Preliminary Study of Deep Learning Algorithms for Metaplasia Detection in Upper Gastrointestinal Endoscopy

Alexandre Neto[1,2], Sofia Ferreira[2], Diogo Libânio[3], Mário Dinis-Ribeiro[3], Miguel Coimbra[1,4], and António Cunha[1,2(✉)]

[1] Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência, 3200-465 Porto, Portugal
acunha@utad.pt

[2] Escola de Ciências e Tecnologia, Universidade de Trás-os-Montes e Alto Douro, Quinta de Prados, 5001-801 Vila Real, Portugal

[3] Departamento de Ciências da Informação e da Decisão em Saúde/Centro de Investigação em Tecnologias e Serviços de Saúde (CIDES/CINTESIS), Faculdade de Medicina, Universidade do Porto, 4200-319 Porto, Portugal

[4] Faculdade de Ciências, Universidade do Porto, 4169-007 Porto, Portugal

Abstract. Precancerous conditions such as intestinal metaplasia (IM) have a key role in gastric cancer development and can be detected during endoscopy. During upper gastrointestinal endoscopy (UGIE), misdiagnosis can occur due to technical and human factors or by the nature of the lesions, leading to a wrong diagnosis which can result in no surveillance/treatment and impairing the prevention of gastric cancer. Deep learning systems show great potential in detecting precancerous gastric conditions and lesions by using endoscopic images and thus improving and aiding physicians in this task, resulting in higher detection rates and fewer operation errors. This study aims to develop deep learning algorithms capable of detecting IM in UGIE images with a focus on model explainability and interpretability. In this work, white light and narrow-band imaging UGIE images collected in the Portuguese Institute of Oncology of Porto were used to train deep learning models for IM classification. Standard models such as ResNet50, VGG16 and InceptionV3 were compared to more recent algorithms that rely on attention mechanisms, namely the Vision Transformer (ViT), trained in 818 UGIE images (409 normal and 409 IM). All the models were trained using a 5-fold cross-validation technique and for validation, an external dataset will be tested with 100 UGIE images (50 normal and 50 IM). In the end, explainability methods (Grad-CAM and attention rollout) were used for more clear and more interpretable results. The model which performed better was ResNet50 with a sensitivity of 0.75 ($\pm$0.05), an accuracy of 0.79 ($\pm$0.01), and a specificity of 0.82 ($\pm$0.04). This model obtained an AUC of 0.83 ($\pm$0.01), where the standard deviation was 0.01, which means that all iterations of the 5-fold cross-validation have a more significant agreement in classifying the samples than the other models. The ViT model showed promising performance, reaching similar results compared to the remaining models.

A. Cunha et al. (Eds.): MobiHealth 2022, LNICST 484, pp. 34–50, 2023.
https://doi.org/10.1007/978-3-031-32029-3_4

Keywords: Deep Learning · Gastrointestinal Metaplasia · Computer Vision · Gastrointestinal Endoscopy

1 Introduction

1.1 Motivation

Upper gastrointestinal endoscopy (UGIE) is a medical procedure that is used to diagnose and treat multiple pathologies. UGIE consists of the introduction of an endoscope through the mouth, allowing the observation of the mucosa of the esophagus, stomach and duodenum, performing endoscopic diagnosis using visual characteristics that are complemented, when necessary, with biopsy and histopathological analysis. Regarding the stomach, endoscopy is used to detect premalignant conditions and to treat pre-malignant lesions and early neoplastic lesions. The pathogenesis of gastric cancer (GC) involves a series of events beginning with *Helicobacter pylori*-induced chronic inflammation (*H. pylori*), progressing to atrophic gastritis (AG), intestinal metaplasia (IM), dysplasia, and eventually GC [1–3]. In Fig. 1 are some examples of these lesions.

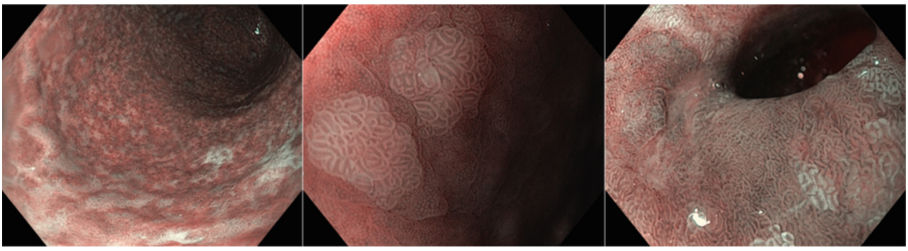


Fig. 1. Precancerous lesions from left to right from our private dataset: AG, IM and dysplasia.

Gastritis consists of inflammation of the stomach mucosa, which can be chronic, most of the time in the setting of *H. pylori* infection. Persistent inflammatory changes can result in gastric atrophy (loss of gastric glands and consequent reduction in the thickness of the gastric mucosa) and in intestinal metaplasia (substitution of gastric glands by intestinal-type glands formed by caliciform mucin-containing cells, paneth cells and absorptive cells). These conditions increase the risk of gastric cancer. Factors such as age, race/ethnicity, family history, environmental factors, and *H. pylori* strain are additional factors that can modify the risk of gastric cancer [3].

Endoscopy is important to detect these precancerous conditions (AG and IM) since patients with extensive precancerous conditions indicate periodical endoscopic surveillance [4]. One of the aims of endoscopic surveillance is the detection of these types of lesions, which can lead to early gastric cancer (EGC). Thus, the detection of AG and IM is a key role to increase the chances of survival since advanced GC has a poor long-term prognosis. However, training for the detection of precancerous conditions is suboptimal and these may be difficult to detect for the majority of endoscopists, even if virtual chromoendoscopy (a technique that uses light filters to increase the contrast of the image) is

used [4]. Moreover, AG and IM show subtle alterations in the mucosa, which requires careful observation during the examination. Due to the nature of these lesions and lack of attention and tiredness, mistakes could be made by endoscopists, leading to missed diagnostics [5, 6].

