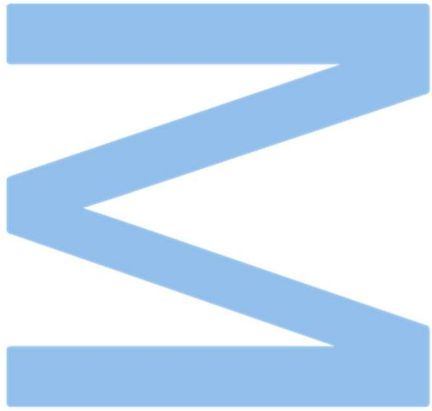
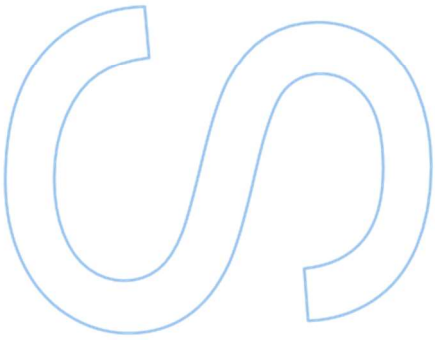
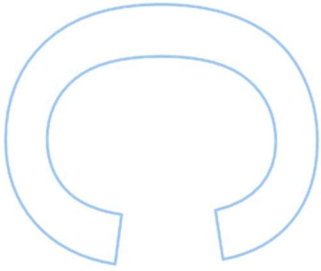
Novel deep learning methods for characterization of precancerous tissue in endoscopic narrow band images





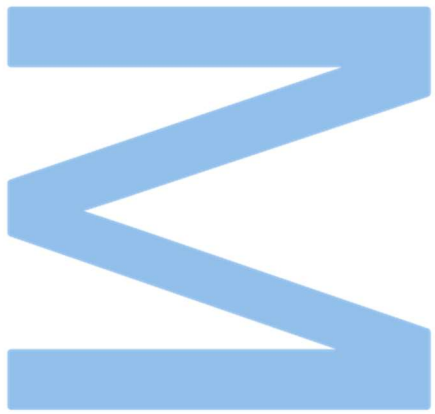


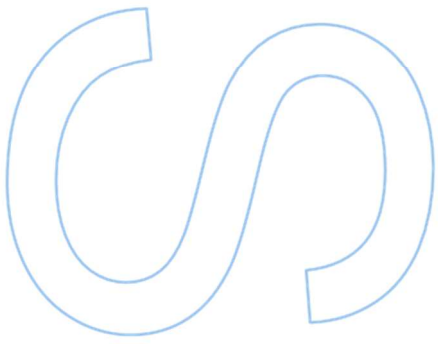
Maria Pedroso da Silva
Master's Degree on Data Science
Department of Computer Science
2023

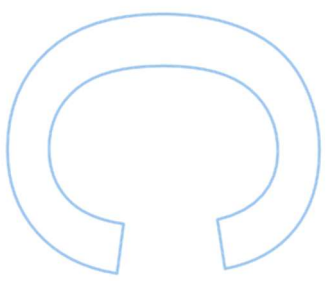
Supervisor
Francesco Renna, Assistant Professor, Faculty of Sciences

Co-supervisor
Miguel Coimbra, Full Professor, Faculty of Sciences


U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO







Abstract

Gastric Cancer (GC) is one of the deadliest forms of cancer worldwide. However, early diagnoses can significantly improve patient survival by enabling frequent monitoring. Gastric Intestinal Metaplasia (GIM) is a precancerous gastric lesion that is related to a high risk of progressing to advanced gastric cancer. Detecting this condition is very challenging due to its characteristics leading to low inter-observer concordance and a high rate of missed diagnoses. Deep Neural Networks (DNN) have emerged as a viable solution for addressing this issue, yielding promising results in similar problems.

Endoscopic images captured during an *Esophagogastroduodenoscopy* (EGD) may display extreme variations in perspective and illumination. Regarding the demanding nature of this task, DNN require a great amount of high-quality data in order to have a good generalisation capability. Yet, endoscopic images are expensive to collect and there are few public datasets available. To account for this, two bilinear models that combine a fractal texture descriptor invariant to these theoretical transformations with DNN are purposed.

All the models were implemented in a dataset with Narrow Band Imaging (NBI) frames, which were preprocessed and augmented. A hyperparameter grid-search was conducted in order to establish a set of baseline models. The performance of all the models was assessed through 5-fold cross-validation. The results showed that the highest Sensitivity (0.867) and Positive Predictive Value (PPV) (0.891) were obtained with the proposed approaches which is a good indicator of the quality of the models in detecting GIM. Given the fast prediction ability and the quality of the results, the proposed bilinear models could be included in a EGD as tool to assist endoscopists to diagnose GIM.

Resumo

O cancro gástrico é um dos tipos de cancro mais mortais globalmente. No entanto, o diagnóstico precoce desta doença pode aumentar a taxa de sobrevivência dos pacientes, uma vez que permite o acompanhamento mais frequente dos mesmos. A metaplasia intestinal é uma lesão gástrica pré-cancerígena, que está diretamente relacionada com um elevado risco de desenvolver cancro gástrico. A deteção desta lesão é bastante desafiante tendo em conta as suas características, o que resulta numa concordância baixa entre os médicos e num elevado número de diagnósticos incorretos. Uma solução para este problema que tem obtido resultados bastantes promissores são as redes de *deep learning*.

As imagens endoscópicas que são obtidas durante a endoscopia são caracterizadas por variações extremas de perspetiva e iluminação. Tendo em conta a dificuldade da tarefa, as redes de *deep learning* necessitam de grandes quantidades de dados de boa qualidade para garantir uma melhor capacidade de generalização. No entanto, as imagens endoscópicas são bastante caras de recolher e existem poucos conjuntos de dados públicos disponíveis. Posto isto, neste tese dois modelos bilineares são propostos que, por sua vez, combinam um descritor fractal de textura invariante a estas transformações teóricas nas imagens com redes de *deep learning*.

Todos os modelos foram implementados num conjuntos de dados com imagens em *Narrow Band Imaging (NBI)*, que foram pré processadas e replicadas usando um processo de *data augmentation*. Com os modelos base houve uma procura da combinação dos melhores hiper parâmetros. Para avaliar a performance de todos os modelos foi usado um método designado *5-fold cross-validation*. Os resultados obtidos mostraram que a Sensitividade (0.867) e *Positive Predictive Value (PPV)* (0.891) mais altos foram obtidos com os modelos propostos. Isto poderá ser indicativo de que o descritor de textura melhorou a capacidade das redes de *deep learning* na deteção de metaplasia intestinal num conjunto de dados pequeno. Tendo em conta que os modelos propostos têm uma capacidade de previsão rápida e obtiveram bons resultados, conclui-se que estes poderiam ser eventualmente integrados num exame endoscópico para ajudar os médicos na deteção desta lesão.

Acknowledgements

When I started my journey at the university, the prospect of completing a thesis seemed too scary, and I doubted about my ability to conquer it. However, this year I have not only embraced the process but have also enjoyed it. The thesis subject helped on this, but it was not the main factor.

The group that I entered was always very friendly and welcoming, and my supervisors were always available to guide me and help me when I needed it. Thus, I want to thank professor Miguel Coimbra and professor Francesco Renna. I want to give a special acknowledgement to my third supervisor Miguel Martins who consistently supported and encouraged me to go further, without judging me for my "innocent" questions. It was a pleasure to work with you all.

Furthermore, I also want to thank my family who always supported me in my studies, my friends who helped me to continue having an intense social life full of good memories, and my boyfriend Ricardo who always supported me in my crises, continued listening about my work even if though not understand it and helped me in everything (even in the most boring ones).

As we say in portuguese "O que é bom acaba depressa" (the good things end to quickly)! Thank you!

Contents

List of Tables

List of Figures

Acronyms

EGCM Early Gastric Cancer Model

ME-NBI Magnifying Endoscopy with Narrow-Band Imaging

CNN Convolutional Neural Networks

M-NBI Magnification Narrow-Band Imaging

GPDFENet Gastric Precancerous Diseases Feature Extractor Network

SVM Support Vector Machine

EGC Early Gastric Cancer

DNN Deep Neural Networks

NPV Negative Predictive Value

GIM Gastric Intestinal Metaplasia

ID Intelligent Diagnostic

IEE Image-Enhanced Endoscopy

HSV Hue Saturation Value

LBP Local Binary Pattern

AG Atrophic Gastritis

ROI Region Of Interest

CAM Class Activation Map

PA Per-pixel Accuracy

IOU Intersection Over Union

CADx Computer-Aided Diagnosis

GLCM Gray-Level Co-occurrence Matrix

MFS Multi-Fractal Spectrum

LFD Local Fractal Dimension

NBI Narrow Band Imaging

WLI White Light Imaging

GC Gastric Cancer

EGD *Esophagogastroduodenoscopy*

UGI Upper Gastrointestinal

NN Neural Networks

ILSVRC ImageNet Large Scale Visual Recognition Challenge

ReLU Rectified Linear Unit

ME Magnifying Endoscopy

PCA Principal Component Analysis

AUC Area Under the Curve

LR+ Positive Likelihood Ratio

LR- Positive Likelihood Ratio

PPV Positive Predictive Value

NPV Negative Predictive Value

TP True Positive

TN True Negative

FP False Positive

FN False Negative

TNR True Negative Rate

TPR True Positive Rate

ROC Receiver Operating Characteristic

FPR False Positive Rate

MLP Multi-Layer Perceptron

GAP Global Average Pooling

SGD Stochastic Gradient Descent

CGT Chronic Gastritis

LGN Low Grade Neoplasia

MCC Matthews Correlation Coefficient

BA Blob-Adapted

GLFD Gradient Local Fractal Dimension

BSA Blob Shape Adapted

BS Blob Shape

MB-LBP Multiscale Block Local Binary Patterns

K-NN K-Nearest Neighbours

LOPO Leave-One-Patient-Out

LBP Local Binary Pattern

CLBP Complete Local Binary Pattern

CLASSNet Cross-Layer Aggregation of Statistical Self-similarity Network

Chapter 1

Introduction

1.1 Overview

Gastric Cancer (GC) ranks as the fifth most common type of cancer globally and accounts for the fourth highest mortality rate among cancer-related deaths [60]. Although the 5-year survival rate of patients with advanced gastric cancer is no more than 30% [1], the survival rate for patients diagnosed with early gastric cancer can be 90% [24], which demonstrates that an early diagnosis can be crucial in improving survival rates and gastric cancer management in general. However, results [45] show that 11.3% of Upper Gastrointestinal (UGI) cancer-related lesions are missed during endoscopic screening up to 3 years before diagnosis.

Considering the gastric precancerous cascade [10], it is possible to notice that Gastric Intestinal Metaplasia (GIM) is the precursor of gastric cancer. This condition arises due to prolonged inflammation in the stomach, resulting in a transformation of the gastric mucosa cells and it is called intestinal since the cells start resembling the intestinal phenotype. Patients with GIM are 10 times more likely to have gastric cancer [33].

GIM can be diagnosed during *Esophagogastroduodenoscopy* (EGD), a medical procedure for capturing images from the UGI, with the detection of an aberrant tissue in the gastric mucosa using a Narrow Band Imaging (NBI) modality. NBI is a lighting modality that improves the quality and the detail of the images captured. However, GIM screening can be very challenging due to the details that characterize this condition. Thus, the diagnosis of GIM has a low rate of interobserver concordance. In fact, L.G Capelle *et al.* conducted two similar studies regarding the performance of GIM diagnosis and they observed high discrepancies in Sensitivity and Specificity for endoscopists with different levels of experience (18% of difference in sensitivity and 35% for specificity) [6].

1.2 Motivation

Regarding the problems enumerated before, it is possible to understand the necessity of finding an automatic procedure unaffected by external factors to detect GIM in endoscopic images. An option that is emerging nowadays is the application of Deep Neural Networks (DNN) to perform this task. These approaches have obtained promising results on landmark detection [8], detection of gastric cancer [2], prediction of invasion status [64], and GIM detection [68].

