggests how important the agreement of computational and human region definitions is for the correct classification of positive-to-cancer samples in the evaluated context. In addition, the correlation values for TP and FN were quite far from each other. Moreover, the same behavior can be observed for the remaining correlation results. For the higher ones, the difference between TP and FN measures was also higher, showing that the sensitivity increasing may also be related to lower FN agreement within human-annotated and computational-annotated regions. The same outcome could be observed in many

Fig. 5. Computational segmentation of TP samples (black) and their respective ground truth delineated area (red) over (a) AlexNet, (b) SqueezeNet, (c) ResNet50, and (d) VGG architectures for Augsburg dataset. Every segmented sample is related to the best XAI interpretation technique obtained for the respective CNN architecture.
other correlation assessments. However, some high sensitivity values did not present strong correlation to the segmentation measures (but moderate), suggesting that not only the agreement region is important for the correct classification of positive samples, but also the attributes highlighted on the outside (see Figs. 5 and 6). Furthermore, even presenting satisfactory human-and-computation agreement in the TP evaluation sets, a deeper look into such defined attributes could be performed to find, perhaps, more discriminative regions for cancerous instance sampling. The fuzzy-regions, defined as areas in which the experts’ annotations do not agree in the manual definition of the ground truth, may be considered as a potential discriminative region for the attribute definition.

Again, for the Augsburg dataset, the achieved results (even for the correlation task) were lower than the MICCAI outcomes, reinforcing that the evaluation of such a dataset is more challenging not only for the classification but also for the interpretation of the results. Even though, still satisfactory outcomes could be achieved, outperforming the state-of-the-art full image classification and interpretation of positive samples.

Finally, the main achieved contributions of the study are presented as follows:

- The interpretation of black-box generalization in BE endoscopic images based on XAI techniques showed up to be promising, presenting trustworthy outputs to be compared to experts’ interpretations of the same problem and encouraging new studies in which cancerous samples must be interpreted after deep learning generalization.

- The saliency technique, based on the interpretation of input’s gradients, achieved the best results, suggesting promising behavior in the interpretation of cancerous tissues.

- The proposed hypothesis about “how related are the computational and human learnings for BE and adenocarcinoma context?” could be answered after the conducted correlation experiments, in a conclusion that, yes, the experts’ annotated regions present from moderate to strong correlation to the correct classification of cancerous samples using black-boxes models, even though outside regions defined as important by such deep learning architectures also present relevance for the correct and incorrect predictions. Moreover, we conclude that the FN-classified samples always presented lower agreement of important regions with experts’ annotations, corroborating the importance of such delineations in the computational learning and classification processes.

Similarly to the results obtained in this work, the study proposed by Souza Jr et al. [15] aimed at understanding the impact of the handcrafted-feature localization on the class prediction of cancerous-tissue over BE samples. For such, the authors evaluated the position and amount of features inside the cancerous region, concluding that the higher the number of features inside such a region, the higher the model’s capability of correctly predicting cancerous samples. Considering the nature of object detection techniques, such as SURF and SIFT (assessed in the mentioned work), it is extremely important not only to perform the same evaluation for CNN architectures (considering the high challenge in solving the black-box learning process) but to highlight that the same interpretation could be achieved, suggesting
once again the importance of defining the correct cancerous region for its correct description, learning and classification.

Regarding future work, we aim to consider more sophisticated and deeper CNN architectures, such as GoogleNet and DenseNet, for the model generalization task and to compare the results with more pixel-wise XAI techniques. Additionally, a layer-wise interpretation will be conducted to assess each layer’s importance in the interpretation of positive sample generalization and classification in BE and adenocarcinoma context. Moreover, we aim at validating the proposed method in more datasets, when available.

Acknowledgements

The authors are grateful to Capes/Alexander von Humboldt Foundation grant number BEX 0581-16-0, CNPq grants 306166/2014-3 and 307066/2017-7, as well as FAPESF grants 2013/0735-0; 2014/12236-1, and 2016/19403-6, 2017/04847-9 and 2019/08605-5.

References

[1] J. Lagergren, P. Lagergren, Oesophageal cancer, BMJ 341 (2010), https://doi.org/10.1136/bmj.c6280.

[2] J. Dent, Barret’s esophagus: a historical perspective, an update on core practicalities and prospects on future evolution. J. Gastroenterol. Hепatol. 26 (2011) 11-30, https://doi.org/10.1111/j.1440-1746.2010.06535.x.

[3] C. Lepage, B. Rachet, V. Jonste, Continuing rapid increase in esophageal adenocarcinoma in California and Wales, 2005. A.cwd, L. A. Ebi (2008) 2694-2699, https://doi.org/10.1157/132486-2008.02191.x.