1.2 State-of-the-Art

Deep learning (DL) in medical image analysis has provided comparable results to humans in multiple classification tasks. This machine learning paradigm, which is based on architectures made of multiple layers, is not intended to replace medical professionals in the diagnostic process but can be integrated into workflows to extract important information from images by recognizing patterns implicit to humans [7]. These DL systems are capable of real-time UGIE lesions detection in endoscopic videos, which can be helpful for endoscopists, pointing to suspicious areas during the examination. It can also be helpful for the endoscopist's learning process. It would allow less experienced endoscopists to learn visual patterns associated with GI lesions [8, 9]. Follows some examples of studies using DL for EGC lesion detection, such as IM and GA, in UGIE images.

The collected studies use well-known convolutional neural networks such as Xception, EfficientNETB4, ResNets, VGGs and DenseNets either to detect AG or IM or both. Of particular relevance, Hongyan Li and colleagues [10] proposed a multi-fusion method for metaplasia detection. First, three feature modules (ResNet50) are used for feature extraction in RGC, HSV and LBP images. The three types of features are fused into one representation based on attention mechanisms throughout an attention fusion module (AFM). The AFM is optimized by a regularization module which combines L1 regularization and label smoothing. Using this module can alleviate the overfitting more effectively and the model overconfidence caused by one hot encoding. Another example is the study of Tao Yan et al. [11] which a system was developed where the first model selects magnifying narrow-band imaging (M-NBI) and narrow-band imaging (NBI) images and then detects the presence of IM using two classifiers, one trained with M-NBI images and another trained with NBI images.

No specific image preprocessing methods are explicit in the described works [10, 12, 13]. Only in [11], after the images were classified as M-NBI or NBI, the black frames were cropped and the resulting image was resized to 224×224 pixels to then detect the IM presence.

To increase the interpretability of model predictions, explainable artificial intelligence (XAI) methods were applied in [11] by using Grad-CAMs for highlighting fine-grained features and in [12] by using class activation maps to indicate the features of the lesion that the CNNs focused on.

Transfer learning methods are usually applied in a variety of classification methods for a good initialization of weights and enhanced robustness against overfitting. All the mentioned works [10–13] used CNN architectures pre-trained using images from the ImageNet datasets.

All the considered works used private UGIE images to train and validate the models and performed 5-fold cross-validation methods, with exception of [13]. Tao Yan [11] used 1880 endoscopic images (1048 IM and 832 non-IM) to train the models. And 477

pathologically images (242 IM and 235 non-IM) to evaluate the models. Ming Xu [13] to train and evaluate these models collected 3 datasets composed of M-NBI and blue laser imaging (BLI) images: 1) collected from two institutions, split into a train (2439 AG and 2017 IM), validation and (internal) test (610 AG and 530 IM) set; 2) collected from 3 institutions, as an external test set (708 AG and 590 IM); 3) a prospective single-centre external test dataset containing 71 precancerous and 36 control images served as a benchmarking for the comparison of the algorithm with trained endoscopists. In the work of Ne Lin [12] 5,883 images were assigned (2,713 AG and 2,912 IM) for training, 606 images (199 AG and 201 IM) for validation and 548 images (180 AG and 188 IM) for testing. Hongyan Li [10] used 1050 UGIE NBI images (585 IM and 465 non-IM) to train and evaluate the models.

For the exclusion criteria, the images affected by artefacts created by mucus, poor focus and motion-blurring were excluded from the datasets used for training and evaluation [10–13]. Depending on the aim of the work, the authors used different types of image modalities. For example, Tao Yan [11] only used M-NBI and NBI images, Ming Xu [13] used M-NBI and BLI images, Ne Lin [12] only used WLI images and Hongyan Lin [10] only used NBI images.

In order to increase the amount of data, all the mentioned works used augmentation processes such as random rotations, flips, shifts and zooms, generating augmented data.

In Tao Yan work [11] it was possible to verify that M-NBI images help the model to achieve a better performance when compared to NBI images, for IM classification. The NBI and M-NBI classifiers achieved a sensitivity of 91%, a specificity of 71%, an accuracy of 83% and a sensitivity of 94%, a specificity of 84%, an accuracy of 89%, respectively.

In Ming Xu study [13], using the VGG16 model, the AG classification achieved 90% of accuracy, 90% of sensitivity and 91% of specificity in the internal test set, 86% of accuracy, 90% of sensitivity and 80% of specificity in the external test set and 88% of accuracy, 97% of sensitivity and 73% of specificity in the prospective video test set. For IM classification achieved 91% of accuracy, 89% of sensitivity and 93% of specificity in the internal test set, 86% of accuracy, 97% of sensitivity and 72% of specificity in the external test set and 90% of accuracy, 95% of sensitivity and 84% of specificity in the prospective video test set. The use of several datasets from different hospitals provides reliability in the robustness of the model since the select methodologies present good results when evaluated in unbiased data. Comparably, Ne Lin [12] used a TResNet achieving an area under the ROC curve (AUC) of 98%, the sensitivity of 96%, specificity of 96% and accuracy of 96% for AG classification and an AUC of 99%, the sensitivity of 98%, specificity of 98% and accuracy of 98% for IM classification. When comparing the results of Ming Xu [13] using M-NBI and BLI images and Ne Lin's [12] results using WLI, it is possible to conclude that a better performance can be achieved by using WLI images instead of using M-NBI and BLI. The TResNet reached better results when compared to the VGG16 using less data, however, in [13] was used more data from different hospitals, providing more robustness to the model for better generalization.

The proposed methodology made by Hongyan Li [10] for IM classification achieved 90% of accuracy, 90% of precision, 93% of recall and 92% of F1-Score.

Table 1 described the most important characteristics of the described works.

Table 1. Summary of the collected state-of-the-art studies.