However, the proposed approaches based on DNN have some limitations. On one hand, the collection of endoscopic images can be a very expensive procedure, which results in a very limited number of public and small datasets. On the other hand, these datasets are characterised by their challenging nature since the scale, perspective, and illumination of the images can vary considerably in a different dataset. Moreover, frames might be populated with blood, bobbles, and undigested food. To guarantee that DNN are resilient to these challenge conditions, the demand for more data that isn't available is superior, which will affect the results obtained by these approaches.

Therefore, the main goal of this thesis is to propose a new approach based on DNN combined with a texture descriptor to solve these demanding requirements for the training data. The idea was to find a technique that captures the irregularity of the tissue present in the images obtained during the EGD. Fractal Geometry emerged as an option since it is proven that it can describe irregular objects not idealize by classic Euclidean Geometry. Mandelbrot in [44] indicated that "clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line". Fractal Geometry is a well-founded alternative and its effectiveness was already validated through empirical evidence in many applied fields such as the characterisation of cells in the biological field [42] [58], detection of lesions in medical images [11] [5] and in material engineering [25].

The objective is to induce a bias in the model using a texture descriptor based on the fractal dimension that is characterised by being invariant to transformations in images such as illumination and perspective. The effectiveness of this combination was enhanced by the fractal dimension's ability to capture texture patterns in natural images [66][49], coupled with the discriminative features derived from a related visual task, particularly in the characterisation of polyps during colonoscopy [21].

1.3 Objectives

We outlined the following objectives for this thesis:

- Analyze the current approaches used for GIM detection in endoscopic images based on DNN

- Develop new deep learning techniques capable of detecting precancerous tissue aberrations in endoscopic images
- Research and develop new deep learning approaches able to leverage fractal geometry techniques to improve classification performance
- Assess the quality of the proposed methods

1.4 Contributions

The work developed result in several different contributions:

- Analysis of the current state-of-the-art methods in using deep Convolutional Neural Networks (CNN) for GIM detection.
- Preprocessing of GIM-NBI dataset for integration with the deep learning approaches.
- Systematic comparison over the proposed dataset of the different methods found in the literature. This contribution led to a peer-reviewed publication in an international conference: Martins, M. L., **Pedroso, Maria**, *et al.* Diagnostic Performance of Deep Learning Models for Gastric Intestinal Metaplasia Detection in Narrow-band Images. 2023 45th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC). IEEE, 2023.
- Proposal of a novel deep learning approach for GIM detection able to combine via bilinear models both features extracted with pre-trained CNN and texture descriptor based on fractal geometry. This contribution led to another peer-reviewed article in an international conference: **Pedroso, Maria**, *et al.* Fractal Bilinear Deep Neural Network Models for Gastric Intestinal Metaplasia Detection. IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI 2023). IEEE, 2023.
- The novel deep learning approach also lead to an article submission at a national conference: **Pedroso, Maria**, *et al.* Diagnosis of Gastric Intestinal Metaplasia in Narrow-Band images with Fractal Bilinear Deep Learning Models. RECPAD Portuguese Conference on Pattern Recognition. (RECPAD 2023)

Chapter 2

Background

This section introduces clinical and mathematical background needed to better grasp the scope and algorithmic techniques described through this thesis. Firstly, in Section 2.1 an introduction to gastric cancer is made with the explanation of this condition, the methods used to detect it, and also the importance of Gastric Intestinal Metaplasia (GIM) in its diagnosis. Then, a brief description of how the Convolutional Neural Networks (CNN) works, CNN that adopted in the literature to detect GIM in endocospic videos and their components are presented (Section 2.2). Also, a method called transfer learning is explained.

2.1 Gastric Cancer Screening

Gastric cancer is a pathological process that impacts the genesis and "death" of cells of the stomach in an uncontrolled way. Most cases are considered carcinomas, which is a type of cancer that starts in the epithelial tissue of the skin or internal organs. Regarding the part where it develops, carcinomas are divided into adenocarcinoma, if it develops in an organ or gland, and squamous cell carcinoma, if it originated in the epithelium [47]. Most cancers of the stomach are adenocarcinomas [13], since they develop from the gland cells in the innermost lining of the stomach named mucosa.

As mentioned in Section 1.1, gastric adenocarcinoma is preceded by premalignant lesions (see Fig.2.1). The first phase is the inflammation of the gastric mucosa, usually caused by *Helicobacter pylori infection*, named non-atrophic gastritis. The next stage in the cascade caused usually by a prolonged inflammatory process is characterised by the loss of normal glandular tissue and is designed as multifocal atrophic gastritis (see image A of Fig.2.2). Then, eventually, this will result in the replacement of the original gland tissue with cellular matter that resembles to the intestinal phenotype, that is the intestinal metaplasia phase associated with a high cancer risk (see image B of Fig.2.2). Consequently, the gland cells begin to have the intestinal phenotype. The final stage before cancer is the dysplasia phase (see image C of Fig.2.2), which is characterised by the abnormal development of cells within the tissue. It results in an irregular shape of the

cells, and occasionally in a bifurcated or branching aspect [10].

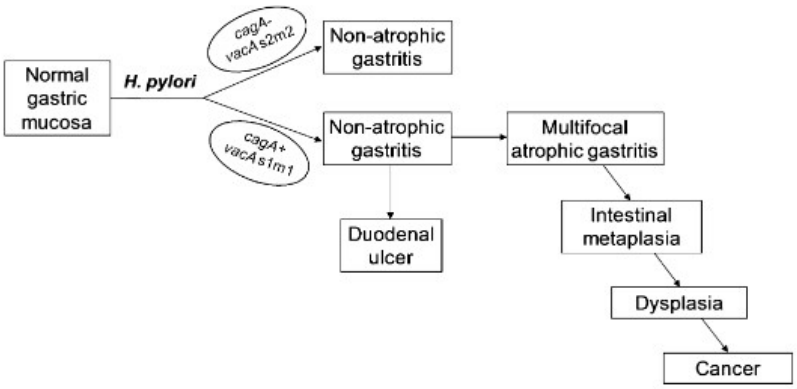


Figure 2.1: Precancerous gastric cascade. Image source: [10]

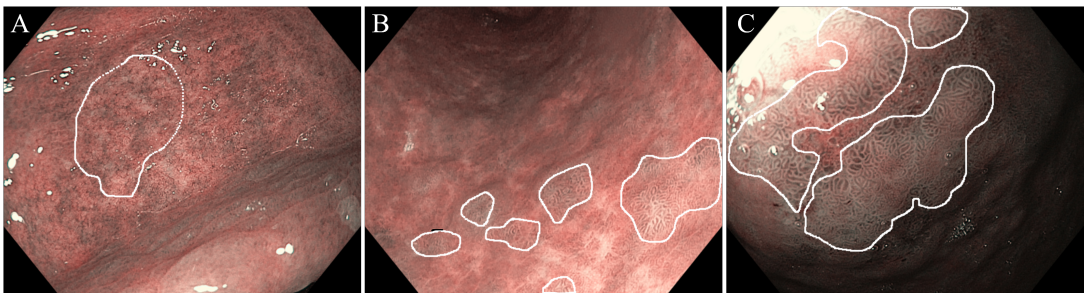


Figure 2.2: Precancerous gastric lesion. A: Atrophic Gastritis; B: Intestinal Metaplasia; C: Dysplasia.

Esophagogastroduodenoscopy (EGD) is the gold standard screening procedure for the above mentioned pathologies. This procedure consists of an exam in which a doctor inserts a flexible and thin tube with a small camera on the tip in the patient's mouth and this tube is guided into the throat, esophagus, stomach, and duodenum. The camera collects videos that are used by the gastroenterologist to detect the lesions (see Fig. 2.3). The phase of the gastric precancerous cascade is recognised regarding the aspect of the tissue. Following the descriptions present in [62], non-atrophic gastritis is characterised by macroscopic nodules, gastric fold hypertrophy, and changes in vascular density. Atrophic gastritis is identified by paleness, the absence of gastric folds, and the prominence of blood vessels. Intestinal metaplasia is distinguished by elevated areas with grayish-white patches, encircled by pale and typically coloured gastric mucosa, or blotchy patchy erythema. Finally, dysplasia is recognized by depressed or raised abnormalities, absence of a discernible vascular pattern, or subtle alterations in colouration (see Fig. 2.2).

The endoscope may support different lighting modalities. The conventional one is White Light Imaging (WLI), which consists in using white light to illuminate the gastric mucosa and then capture the images. Nowadays, Image-Enhanced Endoscopy (IEE) has emerged as a new option, since the diagnostic accuracy of WLI is low [72]. IEE refers to techniques for improving the visualisation and lesion detection on images. One technique is Narrow Band

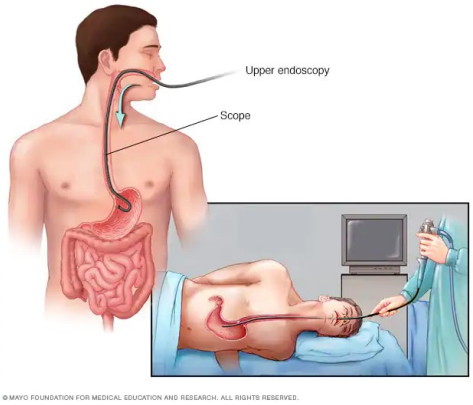


Figure 2.3: EGD procedure. Image source: https://www.mayoclinic.org/tests-procedures/endoscopy/about/pac-20395197

Imaging (NBI) which is based on the idea that the wavelength of light determines how deep the light penetrates the tissue. Blue and green lights are used since their wavelength corresponds to the hemoglobin absorption which results in vessels appearing in dark, highlighting the vasculature of the superficial mucosa (see Fig. 2.4). The potential of NBI is improved with the use of Magnifying Endoscopy (ME) which is a zoom method that can enlarge an image up to 100 times. This method allows a better inspection of the microvascular architecture and the micro surface structure, thus enhancing optical diagnosis of these precancerous lesions [59].

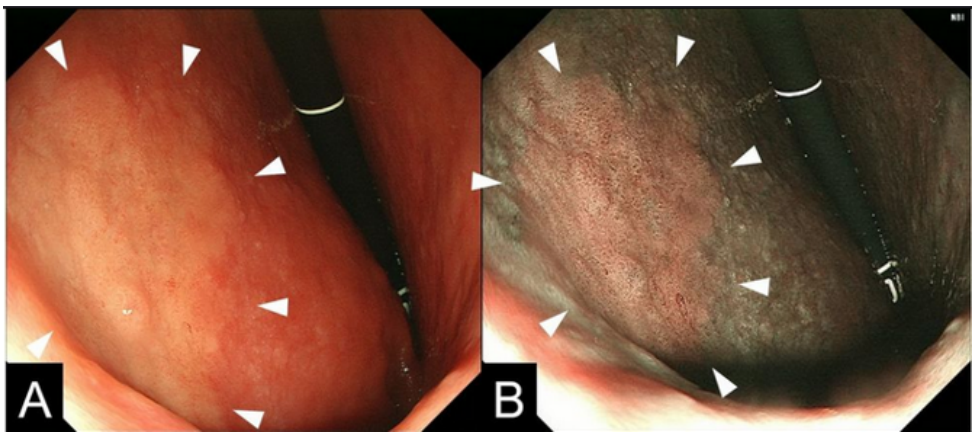


Figure 2.4: Endoscopic image of early gastric cancer. A: WLI image; B: NBI image. Image source: [69]

Lesions are firstly identified through optical diagnosis by a human physician. The detection of GIM requires well-trained endoscopists [65] since this condition is characterised by having a fine-grained variability on the gastric mucosa [68]. Anatomically, this condition lesion can be categorised into two types: limited, if it is present only in one region or extensive if it is scattered in the stomach. Considering the structure of the cells, it can be classified as either complete or incomplete. The worst case is when an incomplete and/or extensive metaplastic tissue is observed in the mucosa [26], which is related to high risk of developing to an advanced gastric

cancer [14] [53].