[4] P. Sharma, J.J. Bergman, K. Goda, M. Kato, H. Messmann, B.R. Alsop, N. Gupta, P. Vennalaganti, M. Hall, V. Konda, A. Koons, D. L. Oden, J. R. Goldblum, J. W. Zhang, M. E. Eisen, M. H. Johnston, J. A. Eastone, J. Horwhat, J. Cartledge, J. S. Mathews, J. Foggy, L. A. de Souza Jr. et al. [12] S. Mirniaharikandei, M. Heidari, G. Danala, S. Lakshmivarahan, B. Zheng, Applying a random projection algorithm to optimise machine learning model for predicting persistent patients in gastric cancer patients using in artificial intelligence, Comput. Methods Progr. Biomed. 200 (2021) 105937, https://doi.org/10.1016/j. cmpb.2020.105937, URL, http://www.sciencedirect.com/science/article/pii/S0169260720317466.

[13] T. Latzchadis, P. Koutla, L. Costardou, I. Pratikakis, Integrating segmentation information into cme for breast cancer diagnosis of mammographic masses, Comput. Methods Progr. Biomed. 200 (2021) 105931, https://doi.org/10.1016/j. cmpb.2020.105931, URL, http://www.sciencedirect.com/science/article/pii/S0169260720317466.

[14] L.A. Souza Jr., C. Palm, R. Mendel, C. Hook, A. Ebigbo, A. Probst, H. Messmann, S. Weber, J.P. Papa, A survey on Barret’s esophageal analysis using machine learning, Comput. Biol. Med. 121 (2020) 103761, https://doi.org/10.1016/j. compbiomed.2020.103761.
A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, J. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: the A. D. Preece, D. Harborne, D. Braines, R. Tomsett, S. Chakraborty, Stakeholders in L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining F. Doshi-Velez, B. Kim, in: Towards a Rigorous Science of Interpretable Machine M. T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you? S. Lapuschkin, S. W. S. Tonekaboni, S. Joshi, M.D. McCradden, A. Goldenberg, What Clinicians Want: L. A. de Souza Jr. et al. k.uni-freiburg.de/Publications/2015/DB15a . org/abs/1810.00184 . arXiv:1810.00184. URL http://arxiv.org/10.1016/j.inffus.2019.12.012 . International Conference on Knowledge Discovery and Data Mining, KDD classifications of any classifier, in: Proceedings of the 22nd ACM SIGKDD Conference on Machine Learning - Volume 79, ICML 17, JMLR.Org, 2017, pp. 3145–3153, https://doi.org/10.5555/3305890.3306006. C. Yu, C. Diaan-reen, L. Yang, C. Lei, Otsu’s thresholding method based on gray level-gradient two-dimensional histogram, in: 2010 2nd International Asia Conference on Informatics in Control vol. 3, Automation and Robotics, 2010, pp. 282–285, https://doi.org/10.1109/CAR.2010.5466878, 2010. M. McFihugh, Interater reliability: the kappa statistic. Biochemica medica : casopis Hrvatskoga društva medicinskih biokemiaka/, HDDM 22 (2012) 276–282, https://doi.org/10.11613/BM.2012.031. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagent classification with deep convolutional neural networks, in: F. Pereziz, C.J.C. Burges, L. Bottou, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25, Curran Associates, Inc., 2012, pp. 1097–1105, https://doi.org/10.1145/3065386. F.N. Iandola, M.W. Moskewicz, K. Ashrafz, S. Han, W.J. Dally, K. Keutzer, SqueezeNet: Alexnet-Level Accuracy with 50x Fewer Parameters and C Model Size, CoRR Abs/1602.07360. arXiv:1602.07360, http://arxiv.org/abs/1602.07360. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR), 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.92. K. Simonay, A. Zisserman, Very deep convolitional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015, URL, http://arxiv.org/abs/1409.1556. San Diego D.P. Kingma, J. Ba, Adam, A method for stochastic optimization, in: 3rd International Conference for Learning Representations, 2015, 2014. D.P. Kingma, J. Ba, Adam, A method for stochastic optimization, in: ICLR (Poster), 2015, URL, http://arxiv.org/abs/1412.6980. A.D. Lovie, Who discovered Spearman’s rank correlation? Br. J. Math. Stat. Psychol. 48 (2) (1995) 255–269, https://doi.org/10.1111/j.2044-8317.1995.tb01063.x. F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bull. 1 (6) (1945) 80–83, https://doi.org/10.2307/3001968. D. Gu, K. Su, H. Zhao, A case-based ensemble learning system for explainable breast cancer recurrence prediction, Artif. Intell. Med. 107 (2020) 101858, https://doi.org/10.1016/j.artmed.2020.101858, https://www.sciencedirect.com/science/article/pii/S0933365719300617. A. Moncada-Torres, M. van Maaren, M. Hendriks, et al., Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival, Sci. Rep. 11 (2021) 6968, https://doi.org/10.1038/s41598-021-86327-7. P. Sabol, P. Sincák, P. Hartono, P. Kocán, Z. Benetinová, A. Bličharová, Ludmila Verbová, E. Štammová, A. Sabolová-Fabianová, A. Jašková, Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images, J. Biomed. Inf. 109 (2020) 103523, https://doi.org/10.1016/j.jbi.2020.103523. https://www.sciencedirect.com/science/article/pii/S0933365719300617. Spearman rank correlation coefficient, in: The Concise Encyclopedia of Statistics, Springer, New York, New York, NY, 2008, pp. 502–505, https://doi.org/10.1007/978-0-387-32833-1_379.