Author	Year	Lesion	Datasets	Image Modality	Model	XAI	Results
Hongyan Li et al. [10]	2021	IM	(Private Dataset) Train and evaluation: 1050 UGIE images (585 IM and 465 non-IM)	NBI	ResNet50	-	Accuracy: 90% Precision: 90% Recall: 93% F1-Score: 92%
Tao Yan et al. [11]	2020	IM	(Private Dataset) Train: 1880 UGIE images (1048 IM and 832 non-IM) Evaluation: 477 UGIEimages (242 IM and 235 non-IM)	M-NBI and NBI	Xception, NASNet and EfficientNetB4	Grad-CAM	IM in NBI Sensitivity: 91%Specificity: 71% Accuracy: 83% IM in M-NBI Sensitivity: 94% Specificity: 84% Accuracy: 89%,
Ne Lin et al. [12]	2021	AG and IM	(Private Dataset) Train: 5,883 images (2,713 AG and 2,912 IM) Validation: 606 images (199 AG and 201 IM) Test: 548 images (180 AG and 188 IM)	WLI	TResNet	Class Activation Maps	AG AUC: 98% Sensitivity: 96% Specificity: 96% Accuracy: 96% IM AUC: 99% Sensitivity: 98% Specificity: 98% Accuracy: 98%
Ming Xu et al. [13]	2021	AG and IM	(Private Dataset) 2149 GA images 3049 IM images	M-NBI and BLI	ResNet50, VGG16, DenseNet169 and EfficientNetB4	-	AG Accuracy: 90% Sensitivity: 90% Specificity: 91% IM Accuracy: 91% Sensitivity: 89% Specificity: 93%

1.3 Contributions

This work will compare the performance of DL models which achieved better results, based on the collected studies, to more recent approaches that rely on self-attention mechanisms, for IM classification. In the end, explainability methods for the models' predictions will be applied to verify if correlated features of the disease are highlighted.

For this, two research questions were formulated:

- **RQ1** - Can novel Vision Transformers (ViT) DL architectures outperform current DL architectures in the task of metaplasia detection in UGIE images?
- **RQ2** - Can explainable techniques highlight regions with clinical relevance and correlate to the metaplasia presence?

By answering these questions, the proposed study will contribute to understanding if these methods, which rely on attention mechanisms, can be applied to IM detection and reach similar performance or outperform the classic CNN architectures. After training and testing the models, in the end, explainability methods will be applied to understand if the features identified by the model have clinical relevance correlated to the output prediction.

This paper is organized into 5 sections, starting with this introduction, in Sect. 1, composed of the motivation, state-of-the-art and contributions. In Sect. 2, the DL architectures applied in this work will be described. In Sect. 3, the methodology used for the proposed study is explained in detail. In Sect. 4 the results will be described and in

Sect. 5 these results will be discussed. Finally, Sect. 6 will point out the conclusions of this work and future perspectives regarding the classification of IM using DL techniques.

2 Deep Learning Models

To select the methods to be used for this work, it was necessary to study in detail some architectures to achieve the initial objectives. Thus, DL models to be used for IM detection in endoscopy images will be analysed in detail, and then use XAI techniques in order to give interpretability and explainability to understand the predictions of the model.

ResNet50 is considered a deep network since it is 50 layers deep (Fig. 2). A residual layer in this stacked layers architecture in the residual block always contains 1×1, 3×3 and 1×1 convolution layers. The first 1×1 convolution will reduce the dimension and then the features will be calculated in the 3×3 bottleneck layer. In the next 1×1 layer, the dimension is increased again. The 1×1 filter is used in this architecture to reduce and increase the dimension of the feature maps before and after the bottleneck layer. Since there is no pooling layer within the residual block, the dimension is reduced in 1×1 convolutions with strides of 2 [14].

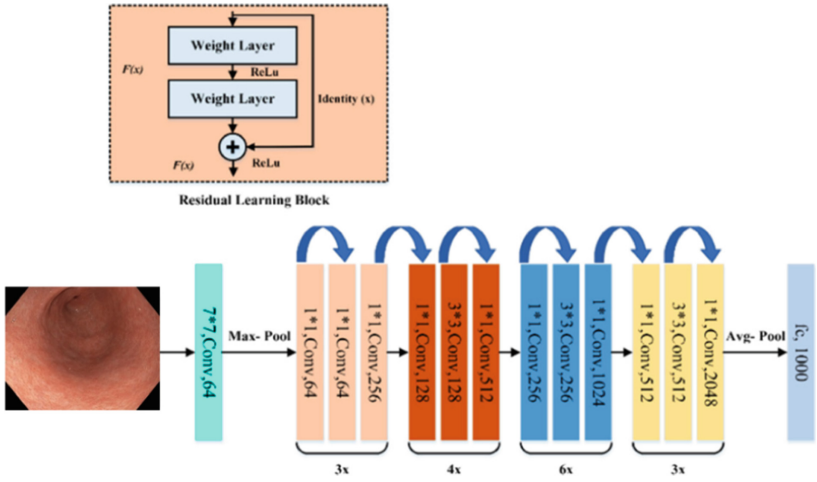


Fig. 2. Block diagram of the ResNet50 architecture, adapted from [15].

Visual Geometry Group (VGG), also known as VGGNet, was created to increase the depth of Convolutional Neural Networks (CNN), with the main goal of increasing model performance. Due to the number of convolutional layers, the VGG architecture is referred to as a deep architecture, i.e., VGG16 means that it has 16 layers (Fig. 3). This architecture is developed with very small convolutional filters. This has then 13 convolutional layers divided into five groups, and a max-pooling layer follows each group and 3 fully connected layers [15]. VGG16 has a 224×224 image as input. The convolution steps are fixed to keep the spatial resolution preserved after resolution.

When it comes to the hidden layers these use ReLu. As mentioned here VGG uses 3 fully connected layers, two of these layers with a size of 4096 neurons and the last one with a size of 1000 neurons [16].

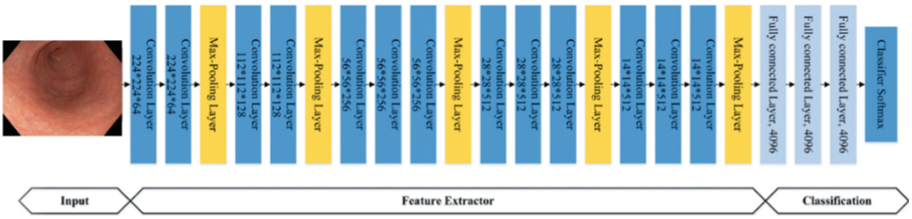


Fig. 3. Block diagram of the VGG16 architecture, adapted from [15].