2.2 Convolutional Neural Networks

CNN are a type of the known Neural Networks (NN). The CNN have mainly two benefits compared with the NN that are fundamental in practice. The number of parameters necessary to the NN to work with complex structures is decreased in the CNN with the weight-sharing process in the convolutional layers. The lack of inductive bias over spatial reasoning which is a consequence of these neural networks being fully connected is tackled by each neuron in CNN only having access to some elements in the neighboring region of the previous layer.

CNN are composed of different types of layers:

- Convolution Layers: These layers are the main part of this architecture and they are responsible for applying a convolution operation to the input data. A convolution operation consists of a sliding window over the input tensor, where the element-wise product with a learnable kernel is made. Each selected region will be characterised by the sum of this element-wise multiplication. The result of the convolution operation is a matrix composed of the results obtained for each region considered, called feature maps (see Fig. 2.5). A layer can have more than one filter, so if it has N filters the result of the convolution layer is N feature maps. This layer can be characterised by the number of filters, the number of input channels, and its filter sizes. Furthermore, it is also possible to define some parameters regarding how the filters slide over the images. These parameters are the strides, that determine the distance (measured in numbers of samples) between the regions where filters are applied, and the padding that consists in adding extra pixels outside the image to change the considered positions affected by the filters.

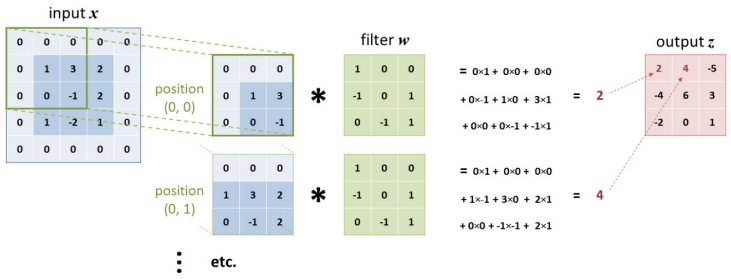


Figure 2.5: Example of a convolution operation. Image source: [50]

- Pooling layers: They consist of sliding over the input matrix and applying a basic function to the selected region. The usual functions used are the maximum or the average function. The result is a matrix with the output of one of these functions applied to each region. These layers are commonly defined with a stride value that determines the size of the receptive field and padding. They are applied to each channel of the input matrix which is a 2D

matrix with different starting weights. Considering this matrix $A = [a_{ij}]_{i=\{1,...,M\},j=\{1,...,N\}}$ for computing the max-pooling with a stride of s and a padding of p we can use the following expression:

$$B = [b_{ij}]_{i=\{1,...,M+p-s+1\},j=\{1,...,N+p-s+1\}}\,, \tag{2.1}$$

where $b_{ij} = \max(a_{i+k,j+l}, k, l \in \{0, ..., s-1\})$. The average pooling can be computed with the expression:

$$C = [c_{ij}]_{i=\{1,...,M+p-s+1\},j=\{1,...,N+p-s+1\}}, \tag{2.2}$$

where $c_{ij} = \frac{1}{S^2}\sum_{k}^{S-1}\sum_{l}^{S-1} a_{i+k,j+l}$.

- Dropout layers: As the name suggests, they consist in a 'drop' technique. Regarding a dropout probability of p established with the layer, the dropout layers will drop p of the neurons in every iteration, and keep $1-p$. The drop is made randomly and it consists in forcing the weights of the deactivated neurons to be zero so that they will be ignored in the next training phases. It has regularisation effect. This results in a network less dependent on specific weights of the neurons, and, consequently, with a better generalisation capability. These layers are used to prevent overfitting and are just used in the training of the network. During inference, dropout is deactivated.

- Fully connected/Dense layers: As the name suggests these layers connect all the neurons in the network. Each neuron is connected to all the output values obtained from the previous layers. These layers are used to perform classification starting from the feature vectors that are extracted from the image using convolution and pooling layers.

With all these types of layers, researchers created different combinations and tried to stack more and more layers to produce new networks to improve the efficiency and accuracy of the networks.

CNN are optimised for a specific task with the available data. This optimisation is done by training the network, which means finding the optimal parameters for the task proposed. The search is made using a loss function, which is a method for assessing the performance of the model in the dataset by measuring the difference between the predicted value and the true one. The goal of this training procedure is to minimise the loss function by finding the best parameters that produce the most accurate predictions. This minimisation is done by applying an optimiser.

There are many loss functions used in these algorithms, but the most common one in classification problems and the one used in this work is the cross-entropy. This function measures the difference between the predicted and the true value using the predicted probability distribution and the actual one. For a classification problem, consider P the target distribution and Q the predicted one. The cross-entropy is computed as follows:

$$H(P, Q) = -\sum_{x \in X} P(x) \log Q(x), \tag{2.3}$$

where $P(x)$ is the probability of the event x in P and $Q(x)$ is the probability of the event x in Q.

Regarding the optimisers, there is also a wide range of options. The most basic one is the gradient descent which is based on computing the first-order derivatives of the loss function with respect to the parameters of the network in order to know which changes are needed to apply to the parameters (Fig. 2.6). Considering the W_l as the weights for layer l, b_l as the bias for layer l and L the loss function for gradient descent we have the following:

$$W_l = W_{l-1} - \frac{\partial L}{\partial W_{l-1}} \tag{2.4}$$

$$b_l = b_{l-1} - \frac{\partial L}{\partial b_{l-1}} \tag{2.5}$$

A variant of gradient descent algorithms is the Stochastic Gradient Descent (SGD) which is computed in a similar manner, but instead of updating the model parameters just once for each iteration, it changes the parameters based on the loss value evaluated for each observation in the dataset. If the dataset has 30 data points, in one iteration of the model the SGD will update the parameters 30 times. This is an advantage compared with the gradient descent since with frequent updates it is more likely to catch local minimums and not lose the optimal. Another variant is the gradient descent algorithm with momentum, which is based in the SGD. The difference is that this method adds the previous updates to the current one controlled by a factor named momentum. Considering the W_t as the weights at time t, b_t as the bias at time t, L the loss function selected and α the learning rate the momentum μ is applied in the following way:

$$W_t = W_t - (\alpha \frac{\partial L}{\partial W_t} + \mu W_{t-1}) \tag{2.6}$$

$$b_t = b_t - (\alpha \frac{\partial L}{\partial b_t} + \mu b_{t-1}) \tag{2.7}$$

The idea is to accelerate the SGD if it is in the same direction as the previous one and decelerate in the opposite case.

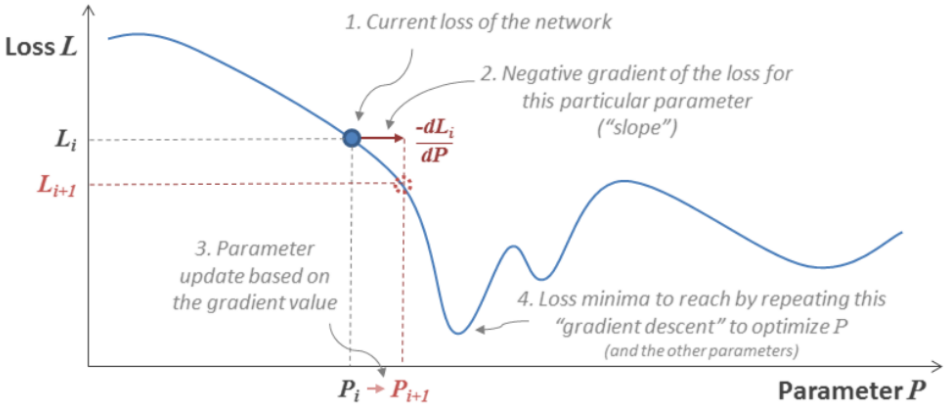


Figure 2.6: Representation of the gradient descent method. Image source: [50]

In equations 2.6 and 2.7, a parameter named learning rate (α) is present. This factor is multiplied by the derivatives to control how much a parameter should be updated. In the previous

optimisers is constant over all the training procedure. In more complex s such as those in the Ada family the learning rate is adaptive over every iteration. This family is composed by three optimisers which are variations of different ideas to change the learning rate: Adagrad, Adadelta, and Adam. Adagrad [12] uses a formula to determine how much the learning rate will decrease regarding the frequency of the features connected to the parameters. Using the same notation of the equations (2.6) and (2.7), Adagrad is computed as follows:

$$W_t = W_{t-1} - \eta'_t \frac{\partial L}{\partial W_{t-1}}, \tag{2.8}$$

where $\eta'_t = \frac{\eta}{\sqrt{\alpha_t + \epsilon}}$, $\alpha_t = \sum_{i=1}^{t} \left(\frac{\partial L}{\partial W_{t-1}}\right)^2$ and ϵ is a small number to prevent the division by 0 does not happen.

Adadelta [70] effectively addresses the issue of a decaying learning rate encountered in the Adagrad optimiser. This problem arises when the learning rate is exceedingly small, causing the network to stop learning. Adadelta solves this problem by controlling the factors used to divide the learning rate for each parameter. For this optimiser the learning rate is updated in the following way:

$$\eta'_t = \frac{\eta}{\sqrt{\mathrm{S_t} + \epsilon}}, \tag{2.9}$$

where $S_t = \gamma \mathrm{S_{t-1}} + (1-\gamma)(\frac{\partial L}{\partial W_{t-1}})^2$ is the exponential average of squares of gradients and γ is a parameter usually set to 0.9. To update the parameters in each time step t the equation described in (2.8) is used with the different η'_t.

Finally, Adam [29] uses the same method to update the learning rate as Adadelta. Furthermore, it also uses the past momentum values to update the parameters. In the Adam optimiser, the parameters are updated as follows:

$$W_t = W_{t-1} - \frac{\eta}{\sqrt{S_t - \epsilon}} * V_t, \tag{2.10}$$

where $V_t = \beta V_{t-1} + (1-\beta)\frac{\partial L}{\partial W_{t-1}}$ is the exponential weighted average of the past gradients and the S_t is the same defined for the Adadelta. The equations to update the bias parameter in the three optimisers are the same as the weights.

An important CNN that launched the success of this techniques, was the AlexNet in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [54]. After AlexNet some new architectures were developed. Next, some of the network architectures adopted in previous works for GIM detection are described. These works are listed and described in more detail in Chapter 3.

2.2.1 VGG-16

The VGG architecture was proposed by K. Simonyan and A. Zisserman in [57]. The authors proposed an architecture composed of five blocks of several convolution layers followed by a max-pooling layer and three final dense layers. The VGG-16 stands for “Oxford Visual Geometry

Group (VGG)", and 16 has to do with the number of layers (13 convolutional layers and 3 dense layers). The scheme is presented in Fig. 2.7. The convolution and the max-pooling layers have the same padding, and the dense layers use the Rectified Linear Unit (ReLU) as the activation function.

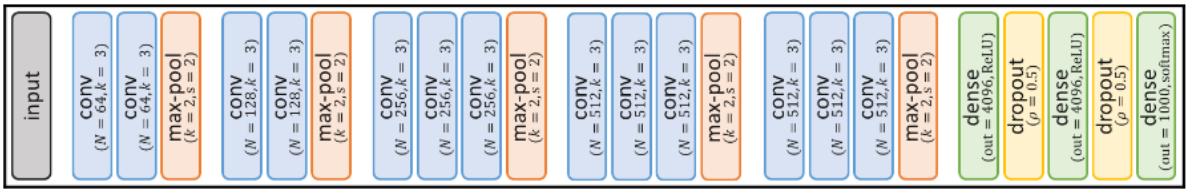


Figure 2.7: VGG-16 architecture. Image source: [50]

The novelty introduced in VGG was the use of convolution layers with small filters, 3×3. This results in a decrease of the number of parameters (compared with AlexNet) and an increase of the non-linearity capacity. Increasing the number of convolution layers in a neural network, and applying a non-linear activation function like ReLU after each layer, enhances the network's ability to learn more complex features.

2.2.2 ResNet-50

ResNet was developed by Kaiming He *et al.* in [16], based on a new concept, the residual modules. They provide a better approach to creating deeper networks. The motivation was that deeper networks are harder to train, due to the vanishing gradient problem, and as the network depth increases the performance can be degraded. Kaiming He *et al.* in [16] pointed out that deeper networks can suffer for a degradation problem since accuracy gets saturated and higher training errors are obtained. This was corroborated with the comparision of a 18-layer-deep CNN with a 34-layer one which resulted in a higher training error throughout the whole training procedure.