The Inception V3 model involves more than 20 million parameters. This model includes symmetric and asymmetric building blocks, where each block is composed of various convolutional, average, and max pooling, concats, dropouts, and fully connected layers (Fig. 4). Batch normalization is usually used and applied in the activation layer in this model. This architecture has 42 layers of depth [17].

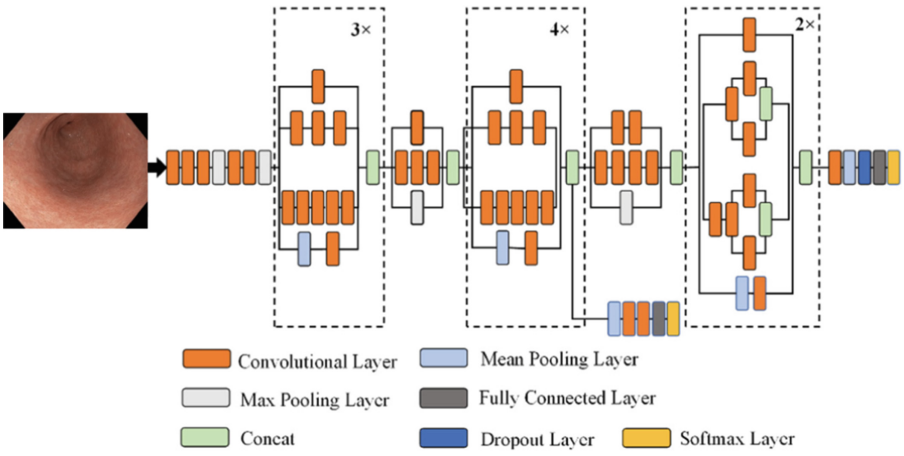


Fig. 4. Block diagram of the InceptionV3 architecture, adapted from [15].

Visual Transformers (ViT) have recently emerged as an alternative to CNN (Fig. 5). Compared to CNNs, it is generally found that the weaker inductive bias of the ViT leads to a greater reliance on model regularization or data augmentation when training on smaller training datasets. Internally, the transformer learns by measuring the relationship between pairs of input tokens. In computer vision, we can use image patches as a token. This relationship can be learned by providing attention to the network. This can be done in conjunction with a convolutional network or by replacing some components of convolutional networks. These network structures can be applied to image classification tasks [18].

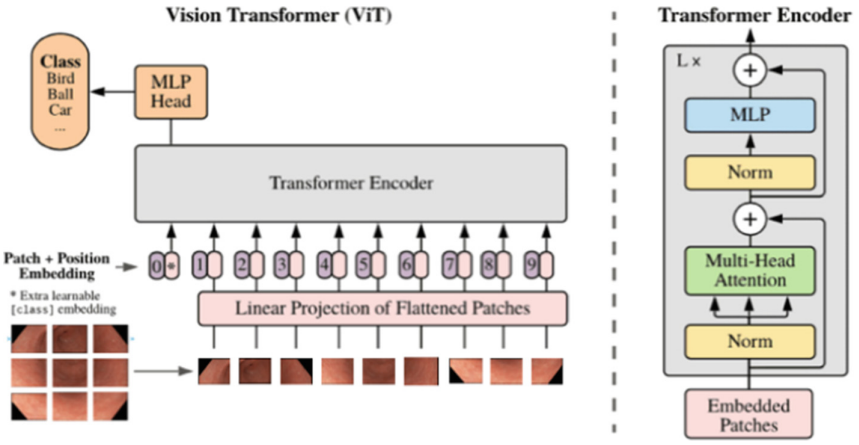


Fig. 5. Overview of ViT model, adapted from [18].

To better understand the model decisions, some XAI techniques were created to provide transparency, explainability and interpretability, helping non-experts to understand which features/attributes are responsible for the model's prediction.

Grad-CAM is a generalization of CAM, since it produces visual explanations for the CNN regardless of the architecture, thus overcoming one of the limitations of CAM. This method is a gradient-base, which uses the class-specific gradient information flowing into the last convolutional layer of a CNN, to design a coarse location map of the important regions in the image, in cases dealing with classification, thus making CNN-based models more transparent. This technique learns the information about the importance of the neurons in the decision process, using the gradient going to the last convolutional layer [19].

For ViT models Abnar and Zuidema [20] introduced the Attention Rollout mechanism which quantifies how the information flows through self-attention layers of transformer blocks. Is an average of the attention weights across all heads and then recursively multiplied by the weight matrices of all layers.

3 Materials and Methods

The workflow of this paper is illustrated in Fig. 6 which will be further explained in this section. Summarily, from the Portuguese Institute of Oncology (IPO) of Porto, a dataset with IM images and healthy gastric tissue was used. All the images suffered preprocessing before being split into the train, test, and validation sets to train and evaluate the different models.

3.1 IPO Dataset for IM Detection

The IPO Post-Map dataset is a private dataset of UGIE images obtained from a total of 170 different exams. The exams were performed in different health institutions: 20

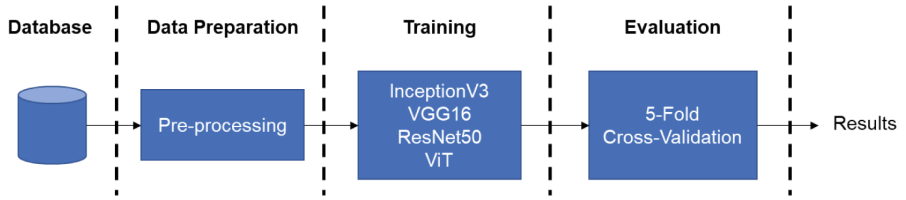


Fig. 6. Pipeline of the developed work.

exams from the IPO of Porto (Portugal), 28 exams from the University of Medicine and Pharmacy TG., Mures (Romania), 14 exams from Queen's Medical Centre, Nottingham (UK), and 75 exams from Hospital Sant 'Andrea, University Sapienza Roma (Italy). The dataset tries to cover various pathologies, allowing the diagnosis of various precancerous lesions and was created using images of three stomach locations (corpus, incisura, and antrum). Samples from patients without indication for biopsy, with significant comorbidities, anticoagulant therapy or coagulation disorders, previous gastric neoplasia or surgery and not being able to perform at least three biopsies during the endoscopy were excluded [21]. It contains a total number of 1355 images, Narrow-Band Imaging (NBI) and White Light Imaging (WLI), of which 499 are high resolution and 856 are low resolution, ranging between 1350 × 1080 to 514 × 476 resolution. It has four classes (AG, dysplasia, IM, and healthy tissue).

3.2 Data Preparation

The images used were resized for the respective input size of each model, for ResNet50, VGG16, and ViT the images were resized to 224 × 224, and for InceptionV3 the images were resized to 229 × 229. Aiming to classify images into one of two classes, only WLI and NBI images of IM and healthy stomach tissue, a total of 818 images, 409 of each class, were selected from the IPO Post-Map dataset. The UGIE images were annotated, by two interns of the IPO of Porto, image-wise using a design annotation tool for this purpose. For the image-based annotation, the UGIE images were classified as normal or IM. Some examples of images from the IPO Post-Map are illustrated in Fig. 7.

Due to the small amount of data a 5-fold cross-validation (CV) was performed to train and evaluate the models. An external dataset, in the end, was used to test the robustness of each model trained in each fold. Figure 8 is present the dataset split used to create and evaluate the four models.

Each network has a different preprocessing method. In the case of ResNet50 and VGG16, the images are converted from RGB to BGR, and then each colour channel is zero-centred concerning the ImageNet dataset, without scaling. For inceptionV3 and ViT the input pixel values are scaled between -1 and 1, sample-wise.

3.3 Training

To execute the binary classification task described previously, multiple state-of-the-art models were tested. VGG16, InceptionV3 and ResNet50 were selected based on their

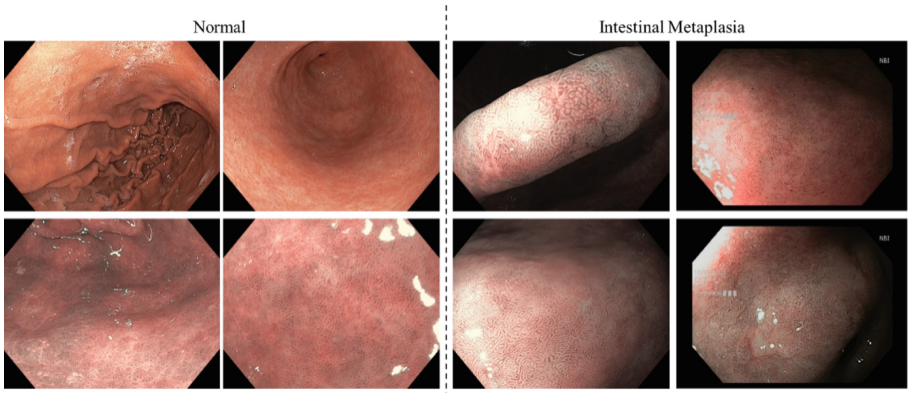
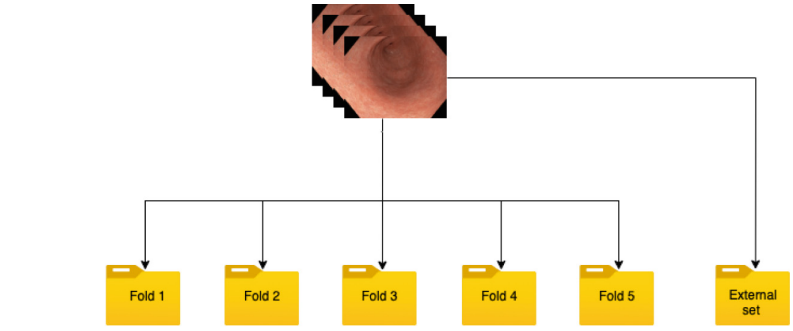


Fig. 7. Examples of WLI and NBI UGIE images of the Post-Map dataset, from normal or intestinal metaplasia samples.



	Train	Validation	Train	Validation	Train	Validation	Train	Validation	Train	Validation	Test
Metaplasia	292	67	299	60	278	81	285	74	282	77	50
Normal	282	77	275	80	296	63	290	69	293	66	50
Total	574	144	574	140	574	144	575	143	575	143	100

Fig. 8. Data organization for the training of the models with 5-fold cross-validation with an external set.

performance in classification tasks related to the aim of this study. These models were compared to a more recent approach that relies on self-attention mechanisms, called Vision Transformers (ViT), which recently emerged as an alternative to CNNs. For this study the Keras API was used integrated into the Tensorflow framework running the different experiments in a Nvidia RTX 3060ti GPU. All the models were trained during 100 epochs, with a dropout layer of 0.3, a learning rate of $1e^{-4}$, Adam optimizer, binary focal loss as loss function and batch size of 16 with exception of ViT model which had a batch size of 8.

3.4 Evaluation

Considering the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) produced by each one of the models in the test set, the following metrics were calculated aiming to evaluate their performance:

The accuracy (1) is the ratio between the number of correct predictions and the total number of input samples.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

Precision (2) is the ratio of correct predictions to the number of positive results predicted by the classifier.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Specificity (3) measures the proportion of the negative cases that were correctly classified. Recall or sensitivity (4) is the number of correctly predicted results divided by the number of all those that should have been classified as positive.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

F1-Score (5) outputs a value between zero and one and tries to find the balance between precision and recall, letting know how accurate the model is and how many samples it correctly classifies. The F1-Score is the harmonic mean of these two.

$$\text{F1-Score} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FN} + \text{FP}} \tag{5}$$

In addition to these measures, other measures were also used, such as rank-based performance measures. These measures rank predictions relative to the probability of an outcome. For this work, measures such as the Receptor Operating Characteristic (ROC) Curve were used.