The ResNet architecture (see Fig. 2.8) is composed of residual blocks where two paths are considered: the non-linear and the identity path. The first one processes the feature maps with some convolution operations and ReLU as the activation function. This path also include a batch normalisation step, which consists in the standardisation of the layer's input resulting in a more stable and faster training process [22]. The second path does not apply any transformation. The outputs of these two branches are merged using element-wise addition. The idea behind this is that if an identity path is used the performance obtained is at least preserved and the transformations obtained by the other path will only be preserved if it benefits the network. These blocks do not contain more parameters than the traditional ones, so they can be used to construct deeper networks without significantly affecting the training process.

Similarly to the VGG architectures, ResNet can have multiple versions regarding the depth considered. A ResNet-50 is characterised by having 50 layers.

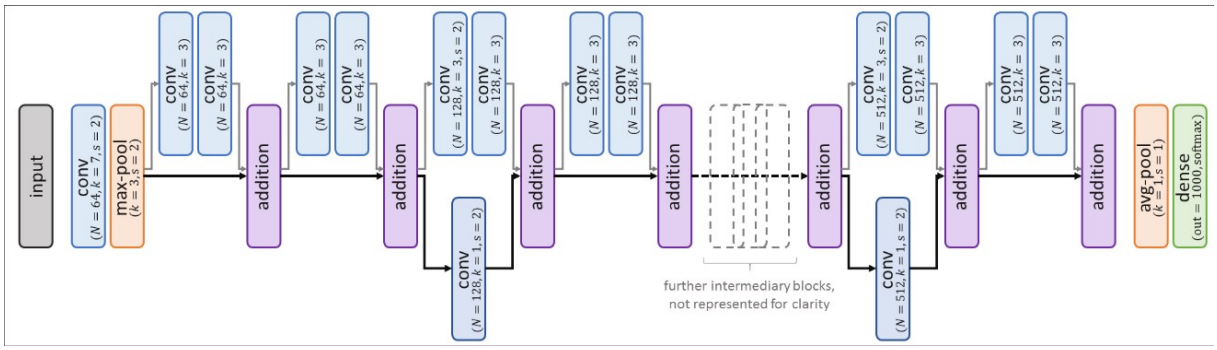


Figure 2.8: Example of the ResNet architecture. Image source: [50]

2.2.3 DenseNet-121

DenseNet, proposed by Gao Huang *et al.* in [20], is a network in which all layers have a direct connection to all the subsequent ones. The idea is to introduce connections between layers with the same feature size in order to maximise the information flow between layers. Moreover, the input information in each layer increases which is then passed on their feature maps to the subsequent ones.

Similarly to the ResNet (2.2.2), the DenseNet uses the idea that some feature maps are redundant and that the network can skip them. However, instead of using element-wise summation to combine the identity function with the output, the DenseNet concatenates the feature maps produced in each layer into a single tensor. Then a composite function composed by a batch normalisation, followed by a ReLU and 3×3 convolution is applied. To simplify the architecture's down-sampling, the DenseNet comprises multiple densely connected blocks connected by transition layers that consist of a batch normalisation layer, a 1×1 convolution layer, and a 2×2 average pooling layer.

A novelty of this network is the introduction of a hyperparameter named the growth rate. This hyperparameter determines the number of input feature maps for each layers. Furthermore, bottleneck layers were also added to reduce the number of input feature maps of each layer and a compression factor was applied to reduce the number of feature maps in the transition layers. Transition layers are the layers between the dense blocks composed by a convolution and a pooling layer (see Fig. 2.9).

This architecture introduces the use of feature reuse, concatenating feature maps learned by different layers, improving network efficiency. The structure is illustrated in Fig. 2.9.

DenseNet can have multiple versions regarding different growth rates and depths. DenseNet-121 stands for the DenseNet architecture with a depth of 121.

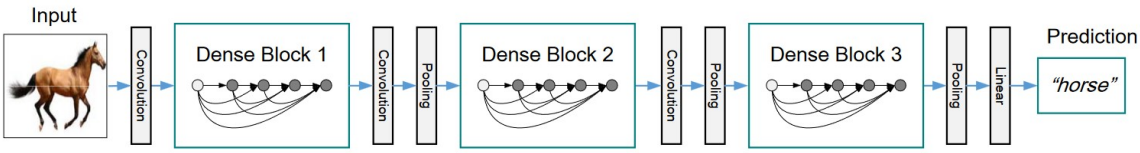


Figure 2.9: Example of the DenseNet architecture. Image source: [20]

2.2.4 EfficientNet-B4

The EfficientNet architecture was proposed by Mingxing Tan and Quoc V. Le. in [61], as an attempt to achieve better accuracy based on a new way of scaling up a network. In order to obtain better results when it is possible networks are scaled up, following empirical procedures that take time to test and, usually, result in a little increase in performance.

Mingxing Tan *et al.* [61] notice that instead of finding the best layer architecture, the main goal was to find the best way of expanding the network length, width, and resolution without changing the established baseline model. In their work, they concluded that these dimensions could be used individually to enhance the network performance. However, to achieve a greater improvement they have to be balanced. The compound scaling method is based on this idea. It uses a compound coefficient that is user-specified regarding the resources available and three constants one for each dimension. They are computed using a grid search to determine how to assign the extra resources.

The building blocks of the EfficientNet are the mobile inverted bottleneck MBConv that are also used in MobileNetv2 [55]. MBConv is a residual block that works in an inverted way, since it starts to increase the size of the input and then it decreases so that the output has the same size as the input, which is the opposite of what happens with a normal residual block (Fig. 2.10 and Fig. 2.11). Thus, it starts by applying a 1x1 convolution, then a 3x3 depthwise convolution (to reduce the number of parameters), and, finally, a 1x1 convolution again to reduce the number of channels.

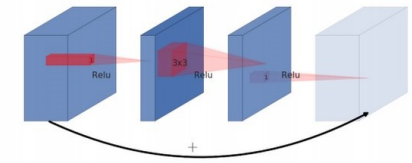


Figure 2.10: Example of a residual block. Image source: [55]

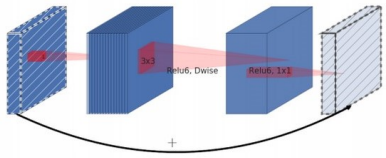


Figure 2.11: Example of a inverted residual block. Image source: [55]

The different versions of EfficientNet, from B0 to B7, were developed regarding different compound coefficients. The structure of the EfficientNet-B0 is illustrated in Table 2.1.

Type of layer	**Kernel size**
Convolution	3×3
MBConvolution1	3×3
MBConvolution6	3×3
MBConvolution6	5×5
MBConvolution6	3×3
MBConvolution6	5×5
MBConvolution6	5×5
MBConvolution6	3×3
Convolution	1×1
Pooling + Fully connected	7×7

Table 2.1: Structure of the EfficientNet-B0.

2.2.5 Transfer Learning

Transfer learning is a technique used in machine learning algorithms that consists in reusing a model trained for a task as a starting point for another model with a different goal. This is a common method used in deep learning for computer vision problems since these models are computationally heavy to train from scratch and time-consuming.

Thus, the implementation of transfer learning in CNN is based on the idea that the first layers of a network learn the generalistic low-level features. In deep learning models for image data, the common procedure is to apply a pre-trained model in the ImageNet dataset [54] composed of 14197122 annotated natural images distributed by 1000 classes. There are mainly two procedures that can be applied:

- Modify the classifier, which consists of the final layers of the network, to suit the new classification task. Thus, the classifier will be trained from scratch using the feature maps obtained by the pre-trained model in the other dataset. All the layers that are before the classifier are "frozen", which means that they will not be trained. Since the base convolutional network was trained with images, these layers are expected to contain the generic features that are common for images and be reused.
- Instead of just training the classifier, some final layers can also be unfrozen. The idea is that these layers contain more specific features of the images. Thus, it makes sense to train them for the new specific task since they need to be more specific for the new task.

Chapter 3

Literature Review

In order to have an organised and useful review, a query was used on a set of databases to search the articles, and a selection of the most relevant ones was made regarding some criteria. This chapter is organised as follows: first, Section 3.1 contains a description of the model and the approach used to choose the most relevant works. Then, Section. 3.2 contains a description of the articles selected and, finally, in Section 3.3 conclusions are drawn.

3.1 Search and Review Methodology

The articles included in this review were found using a given query in the following databases: PubMed, Scopus, and IEEE. In order to have an efficient search, a model named PICO [46] was used to create the query. This models consists of a format of organising keywords into a clinical question regarding four categories: **P**acient/problem, **I**ntervention, **C**ontrol/comparison and **O**utcome.

The chosen keywords were organised into four categories to create the final query.

P - Pacient or problem

(“Early Gastric cancer” OR “Gastric cancer” OR “Gastric mucosal lesions” OR “Gastric precancerous disease” OR “Chronic atrophic gastritis” OR “Intestinal Metaplasia” OR “Neoplasia” OR “Dysplasia”)

I - Intervention

(“Deep learning” OR “Convolutional Neural Networks” OR “Computer vision” OR “M-NBI” OR “ME-NBI” OR “ME-WLI” OR “ME-BLI” OR “Narrow band imaging ” OR “Magnifying endoscopy”)

C - Comparison

-

O - Outcome

(“Image classification” OR “Artificial intelligence” OR “Metaplasia detection” OR “Anomaly detection”)

So, the final query was the following:

((“Early Gastric cancer” OR “Gastric cancer” OR “Gastric mucosal lesions” OR “Gastric precancerous disease” OR “Chronic atrophic gastritis”OR “Chronic atrophic gastritis” OR “Intestinal Metaplasia” OR “Neoplasia” OR “Dysplasia”) AND (“Deep learning” OR “Convolutional Neural Networks” OR “Computer vision” OR “M-NBI” OR “ME-NBI”OR “ME-WLI” OR “ME-BLI” OR “Narrow band imaging ” OR “Magnifying endoscopy”) AND (“Image classification” OR “Artificial intelligence” OR “Metaplasia detection” OR “Anomaly detection”))

With the query, the following number of articles was obtained:

Repository	Nº of results
IEEE	71
PubMed	203
Scopus	61

Table 3.1: Number of articles obtained in each database.

Given that 264 articles were obtained, a selection was made and some attributes were chosen to organise the important information.

Regarding the importance of the fractal dimension for the project, it was necessary to search for articles about this topic. A query was also created but given the novelty of this subject in this area, no article was found. Thus, a more general search was made and two articles that use fractal dimension for the classification of medical images were selected.

3.1.1 Selection of the Articles

The articles were selected regarding relevance, first considering the title and then the abstract. The inclusion criteria was the implementation of Convolutional Neural Networks (CNN) for detecting early and advanced gastric cancer in endoscopic images. Thus, after this filtering step the number of articles obtained for each database is reported in Table 3.2.

3.1.2 Extraction of the Information

In order to extract the relevant information from the articles in an efficient way and to allow the comparison between the methodologies implemented, some attributes were chosen to characterise them (see Table 3.3).

Repository	Nº of results
IEEE	2
PubMed	11
Scopus	4

Table 3.2: Number of articles obtained in each repository after filtering.

Attributes	**Meaning**
Main goal	The main goal of the article
Algorithm	Machine learning algorithms used in the article
Techniques	Different techniques used (e.g., data augmentation or transfer learning)
Nº of images	Number of images of the dataset used in the article
Nº of patients	The number of patients analysed to obtain the images
Performance	Results of the evaluation metrics for the best model obtained

Table 3.3: Attributes used to extract the relevant information of the articles.

3.2 Results

The relevant information of each article regarding the criteria presented in Table 3.3, is described in Table 3.4.

With the information gathered, it was possible to understand that the articles obtained can be grouped into three categories: i) base deep learning models for classification, ii) classification with the addition of segmentation methods, and iii) the use of fractal dimension for image classification. These three categories will be explained in the following sections.

3.2.1 Classification

Despite the main goal of the project being to detect intestinal metaplasia in endoscopic images, some of the articles considered are about detecting different kinds of lesions, or just differentiating between cancerous and non-cancerous tissue. However, a lot of work has been done in this area and some methodologies used and some of the conclusions obtained were considered useful for the scope of this thesis.

- H. Hu *et al.* [19] developed a computer-aided model named Early Gastric Cancer Model (EGCM) to detect lesions of early gastric cancer in Magnifying Endoscopy with Narrow-Band Imaging (ME-NBI) images based on a VGG-19 pre-trained on the ILSVRC-2012 datasets [31]. They fine-tuned the network and they used the Grad-CAM [56] to produce heat maps to understand the area that most contributed to EGCM's prediction. They also

compared the performance of the model with experient endoscopists and concluded that the best results were obtained when the neural networks and the endoscopists were used.

- X. Liu *et al.* [40] showed different CNN approaches to classify Magnification Narrow-Band Imaging (M-NBI) endoscopic images into three classes: Chronic Gastritis (CGT), Low Grade Neoplasia (LGN), and Early Gastric Cancer (EGC). So, they conducted three main experiences: testing different CNN, assessing the difference in the performance using CNN features or traditional handcraft features, and evaluating the usefulness of transfer learning. For the first experiment they concluded that the performance of ResNet50 was the best, with the second one they noticed that features extracted by fine-tuning pre-trained CNN were more suitable for M-NBI image classification problems than traditional handcraft features and, finally, with the last experiment, they concluded that transfer learning of CNN are more efficient in contrast to directly training CNN architectures.

- L. Li *et al.* [35] described the application of a CNN to classify images as cancer or non-cancer. They used an Inception-v3 network to do that and then they compared the metrics obtained by the CNN with the classification made by experts and non-experts. They visualised that the results were better for the CNN and that CNN can work as a second diagnosis that may help reduce diagnostic errors made by the endoscopist.

- H. Borgli *et al.* [4] presented the largest image and video dataset of the gastrointestinal tract, *HyperKvasir*, available today. After explaining all the characteristics of the dataset, five different models were applied to the dataset to classify the images into 23 different classes and to show the quality of the dataset created.

- J. Islam *et al.* [23] presented a model to detect gastric precancerous diseases composed by two different models: a Gastric Precancerous Diseases Feature Extractor Network (GPDFENet) that works as feature extractor and a Support Vector Machine (SVM) that receives these features and makes the classification of the images in erosion, polyp or ulcer. They verified that in terms of accuracy, the proposed model outperformed previous pre-trained networks.

- Y. Horiuchi *et al.* [18] described the design of a CNN to differentiate between EGC and gastritis using ME-NBI images. The Deep Neural Networks (DNN) used was a GoogLeNet and transfer learning was implemented. They verified that the model was able to differentiate between EGC and gastritis in a short time with high sensitivity and Negative Predictive Value (NPV).

- X. Liu *et al.* [39] developed a transfer learning approach by fine-tuning deep convolutional neural networks to classify M-NBI images into two classes: normal gastric and EGC. Furthermore, with this model, they also conducted 4 experiences to conclude that: 1) transfer learning of CNN architectures has a strong potential in M-NBI image classification; 2) the performance of original coarse dataset trained by the deep CNN is better than fine M-NBI images; 3) it is more suitable to build or reform the deep CNN by using optimal feature extraction modules for M-NBI images; 4) size of the input image has a great influence on the performance of the deep CNN.

- T. Hirasawa *et al.* [17] presented the use of a CNN based on a different algorithm called Single Shot MultiBox Detector, to detect gastric cancer in endoscopic images. This network is designed for object detection, which means that instead of giving a prediction the output is a bounding box that contains the lesion. The model was capable of analysing 2296 images in just 47 s achieving very good results (sensitivity of 92,2%). Furthermore, the undetected cases were considered difficult to distinguish from gastritis even for professionals.

In addition, some work were focused on detecting intestinal metaplasia and atrophic gastritis in endoscopic images, which helped understand the common models used for this tasks and how to improve them, and also some preprocessing methods that it is common to use with this kind of data.

- P. Guimarães *et al.* [15] developed a deep learning approach that overcomes the limitations of white-light endoscopy in diagnosing atrophic gastritis. They decided to focus on the proximal stomach, and since the dataset was small they used a fine-tuned CNN and data augmentation. The best model obtained was a VGG-16. In the end, they obtained an algorithm that was not dependent on high-quality images and they concluded that the deep learning approach developed was better than the experts in the diagnosis of atrophic gastritis.
- T. Yan *et al.* [68] developed a system based on narrow-band and magnifying narrow-band images to detect Gastric Intestinal Metaplasia (GIM), using an EfficientNetB4 with some improvements in the structure. Data augmentation was used to improve the performance and heat maps to determine where the model focused to make the classification. They concluded that GIM diagnosis produced by the Intelligent Diagnostic (ID) system and human experts showed no significant difference in terms of diagnostic performance. However, the diagnosis time of the ID system was much faster than that of the experienced endoscopists.
- M. Xu *et al.* [65] constructed a deep convolutional neural network system, named ENDOANGEL, to detect intestinal metaplasia and gastric atrophy by Image-Enhanced Endoscopy (IEE). They decided to use VGG-16 to develop the system combined with transfer learning. To prevent overfitting they used dropout, early stopping, and data augmentation. The system was compared with experts and non-experts, obtaining results similar to the first ones and better than the second ones, respectively.
- H. Li *et al.* [34] proposed a novel multi-feature fusion model to identify GIM that considers the different features of intestinal metaplasia images such as shape, colour, texture, etc. With this, they used the same three-way pre-trained ResNet model as a parallel network, and take RGB images, Hue Saturation Value (HSV) images, and Local Binary Pattern (LBP) texture features as the input of the three-way network. Then the extracted three types of features are fused based on an attention mechanism, and finally, the GIM is identified through a regularisation module. They compared their method with single-featured networks and existing methods in terms of recognition accuracy and concluded that the results obtained were superior.

- Y. Zhang *et al.* [71] designed a convolutional neural network to improve the diagnostic rate of chronic atrophic gastritis. The model used was a DenseNet and they also implemented heat maps in the images to verify whether the deep learning model had learned information regarding the detection of chronic atrophic gastritis. The results obtained indicated that the detection rates for moderate and severe atrophic gastritis by CNN were higher than those for mild atrophic gastritis and that lesions detected by the model were consistent with those detected by doctors, verifying the accuracy and validity of the model trained.

- N. Lin *et al.* [36] proposed a CNN for simultaneous recognition of Atrophic Gastritis (AG) and GIM. The CNN model used was TResNet and due to the small size of the dataset, data augmentation was applied. They concluded that the CNN system based on endoscopic white light images achieved high sensitivity, specificity, and accuracy for recognition of AG and GIM.

3.2.2 Classification and Segmentation

Some authors decided to go further and present some approaches that implement two methodologies: classification and segmentation of lesions in endoscopic images. These articles showed how it was possible to combine these two tasks in order to achieve better results and also the common models used.

- J. Y. Nam *et al.* [48] reported the development of a model for the diagnostic of endoscopic images named the AI-Scope model. This model consists of 3 consecutive models: one for detecting gastric mucosal lesions (AI-LD); another to provide differential diagnoses for gastric mucosal lesions such as benign gastric ulcers, early gastric cancers, and advanced gastric cancers (AI-DDx); and another for estimating early gastric cancer invasion depths (AI-ID). The first model used, respectively, was a U-Net and the other two were CNN. The results obtained showed that the AI-Scope model was superior in detecting abnormal mucosal lesions, that it differentiated the lesions with a higher performance than endoscopists with intermediary experience, and it estimated the invasion depth of early gastric cancer better than endoscopic ultrasound.

- L. Ma *et al.* [43] described the development of two algorithms to detect early gastric cancer: one for classification and one for segmentation. For the classification method, they obtained a CNN based on ResNet-50 that was called GAIN-ResNet-50 and the segmentation model designed was CA-U-Net that was based on U-Net. To preprocess the images they implemented a method called Region Of Interest (ROI). Furthermore, to evaluate the classification model they used the Class Activation Map (CAM), the usual metrics, and for the segmentation, they applied two metrics: Per-pixel Accuracy (PA) and Intersection Over Union (IOU). Finally, they concluded that the results obtained for the two developed models were very good compared with other usual models.

- T. Kanesak *et al.* [27] developed a Computer-Aided Diagnosis (CADx) system to identify and delineate early gastric cancer in M-NBI images. They started to preprocess the images with histogram equalisation, Gaussian filtering, and partition them into blocks. Then they determine Gray-Level Co-occurrence Matrix (GLCM) features that result in the variation vector that is given to the first classifier (SVM). This classifier determines if the image is abnormal or normal, and if it is normal the P and Q GLCM feature vectors from the cancerous and noncancerous blocks are collected and another classifier is implemented to detect the cancerous region. The results obtained showed that the system has the potential to assist endoscopists in real-time diagnosis and delineation of EGC in M-NBI images.

3.2.3 Fractal Dimension

CNN are characterised by having a strong dependence on the quantity of data needed to train [31]. Furthermore, the generalisation capability of these networks is also determined by the diversity of the data used to train. However, due to legal constraints and structural problems in data collection, gastric endoscopic images are scarce and of varying quality and sometimes it is difficult to classify these images using just deep learning models or obtain models capable of performing well in different datasets.

One way to solve these problems is based on a key component in image analysis: texture. Results from many articles have shown that this component can give very good insights about the analysis of different medical images [41]. Y. Xu *et al.* [66] indicates that most texture descriptors are sensitive to changes in viewpoint and lighting. Thus, they suggested a novelty texture descriptor named Multi-Fractal Spectrum (MFS) that is based on fractal geometry. They showed that this descriptor is globally invariant under the bi-Lipschitz transform, a very general transform that includes perspective transforms and general texture surface deformations.

Some studies have already combined fractal geometry with CNN in their classification methodologies [67] [9]. G. F. Roberto *et al.* [52] described the development of a computer-aided diagnosis system for histological image analysis. This strategy is composed by two modules: extraction of local features by applying fractal techniques and the implementation of two ResNet-50 whose goal is to perform classifications to obtain an array of probabilities. The input of the first CNN is an artificial image generated from the features extracted and the second CNN receives the original image, wherein the class probabilities obtained from the classification of the such image are summed to the respective class probabilities from the first CNN. The values for the accuracy in the different datasets were high and a shorter training time was verified in comparison with other approaches.

Although the idea of having an invariant texture descriptor to transformations like lighting, texture deformations, or perspective changes in endoscopic images could be a possible solution to the lack of high-quality gastric endoscopic images needed to obtain a good performance with CNN this idea has not been extensively explored. Just, M. Häfner *et al.* [21] presented an approach to classify colonic polyps in endoscopic images using texture analysis methods that are

based on computing Local Fractal Dimension (LFD). Eight CC-i-Scan databases and an Narrow Band Imaging (NBI) database were used to compare different LFD based approaches and also to evaluate the difference between this method and other colonic polyp classification methods not based in LFD. Furthermore, a public texture image database, the UIUCtex database [32], is used to test the viewpoint invariance of the methods. The results showed that some of the LFD based methods were the best performing methods or at least were among the best ones. LFD based approaches are more viewpoint invariant than the other approaches.

3.3 Discussion

Even though the main objective of this thesis is the detection of Gastrointestinal Metaplasia (GIM) in endoscopic images, many articles about detecting other kinds of lesions were considered since our search did not yield sufficient results (just four) in this specific task.

The set of articles regarding classification and segmentation was useful in understanding that ResNet, VGG, and Inception were the most used models for detecting lesions in endoscopic images. In this area is very common to use transfer learning and cross-validation. Some approaches also used data augmentation to compensate for the lack of data. Furthermore, it was possible to notice that the results obtained were, in general, promising with an accuracy and sensitivity higher than 85% (except in two works). Nonetheless, two limitations were prevalent across many of the articles: the need to validate the model in a bigger dataset to assess its generalisation capability and the need to ensure images with optimal views for optimizing the training process.