The ROC curve is a performance measure for classification problems at various threshold parameters. ROC is a probability curve and AUC represents the degree or measure of separability. Essentially this is how well the model can distinguish between classes. It can be concluded that the higher the AUC the better the model distinguishes between patients with and without the disease. The ROC curve is plotted with TPR (sensitivity) against FPR (False Positive Rate).

4 Results

Throughout this project, several experiments were carried out to evaluate the different state-of-the-art architectures in the binary classification task.

The results will be discussed and compared with the study described in Sect. 1. In the end, will be performed a brief comparison of the standard methodologies and the ViT methodology.

Table 2 shows the average results obtained in each fold of the 5-fold CV validation set.

Overall, the model with the best performance was the InceptionV3 with 0.82 (±0.03) of accuracy, 0.82 (±0.07) of sensitivity, 0.80 (±0.11) of specificity, 0.82 (±0.04) of precision, 0.82 (±0.04) of F1-Score and 0.84 (±0.05) of AUC. The ResNet50 achieved very similar results, with slight changes, and all the models reached an AUC over 0.80.

The ViT model reached similar results when compared to the VGG16, with 0.77 (±0.03) of accuracy, 0.76 (±0.07) of sensitivity, 0.80 (±0.06) of specificity, 0.79 (±0.08) of precision, 0.77 (±0.03) of F1-Score and 0.82 (±0.03) of AUC.

Table 2. Results for the validation set in 5-fold CV. The higher results are highlighted in bold.

Metrics	ResNet50	VGG16	InceptionV3	ViT
Accuracy	0.80 (±0.02)	0.75 (±0.03)	**0.82 (±0.03)**	0.77 (±0.03)
Sensitivity	0.77 (±0.04)	0.75 (±0.10)	**0.82 (±0.07)**	0.76 (±0.07)
Specificity	**0.83 (±0.04)**	0.74 (±0.15)	0.80 (±0.11)	0.80 (±0.06)
Precision	**0.82 (±0.07)**	0.76 (±0.06)	**0.82 (±0.04)**	0.79 (±0.08)
F1-Score	0.79 (±0.03)	0.75 (±0.05)	**0.82 (±0.04)**	0.77 (±0.03)
AUC	**0.84 (±0.02)**	0.80 (±0.05)	**0.84 (±0.05)**	0.82 (±0.03)

Finally, to evaluate the robustness of these models was presented an external test set to further test the performance of these models in data never used for validation or training (Table 3). The ResNet50 model reached the best results, with 0.79 (±0.01) of accuracy, 0.75 (±0.05) of sensitivity, 0.82 (±0.04) of specificity, 0.81 (±0.03) of precision, 0.77 (±0.02) of F1-Score and 0.83 (±0.01) of AUC, followed by the InceptionV3, once again, with a very similar performance. The VGG16 and the ViT model achieved comparable performances between them, with an AUC of 0.79 when tested with the external dataset.

Table 3. Results for the external set in 5-fold CV. The higher results are highlighted in bold.

Metrics	ResNet50	VGG16	InceptionV3	ViT
Accuracy	**0.79 (±0.01)**	0.76 (±0.02)	0.77 (±0.01)	0.75 (±0.02)
Sensitivity	0.75 (±0.05)	0.76 (±0.04)	**0.78 (±0.04)**	0.66 (±0.04)
Specificity	0.82 (±0.04)	0.76 (±0.07)	0.76 (±0.04)	**0.85 (±0.04)**
Precision	**0.81 (±0.03)**	0.76 (±0.05)	0.77 (±0.02)	**0.81 (±0.03)**
F1-Score	**0.77 (±0.02)**	0.76 (±0.01)	**0.77 (±0.01)**	0.73 (±0.02)
AUC	**0.83 (±0.01)**	0.79 (±0.02)	0.81 (±0.01)	0.79 (±0.03)

5 Discussion

Overall the results are very similar between all the models. Generally, for the validation set, InceptionV3 achieved the best values in almost all used metrics with exception of the specificity. Either way, the results are very similar when compared to the ResNet50 model and both reached an AUC of 0.84 with a small standard deviation which proves the agreement in the different folds. Although the models did not reach comparable results with the studies presented in Sect. 1 mostly due to the huge gap in data that exists between this work and the described state-of-the-art works.

The ViT achieved comparable performance, especially when compared to the ResNet50 model, reaching an AUC of 0.82 and proving to have a better performance than the VGG16. In overall, the models show consistency in predicting either normal or IM images.

In Fig. 9 are illustrated the ROC curves with the respective AUC values of each model, with the mean ROC for the 5-fold for each model and, in grey, the standard deviation between all folds of each model.

Analysing the ROC curves can be seen that the ResNet50 and InceptionV3 curves are very similar but the ResNet50 presents less grey area, consequently a smaller standard deviation which represents a bigger agreement between the different folds of this model.

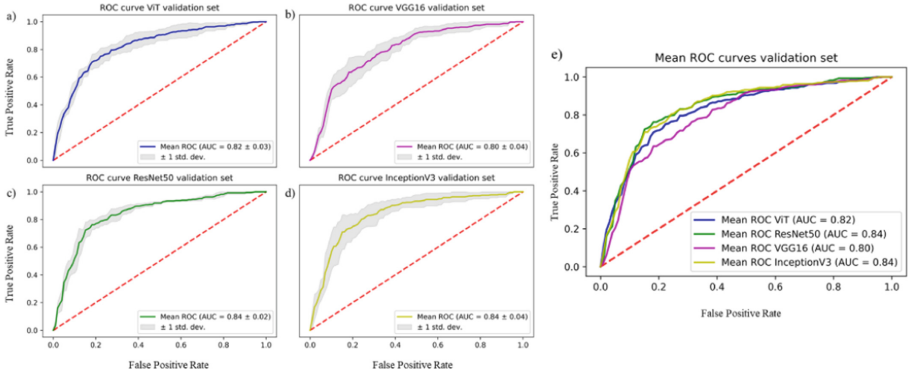


Fig. 9. ROC curves of the validation set of the models: (a) ViT, (b) VGG16, (c) ResNet50, (d) InceptionV3 and (e) is the mean of all ROC curves.