Thus, the idea is to use texture descriptor in order to mitigate the above-mentioned limitations specifically the MFS, which show to be invariant to lighting and perspective [66]. However, the work made in the context of endoscopic data using this descriptor is scarce. Only M. Häfner *et al.* [21] used this idea for detecting colonic polyps in endoscopic images. Given these motivations, in this thesis algorithmic solutions that bridge the gap between CNN and multifractal descriptors are proposed.

Article	Main goal	Algorithm	Techniques	Nº of images	Nº of pacients	Performance
[68]	Detection of gastric intestinal metaplasia in NBI and M-NBI images	Xception; NASNet; EfficientNet-B4	Cross-validation; Data augmentation; Transfer learning; Grad-CAM method	1880	336	. Sensitivity: 91.9% . Specificity: 86.0% . Accuracy: 88.8%
[19]	Detection of early gastric cancer in ME-NBI images	VGG-19	Transfer learning; Grad-CAM method	1777	295	. Accuracy: 77,0 % . Sensitivity: 79,2 % . Specificity: 74,5 % . Positive Predictive Value (PPV): 77,2 % . NPV: 76,7 %
[65]	Detection of intestinal metaplasia and gastric atrophy by IEE using a deep convolutional neural network	ResNet-50; VGG-16; DenseNet-169; EfficientNet-B4	Transfer learning; Dropout; Early stopping; Data augmentation	6250 images and 98 videos	837	Gastric Atrophy: . Accuracy: 86,9% . Sensitivity: 87,3% . Specificity: 86,1% . PPV: 92,5% . NPV: 77,5% Intestinal metaplasia: . Accuracy: 88,8% . Sensitivity: 90,1% . Specificity: 86,1% . PPV: 92,8% . NPV: 81,6%
[48]	Lesion detection, differential diagnosis, and invasion depth estimation of gastric mucosal lesions	U-Net; CNN	Grand-CAM method	1366	1366	AI-LD model: . Dice similarity coefficient: 0,93 AI-DDx model: . Area Under the Curve (AUC): 0,86 AI-ID model: . AUC: 0,73
[40]	Classification of M-NBI images regarding three classes: chronic gastritis, low grade neoplasia and early gastric cancer.	VGG-16; Inception-V3; ResNet-50; Inception-ResNet-V2; SVM	Transfer learning; Cross-validation	3871	-	. Accuracy: 96% Precision: . CGT: 92,8% . LGN: 90,7% . EGC: 98,9% Recall: . CGT: 91,5% . LGN: 92,2% . EGC: 99,0%
[34]	Detection of gastric intestinal metaplasia in NBI images using a multi-feature fusion model	ResNet-50	Transfer learning; Cross-validation; Attention fusion module; Regularization module	1050	242	. Accuracy: 90,28% . Precision: 89,80% . Recall: 93,16% . F1-score: 91,45% . AUC: 0,949
[43]	Segmentation and detection of early gastric cancer in endoscopic images	ResNet-18; ResNet-34; ResNet-50; GoogLeNet; DenseNet-121; U-Net; SegNet; U-Net++; CE-Net	Transfer learning; ROI extraction; CAM;	4697	387	GAIN-ResNet-50: . Recall: 97,38% . Precision: 99,00% . Accuracy: 98,84% . Specificity: 99,53% . F1-score: 98,18% CA-U-Net: . PA: 83,51% . IOU: 0,64
[35]	Detection of gastric mucosal lesions in M-NBI images	Inception-v3	Data augmentation; Transfer learning	2430	-	. Sensitivity: 91,18 % . Specificity: 90,64 % . PPV: 90,64 % . NPV: 91,18 % . Accuracy: 90,91 %
[4]	Show the largest image and video dataset, HyperKvasir, of the gastrointestinal tract	ResNet-50; ResNet-152; DenseNet-161; Multi-Layer Perceptron (MLP)	Transfer learning;	110079 images and 374 videos	-	Macro Average: . Precision: 63,3% . Recall: 61,5% . F1-score: 61,7% Micro Average: . Precision: 91,0% . Recall: 91,0% . F1-score: 91,0% Matthews Correlation Coefficient (MCC): 90,2%
[71]	Detection of chronic atrophic gastritis in endoscopic images	AlexNet; VGG-19; Inception-V3; ResNet-152; DenseNet-121	Cross-validation; Transfer learning; CAM	5470	1699	. Accuracy: 94,24% . Sensitivity: 94,58% . Specificity: 94,01% . AUC: 0,99
[15]	Detection of atrophic gastritis in real-world endoscopic images from the proximal stomach	VGG-16	Cross-validation; Transfer learning; Data augmentation	270	136	. Accuracy: 92,9% . Balanced accuracy: 93,8% . Sensitivity: 100% . PPV: 85,7% . NPV: 100% . F-score: 93,3% . AUC: 0,981
[36]	Detection of atrophic gastritis and gastric intestinal metaplasia in white light images	TResNet	Data augmentation; Cross-validation; CAM	7037	2741	Atrophic Gastritis: . AUC: 0,98 . Sensitivity: 96,2% . Specificity: 96,4% . Accuracy: 96,4% Gastric Intestinal Metaplasia: . AUC: 0,99 . Sensitivity: 97,9% . Specificity: 97,5% . Accuracy: 97,6%
[23]	Detection of gastric precancerous diseases in endoscopic images	GPDFENet; ResNet-101; ResNet-50; Inception-V3; AlexNet; SVM	Cross-validation	3673	-	. Sensitivity: 93,22% . Specificity: 96,61% . Precision: 93,21% . Accuracy: 93,22%
[17]	Detection of gastric precancerous diseases in endoscopic images	Single Shot MultiBox Detector	-	15880	69 (only the test set)	. Sensitivity: 92,2% . PPV: 30,6%
[18]	Differentiate early gastric cancer from gastritis with ME-NBI images	GoogLeNet	Transfer learning	2828	-	. Accuracy: 85,3% . Sensitivity: 95,4% . Specificity: 71,0% . PPV: 82,3% . NPV: 91,7%
[21]	Classification of colonic polyp in images using texture analysis methods that are based on computing local fractal dimension	MFS-LFD; MRS-LFD; Blob-Adapted (BA)-LFD; 4 new variation of the BA-LFD: BA-Gradient Local Fractal Dimension (GLFD), Blob Shape Adapted (BSA)-LFD, BSA-GLFD and Blob Shape (BS); SIFT; DT-CWT; LBP; Vascularization features; Multiscale Block Local Binary Patterns (MB-LBP); K-Nearest Neighbours (K-NN)	Leave-One-Patient-Out (LOPO); McNemar test; Cross-validation; Wilcoxon rank-sum test	CC-i-Scan database: . 818 NBI dataset: . 908	CC-i-Scan database: . 540	CC-i-Scan database: . BSA-LFD: 79 . MRS-LFD: 79 . SIFT: 79 NBI dataset: . BSA-GLFD: 88.2%
[52]	Classification of histology images using an ensemble model based on handcrafted fractal features and deep learning	ResNet-50; Gliding-box alghorithm	Fractal Dimension; Lacunarity; Percolation; Transfer learning; Cross-validation	UCSB (breast tumours): 58 CR (colourectal tumours):165 NHL (non-Hodgkin lymphoma): 173 LG (liver tissue): 265 LA (liver tissue): 529	-	Accuracy: . NHL: 95,55% . CR: 99,39% . UCSB: 89,66% . LG: 99,62% . LA: 99,62% F-score: . NHL: 0,864 . CR: 0,994 . UCSB: 0,895 . LG: 0,996 . LA: 0,996
[39]	Classification of early gastric cancer in M-NBI images	VGG-16; Inception-V3; Inception-ResNet-V2; LBP; Complete Local Binary Pattern (CLBP); Gabor; GLCM	Data augmentation; Transfer learning; Cross validation	2331	-	. Accuracy: 0,985 . Sensitivity: 0,981 . Specificity: 0,989
[27]	Identify and delineate early gastric cancer in M-NBI images with a CADx system	SVM	GLCM features	127	127	Diagnostic performance: . Accuracy: 96,3% . Precision: 98,3% . Recall: 96,7% . Specificity: 95% Delineation performance: . Accuracy: 73,8% . Precision: 75,3% . Recall: 65,5% . Specificity: 80,8%

Table 3.4: Relevant information about each article regarding the criteria chosen.

Chapter 4

Methods

This chapter outlines the methods used in this project. First, in Section 4.1 the pre-processing procedure implemented is described. Then, in Section 4.2 the normalisation of the data is explained, and in Section 4.3 the augmentation methods applied are presented. Finally, in Section 4.4 and 4.5, a theoretical explanation of the novel texture descriptor named Multi-Fractal Spectrum (MFS) and of the Bilinear Models, respectively, is given.

4.1 Pre-Procesing

The preprocessing procedure started with the removal of unnecessary information for each frame of the image. Namely, two problems were detected: excessive black borders and the presence of text (see the first image of Fig. 4.1). Thus, the ammount of redundancy in the dataset was minimised in order to facilitate the training process.

Firstly, the excessive black borders present in the collected *Esophagogastroduodenoscopy* (EGD) frames were removed. Thus, the minimum lower, upper, right, and left limit where the black borders end were found. The idea was to scan through all the pixels horizontally and vertically in order to detect where the image's RGB value transits from black to another colour. Considering a coordinate system like the one in the middle image of Fig. 4.1, the objective is to find x_1, x_2, y_1 and y_2. To detect if the colour was not black, the Euclidean distance between $(0, 0, 0)$ and the colour channel of every pixel considered was computed, and if was higher than 55 it was defined as not black. Considering that the shape of the original image is $N \times M$, the result image will be $N - (x_2 - x_1) \times M - (y_2 - y_1)$. All the values used as thresholds were obtained by an empirical analysis of this dataset.

Concerning redundant textual information, only the text present in the black triangles was considered because the removal of the text in the central image could distort important regions. This procedure was applied after the prior removal of excessive black borders. The equation of the hypotenuse of each triangle was foundnd. With this it was possible to have the points where the image starts on the left side and equal the pixels that are behind to $(0, 0, 0)$, obtaining the

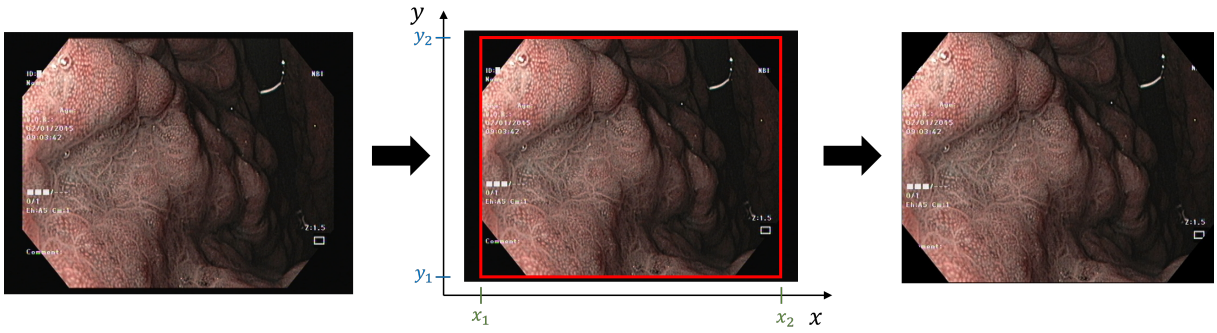


Figure 4.1: Pre-processing procedure. The first image is the original, the second represents the black borders removal procedure where the x_1, x_2, y_1 and y_2 are the values found by the procedure and the red square is delineating the resulting image. The last image is the result of the pre-processing procedure.

region considered all in black. On the right side was the opposite. To compute the equation of the hypotenuse, the criteria to find the two points was the same as in the black borders method, but instead of considering horizontal or vertical regions all the points were considered individually. The upper and lower triangles were assumed to be equal. The preprocessing procedure is represented in Fig. 4.1.

Before applying any of the models, all the images were resized to $224 \times 224 \times 3$.

4.2 Data Normalisation

Initially, the idea was to employ the standardisation method for image normalization across all models. However, the models performance turned out to be highly unsatisfactory with this method. Better results were obtained with the specific normalisation procedure that were used in TensorFlow 2.11.0 for pre-train each model in the ImageNet dataset.

The normalisation methods used for the different models considered were the following:

- ResNet-50: the pixel values were zero-centered and converted from the RGB to the BGR colour space.
- VGG-16: the pixel values were zero-centered.
- DenseNet-121: the pixel values were scaled to the range $[0, 1]$ and each channel was separately normalised.
- Efficientnet-b4: normalisation was not used during pre-training, thus there was no need to normalise the images in our dataset.

4.3 Data Augmentation

Deep Neural Networks (DNN) benefit from large quantity of data. The dataset considered for this thesis contains only 125 images, so data augmentation was applied. The dataset will be explained in the Section 5.1.

The images display significant scale, perspective, occlusion, and colour variations. Therefore, standard methods such as random cropping or colour variations were not considered in order not to compromise the visibility of GIM in the images. Furthermore, since the image is bound by a rectangular frame, it will be affected if a rotation is applied with an angle different from π radians.

Therefore, a simple data augmentation procedure was applied. Each image was replicated one time using a procedure named Fancy Principal Component Analysis (PCA), and then a random method was applied that could result in 5 to 15 augmented images. In this method, each image could be replicated using random horizontal or vertical flips, and also by applying again the Fancy PCA. The procedure is described in algorithm 1, where *AugmentedImages* is the list with all the images obtained.

Algorithm 1 Data Augmentation Procedure

```
Image1 ← FancyPCA(Image)
AugmentedImages ← Image1
for K in {1, 2, 3, 4, 5} do
    a ← Random number in [0; 1]
    b ← Random number in [0; 1]
    if a > 0.5 then
        Image ← FlipLeftRight(Image)
        AugmentedImages ← Image
    else if b > 0.5 then
        Image ← FlipUpDown(Image)
        AugmentedImages ← Image
    end if
    Image ← FancyPCA(Image)
    AugmentedImages ← Image
end for
```

Fancy PCA was applied following the procedure described in [30]. This method consists in changing the values of the RGB channels. It starts with computing the PCA on the RGB pixel values for each image, obtaining a 3×3 covariance matrix. Considering an image of shape $N \times M$ and the pixel values $p(i,j) = (R_{(i,j)}, G_{(i,j)}, B_{(i,j)})$, where $0 < i < N$ and $0 < j < M$, the output pixel values $p'(i,j)$ are obtained with the following:

$$p'(i,j) = (R_{(i,j)}, G_{(i,j)}, B_{(i,j)})^{\mathrm{T}} + (v_1, v_2, v_3)(\alpha_1\lambda_1, \alpha_2\lambda_2, \alpha_3\lambda_3)^{\mathrm{T}}, \tag{4.1}$$

where $\mathbf{v}_k$ and the λ_k are the eigenvector and the eigenvalue, respectively, of the covariance matrix

obtained, and α_k is a random variable obtained from three independent Gaussian distributions with mean 0 and standard deviation of 0.1. A different α_k is computed for each image.

4.4 Multi-Fractal Spectrum

In Computer Vision, most of the texture descriptors used are not invariant to scale and illumination changes in images [66]. However, the importance of finding an invariant one was already recognized and some researchers have already proposed some approaches for tackling this [7]. In this project, a texture descriptor proposed by Yong Xu *et al.* in [66] is used, which is theoretically invariant to these transformations.

This descriptor is an extension of the fractal dimension, which is the key quantity to describe perfectly self-similar fractal objects. Fractal geometry can describe irregular and complex objects that are indescribable by classical Euclidean geometry [21], and its central concept is the self-similarity. A self-similar object can be divided into N self-similar parts scaled by a factor δ, where $\delta = \frac{1}{\sqrt[D]{N}}$ and D is the dimension of the space [3]. This concept is described in Fig. 4.2. Thus, the similarity dimension which is the same as the fractal dimension, can be obtained with the following:

$$D = \frac{\log(N)}{\log(\frac{1}{\delta})} \tag{4.2}$$

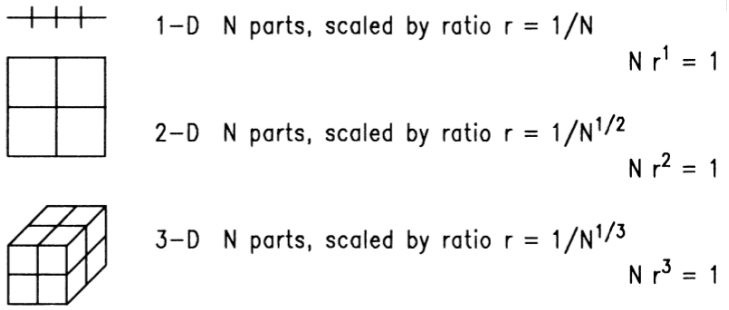


Figure 4.2: Divison of self-similar objects into N self-similar parts of different dimension D. Image source: [3]

The fractal dimension is not always an integer. For example in Fig. 4.3, one can divide a square in 16 similar parts so $N = 16$ and $\delta = \frac{1}{4}$. Then, $D = -\frac{\log(16)}{\log(4)} = 2$. But, if the Koch curve (see Fig. 4.4) is used the result is not an integer. The Koch curve starts as equilateral triangle in which each side is divided in 4 similar parts, by a factor of $\delta = \frac{1}{3}$. Then, this is applied successive times. The fractal dimension of this object will be given by $D = -\frac{\log(4)}{\log(\frac{1}{3})} \simeq 1.26$. This value show the irregularity of this object, and that this curve "fills more of a space than a line ($D = 1$), but less than a Euclidean area of the plane ($D = 2$)" [3]. Fractal dimension is based on this idea, since it gives information of how the curves change, with lower values to more "line-like" curves and higher values to curves that "occupy" more space.

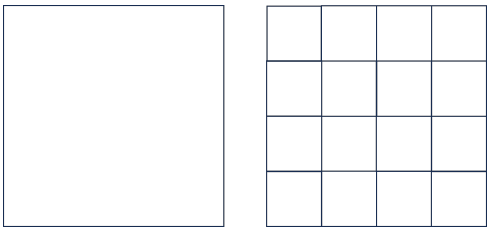


Figure 4.3: Division of a square in $N = 16$ self-similar parts. The fractal dimension is $D = 2$.

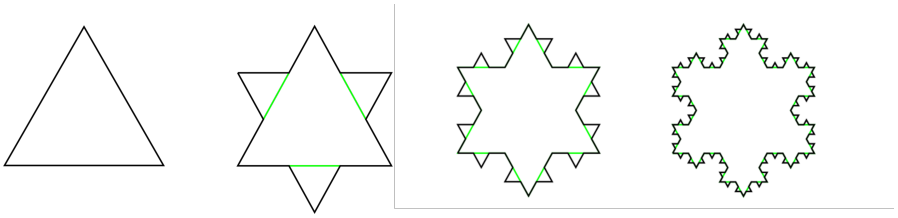


Figure 4.4: First four interactions of the Koch curve. The fractal dimension is $D = 1.26$. Image source: https://en.wikipedia.org/wiki/Koch_snowflake

This concept can be extended to statistically self-similar objects like the coastline represented in Fig. 4.5. The idea is as basic as thinking that if we walk along the beach is longer than driving along the corresponding coast highway. Thus, if a ruler δ is used to measure this dimension the result will be the ruler multiplied by the times that it was used (for example, the number of steps that a person gives when is walking around the beach). Fig. 4.5 shows that as the stick decreases the length of the coastline increases.

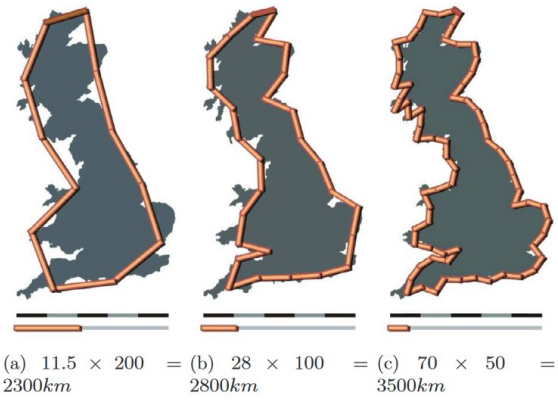


Figure 4.5: Example of the concept of fractal dimension, where delta is represented by the stick. As the length of the stick decreases the dimension of the coastline increases. Image source: [21]

The coastline example gives an intuition that a large scale can be insufficient to capture all the details of an object. The fractal dimension characterises how these details change with the length of the scales. Following the definition of Young Xu *el al.* [66], considering objects in $\mathbb{R}^2$ and an irregular point set E defined in $\mathbb{R}^2$, the fractal dimension of E is

$$\dim(E) = \lim_{\delta \to 0} \frac{\log N(\delta, E)}{-\log \delta} \tag{4.3}$$

where $N(\delta, E)$ is the smallest number of sets of diameter less than δ that cover E. In practice, the space is usually divided into boxes of size δ and the result is the number of boxes occupied by the point set.

Young Xu *el al.* [66], conclude that a global texture descriptor like the fractal dimension was not sufficient. Thus, they proposed the novel texture descriptor based on multifractal analysis, the MFS, that uses local and global information. In this project, the definition proposed in [66] is used since it is proved that it is invariant under any spatial *bi-Lipschitz* transformation. This descriptor was defined considering a point set categorisation function regarding some criteria, that result in several sets. Then, the fractal dimension of each set is computed and the MFS is a vector composed by the results.

The point set categorisation function was given by a local density function. Consider an image I defined over $\Omega \subseteq \mathbb{R}^2$ and let μ be a measure over Ω so that $\mu(\mathbf{x}, r) = kr^{\hat{\beta}(I,\mathbf{x})}$ for $\mathbf{x} \in \Omega$, where $\hat{\beta}(I, \mathbf{x}) \in \mathbb{R}$ is a density function and $k \in \mathbb{R}$, the local density function is defined as:

$$\hat{\beta}(I, \mathbf{x}) = \lim_{r \to 0} \frac{\log \mu\big(B(I(\mathbf{x}), r)\big)}{\log r}, \tag{4.4}$$

where $B(I(\mathbf{x}), r)$ is a closed disk of length r around coordinate x in I. Furthermore, for each $\alpha \in \mathbb{R}$, the set of all points with the same local density function is defined as

$$E_\alpha = \left\{ x \in \Omega : \hat{\beta}(I, \mathbf{x}) = \alpha \right\}. \tag{4.5}$$

Finally, the MFS is obtained computing the fractal dimension of each set:

$$MFS(I) = \{\dim(E_\alpha) : \alpha \in \mathbb{R}\}, \tag{4.6}$$

where $dim(E_\alpha)$ is obtained using equation (4.3). Regarding the equation (4.6), the measure function μ is still not defined. We continue following the definition proposed in [66], where three different measure functions were considered. The first one was:

$$\mu_1\big(B(I(\mathbf{x}), r)\big) = \int_{B(I(\mathbf{x}),r)} G_r * I(\mathbf{x})\, d\mathbf{x}, \tag{4.7}$$

where $*$ is representing a 2D convolution operator and G_r is a Gaussian smoothing kernel with variance r. The second function was:

$$\mu_2\big(B(I(\mathbf{x}), r)\big) = \int_{B(I(\mathbf{x}),r)} \sum_{i=1}^{4} g_i(G_r * I(\mathbf{x}))\, d\mathbf{x}, \tag{4.8}$$

where g_1, g_2, g_3 and g_4 are four differential operators along the vertical, horizontal, diagonal, and anti-diagonal directions, respectively. Finally, the third one was the sum of the Laplacian filter of the image inside $B(I(\mathbf{x}), r)$ that is given by:

$$\mu_3\big(B(I(\mathbf{x}), r)\big) = \int_{B(I(\mathbf{x}),r)} |\nabla^2(G_r * I(\mathbf{x}))| d\mathbf{x}. \tag{4.9}$$

This filter is used to identify changes in image intensity and detect edges.