For the external test set, ResNet50 and InceptioV3 achieved overall better results, although this time, the ResNet50 has slightly better performance. As said before, the ResNet50 shows a bigger agreement between all the folds, reaching the best values for all the metrics except for the sensitivity and specificity.

The ViT proved one more time to be a reliable model with similar results to the VGG16. In Fig. 10 the ROC curves for the external test set are represented. The ROC curves follow the same pattern as seen in Fig. 9, where the InceptionV3 and ResNet50 achieved the best AUC values and have the best ROC curves shape. The ResNet50 has the smallest standard deviation, proving again the most agreement between the different

folds, now to an external test set, reinforcing the best generalization and robustness of all models.

When comparing the state-of-the-art results it is possible to see that although the performance of these models is good, they did not achieve the results presented in the literature review, mostly due to the lack of data.

As far as transformers are concerned, so far no other studies were found using ViT models in the IM classification of UGIE images. The ViT model in this presented study proves to achieve similar results compared to the remaining models, even reaching a better performance than the VGG16 model. These self-attention models rely even more upon great amounts of data when compared to traditional CNN architectures, which could be the problem for not reaching even better results.

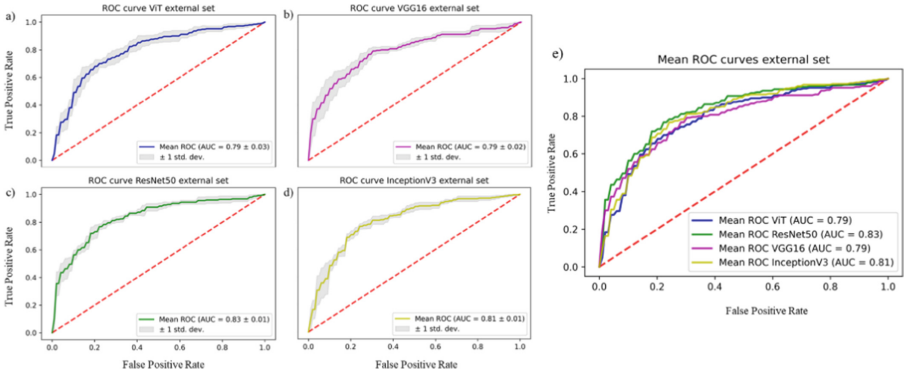


Fig. 10. ROC curves of the external set of the models: (a) ViT, (b) VGG16, (c) ResNet50, (d) InceptionV3 and (e) is the mean of all ROC curves.

CNNs can be extremely difficult for non-experts to explain and understand, so was decided to apply XAI methods, which help to create transparency, interpretability, and explainability as a basis for the output of the neural networks.

Figure 11 presents the activation maps for the different models, using Grad-CAMs for traditional CNNs and Attention Rollout for the ViT model.

The Grad-CAMs attention maps point to important areas that lead the model for predicting IM presence in the images, by using the gradient of the model of the final convolutional layer. The attention rollout uses the weights average across the different attention blocks to highlight the most important regions for the ViT decision.

In most cases, the highlighted areas are common for the different models and reveal IM-related characteristics. For example, the different models focus on flat and patchy regions with irregularities and in NBI images bluish-white appearance of the gastric mucosa is highlighted as well, which is a feature of IM presence. In other cases, the ViT model highlight more accurate regions, more correlated and indicative of IM presence. For instance, in Fig. 11 mid-row images, the ViT model points to more important regions with more clinical indications of IM presence, when compared to the regions highlighted by the CNN models, with no presence of IM characteristics.

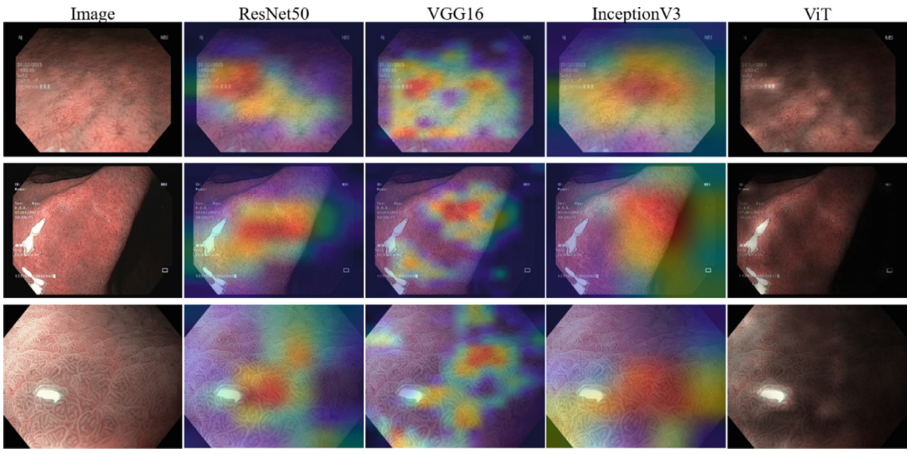


Fig. 11. Grad-CAM and attention rollout of the 5-Fold CV for the CNN and ViT model, respectively. All the UGIE images are IM examples.

This type of explainability is extremely important for trustworthy models so that the physicians can understand why is made certain predictions from the models and which features are related to that decision. More trustworthy models will understand which features are more indicative of IM presence, thus highlighting regions with these characteristics.

Regarding RQ1, the ViT models reached similar performance when compared to more traditional CNNs, however, did not outperform ResNet50 and InceptionV3 architectures. More adaptations can be applied for ViT performance enhancement, such as shifted patch tokenization or/and locality self-attention to tackle the problem of locality inductive bias present in ViTs. Nonetheless, regarding RQ2, the ViT activation maps highlight IM characteristic with more clinical relevance. Thus, despite not outperforming the ResNet50 and InceptionV3 models, the ViT architecture has promising applications in IM detection due to the relation process at the pixel level, understanding different features which are not identified by traditional CNNs.

6 Conclusions

This paper focuses on the evaluation of multiple state-of-the-art architectures in image classification in IM and normal gastric mucosa. The results achieved in the multiple experiments were satisfactory compared to the described works in Sect. 1, although not reaching the same performance mostly due to the gap of data between works. Additionally, self-attention models such as ViT were tested and reached comparable results to the standard CNNs. This type of architecture which relies on transformer blocks needs even more data than a CNN. However, with the data available, it was possible to achieve satisfactory results similar to the performance of the CNNs. For all the models activation maps were computed to give interpretability and explainability of the results.

This is a required step nowadays when DL models are applied especially to the healthcare field in order to justify the algorithmic results to non-experts and highlight which features/attributes are strongly related to the model decision.

In future works, increasing the data will be essential and this can be done using generative adversarial methods, creating synthetic UGIE images. Adaptation in ViT models can enhance the performance of IM detection, especially in small datasets, by using shifted patch tokenization or/and locality self-attention. Further explainability methods should be explored to provide even more transparency to DL models, such as local interpretable model-agnostic explanations which are more suitable for local explanations.

Acknowledgements. This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project PTDC/EEI-EEE/5557/2020.

References

1. ASGE Standards of Practice Committee, et al.: Appropriate use of GI endoscopy. Gastrointest. Endosc. **75**(6), 1127–1131 (2012). https://doi.org/10.1016/j.gie.2012.01.011
2. Evans, J.A., et al.: The role of endoscopy in the management of premalignant and malignant conditions of the stomach. Gastrointest. Endosc. **82**(1), 1–8 (2015). https://doi.org/10.1016/j.gie.2015.03.1967
3. Peixoto, A., Silva, M., Pereira, P., Macedo, G.: Biopsies in gastrointestinal endoscopy: when and how. GE Port. J. Gastroenterol. **23**(1), 19–27 (2016). https://doi.org/10.1016/j.jpge.2015.07.004
4. Pimentel-Nunes, P., et al.: Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. Endoscopy **51**(04), 365–388 (2019). https://doi.org/10.1055/a-0859-1883
5. Sitarz, R., Skierucha, M., Mielko, J., Offerhaus, G.J.A., Maciejewski, R., Polkowski, W.P.: Gastric cancer: epidemiology, prevention, classification, and treatment. Cancer Manag. Res. **10**, 239–248 (2018). https://doi.org/10.2147/CMAR.S149619
6. Moon, H.S.: Improving the endoscopic detection rate in patients with early gastric cancer. Clin. Endosc. **48**(4), 291 (2015). https://doi.org/10.5946/ce.2015.48.4.291
7. e Gonçalves, W.G., Dos Santos, M.H.D.P., Lobato, F.M.F., Ribeiro-dos-Santos, Â., de Araújo, G.S.: Deep learning in gastric tissue diseases: a systematic review. BMJ Open Gastroenterol. **7**(1), e000371 (2020). https://doi.org/10.1136/bmjgast-2019-000371
8. Renna, F., et al.: Artificial intelligence for upper gastrointestinal endoscopy: a roadmap from technology development to clinical practice. Diagnostics **12**(5), 1278 (2022). https://doi.org/10.3390/diagnostics12051278
9. Arribas, J., et al.: Standalone performance of artificial intelligence for upper GI neoplasia: a meta-analysis. Gut **70**(8), 1458–1468 (2021). https://doi.org/10.1136/gutjnl-2020-321922
10. Li, H., et al.: A multi-feature fusion method for image recognition of gastrointestinal metaplasia (GIM). Biomed. Signal Process. Control **69**, 102909 (2021). https://doi.org/10.1016/j.bspc.2021.102909
11. Yan, T., Wong, P.K., Choi, I.C., Vong, C.M., Yu, H.H.: Intelligent diagnosis of gastric intestinal metaplasia based on convolutional neural network and limited number of endoscopic images. Comput. Biol. Med. **126**, 104026 (2020). https://doi.org/10.1016/j.compbiomed.2020.104026

12. Lin, N., et al.: Simultaneous recognition of atrophic gastritis and intestinal metaplasia on white light endoscopic images based on convolutional neural networks: a multicenter study. Clin. Transl. Gastroenterol. **12**(8), e00385 (2021). https://doi.org/10.14309/ctg.0000000000000385
13. Xu, M., et al.: Artificial intelligence in the diagnosis of gastric precancerous conditions by image-enhanced endoscopy: a multicenter, diagnostic study (with video). Gastrointest. Endosc. **94**(3), 540–548 (2021). https://doi.org/10.1016/j.gie.2021.03.013
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv: arXiv:1512.03385 (2015). http://arxiv.org/abs/1512.03385. Accessed 05 Jun 2022
15. Ali, L., Alnajjar, F., Jassmi, H.A., Gocho, M., Khan, W., Serhani, M.A.: Performance evaluation of deep CNN-based crack detection and localization techniques for concrete structures. Sensors **21**(5), 1688 (2021). https://doi.org/10.3390/s21051688
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition, p. 14 (2015)
17. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826. Las Vegas, NV, USA (2016). https://doi.org/10.1109/CVPR.2016.308
18. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv: arXiv:2010.11929 (2021). http://arxiv.org/abs/2010.11929. Accessed 02 Jun 2022
19. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: a review of machine learning interpretability methods. Entropy **23**(1), 1–45 (2021). https://doi.org/10.3390/e23010018
20. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. arXiv (2020). http://arxiv.org/abs/2005.00928. Accessed 18 Jul 2022
21. Pimentel-Nunes, P., et al.: A multicenter prospective study of the real-time use of narrow-band imaging in the diagnosis of premalignant gastric conditions and lesions. Endoscopy **48**(08), 723–730 (2016). https://doi.org/10.1055/s-0042-108435