4.4.1 Application of the Multi-Fractal Spectrum

Initially, the local density function is calculated for each pixel by adjusting the slope of the line in the scaling plot $\log \mu\big(B(I(\mathbf{x}), r)\big)$ *vs.* $\log r$. Then, $\{\alpha_i : i = 1, ..., N\}$ is selected from a discrete sample from the interval $[1, 4]$ and E_{α_i} is defined through all the pixels that have a local density function close to a certain α_i considering a given threshold. With each E_{α_i}, it is possible to obtain the MFS by computing the fractal dimension, using the same process used for the local density function that is adjusting the slope of the scaling plot of the $\log N(\delta, E)$ *vs.* $\log -\delta$.

Regarding the texture captured in the endoscopic images with Narrow Band Imaging (NBI) highlighted in Section 2.1, the idea is to use this descriptor as a feature vector that helps the model learn how to detect Gastric Intestinal Metaplasia (GIM), through its texture pattern visible in the images (see Fig. 2.2). However, note that this texture descriptor is commonly used in images in which just one texture pattern is visible [66]. Yet, it is worth noting that endoscopic images display drastically more complicated scenes, where multiple textural patterns might be present, such as those of bubbles, blood, or of the endoscopic tube (see Fig. 4.6). Furthermore, GIM commonly appears in small sections of the image (see image B from Fig. 2.2). To account for this, we compute the MFS for several non-overlapping patches of the input image obtaining a more accurate description of the fragments with a lesion.

In order to facilitate computation, the spatial resolution of the dataset was normalised to $1078 \times 1351 \times 3$ so that the images are partionable in 7×7 non-overlapping patches of resolution $154 \times 193 \times 3$ (see Fig. 4.7).

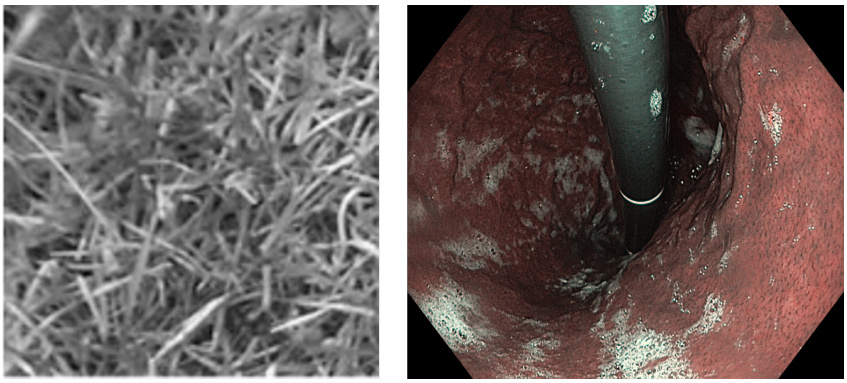


Figure 4.6: Examples of two images to show the difference between the quantity of texture patterns present in each. The image with the grass texture was obtained from [66].

4.5 Bilinear Models

Detecting GIM in endoscopic images can be considered a fine-grained recognition task, since the difference between a healthy and an unhealthy mucosa is minimal. The difference is based on minor details that commonly appear in a small region which results in very similar images like

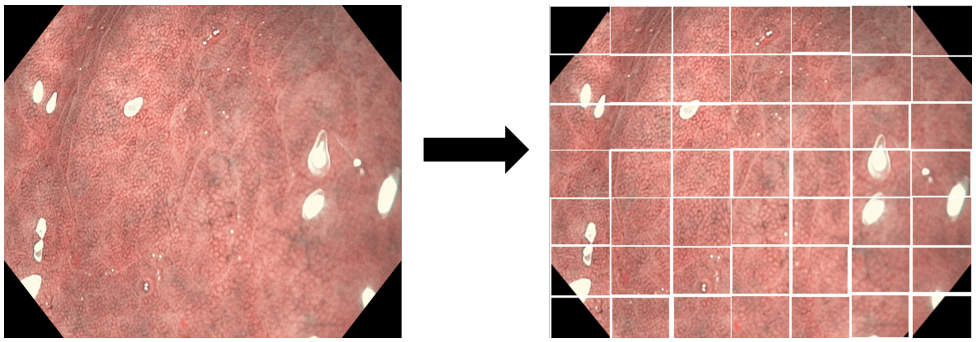


Figure 4.7: Example of the division of the image into patches.

the ones present in Fig. 4.8. Tsung-Yu Lin *el al.* [37] proposed a bilinear model to solve this type of tasks, obtaining very promising results in different experiences tested.

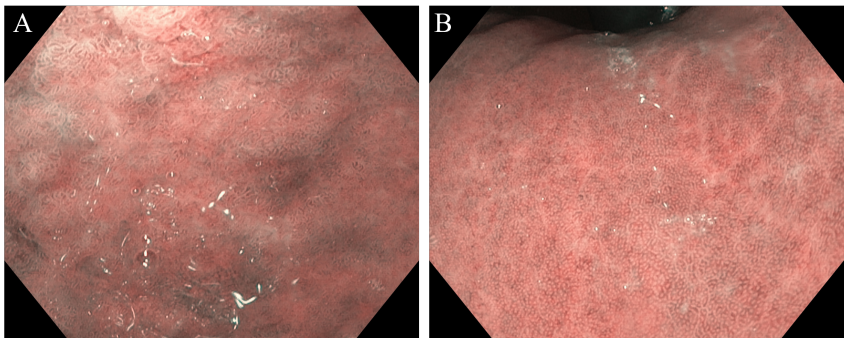


Figure 4.8: Examples of two endoscopic images that appear to be similar but in (A) the mucosa is with GIM and in (B) is normal.

A bilinear model $\mathcal{B}$ can be defined by a quadruple as $\mathcal{B} = (f_a, f_b, \mathcal{P}, \mathcal{C})$, where $f_a : \mathcal{L} \times \mathcal{I} \rightarrow \mathbb{R}^{k \times M}$ and $f_b : \mathcal{L} \times \mathcal{I} \rightarrow \mathbb{R}^{k \times N}$ are feature functions that takes as input an image $\mathcal{I}$ and a location $\mathcal{L}$ and that gives as an output feature vectors, $\mathcal{P}$ is a pooling function and $\mathcal{C}$ is a classification function. An example of a feature function is a Convolutional Neural Networks (CNN).

Firstly, the outputs of the f_a and f_b are combined at each location through the computation of the matrix outer product:

$$\mathbf{x} = \mathrm{A}^T \mathrm{B}, \tag{4.10}$$

where $\mathrm{A}^T = [f_a(1, I), f_a(2, I), ..., f_a(|\mathcal{L}|, I)]$, $\mathrm{B} = [f_b(1, I); f_b(2, I); ...; f_b(|\mathcal{L}|, I)]$ and $I \in \mathcal{I}$. With the matrix outer product the elements of the output of f_a are conditioned on the output of f_b. Thus, the outer product effectively captures all the pairwise interactions between the two feature functions (see Fig. 4.9).

Then, the pooling function is used to aggregate the result. Considering that the sum pooling is used the result obtained is the following:

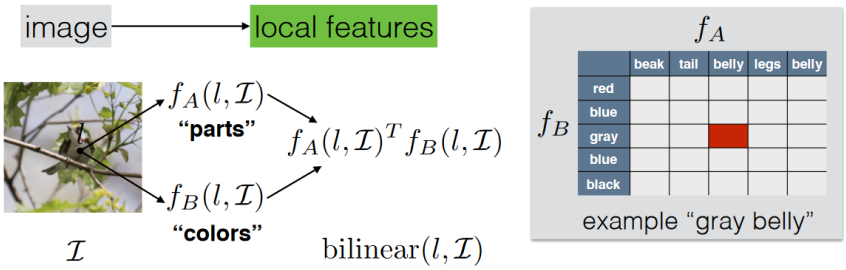


Figure 4.9: Scheme to represent the outer product capturing the pairwise interactions between the feature functions f_a and f_b. In the example the f_a captures the parts of the body of the bird while f_b retains the colours. The result of the outer product is the combination of these two local features. Image source: [38]

$$\mathcal{P}(\mathbf{x}) = \sum_{l \in \mathcal{L}} f_a(l, I)^{\mathrm{T}} f_b(l, I) \tag{4.11}$$

$\mathcal{P}(\mathbf{x})$ is considered an orderless representation since the pooling function ignores the location of the features. The first dimension k of the feature functions has to be the same, and the shape of the resulting pooling function is $M \times N$. Note that $\mathbf{x} = \mathcal{P}(\mathbf{x})$ since the inner product of any two (multipliable) matrices is the same as the summation of the outer product of the columns of the matrix in the right and the rows of the matrix in the left. Thus, $\mathrm{A^TB}$ can be described in the following way:

$$\mathrm{A^TB} = \sum_{\updownarrow \in \mathcal{L}} f_a(l, I)^{\mathrm{T}} f_b(l, I) \tag{4.12}$$

Following the procedure described in [37], after the pooling the output is reshaped to $MN \times 1$ and normalised in the following way:

$$y = sign(\mathcal{P}(\mathbf{x}))\sqrt{\mathcal{P}(\mathbf{x})} \tag{4.13}$$

$$z = \frac{y}{\|y\|_2}. \tag{4.14}$$

Then, the predicted class is obtained using a classification function $\mathcal{C}$ applied to z.

Chapter 5

Experimental Methodology

In this chapter, we provide an explanation of the experiments conducted, starting with the description of the materials used (Section 5.1), and the definitions of the performance metrics applied to evaluate the proposed models (Section 5.2). Then, in Section 5.3 the procedure to apply the baseline models is described, and, finally, in Section 5.4 the proposed approaches are explained.

5.1 Materials

For the experiences conducted on this project, a dataset collected at the Gastroenterology department of Instituto Português de Oncologia, Porto (IPO-Porto) was used. The dataset is composed of 883 high-resolution images of three distinct modalities: White Light Imaging (WLI), Narrow Band Imaging (NBI) and Magnifying Endoscopy with Narrow-Band Imaging (ME-NBI). The images represent 4 different classes: normal, Gastric Intestinal Metaplasia (GIM), dysplasia and atrophic gastritis (see Fig.5.1 and Table 5.1).

Classes	Nº of images
Normal	808
GIM	64
Dysplasia/Carcinoma	10
Atrophic Gastritis	1

Table 5.1: Number of images for each class of the dataset.

After a filtering procedure made by an endoscopist regarding incorrect diagnosis of GIM, low resolution, and frames captured in WLI, the final version of the dataset was obtained with 125 high-quality NBI/ME-NBI images, 65 classified as normal (- class) and 60 as GIM (+ class). Notice that some foam, bubbles, bile, blood, and some pathological findings such as polyps can be found in this dataset, since the frames were captured in a standard clinical practice (see Fig. 5.2).

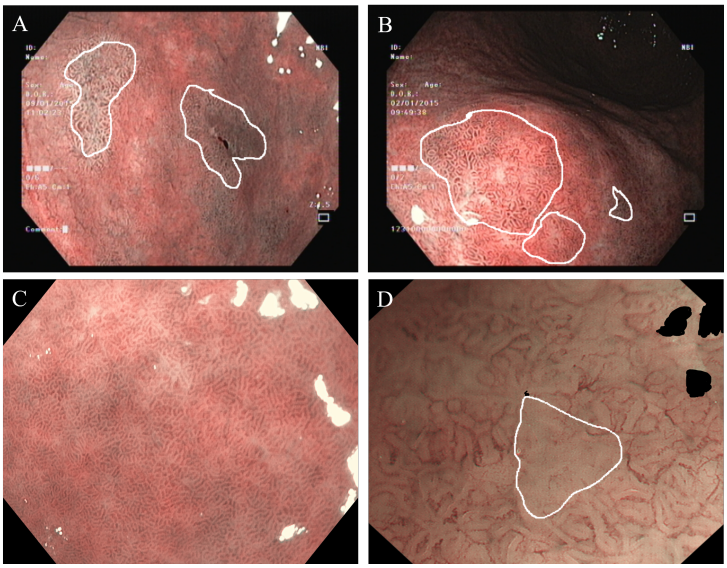


Figure 5.1: Representative images of each class with the lesion outlined in white. A: Dysplasia/Carcinoma; B: Intestinal Metaplasia; C: Normal; D: Atrophy

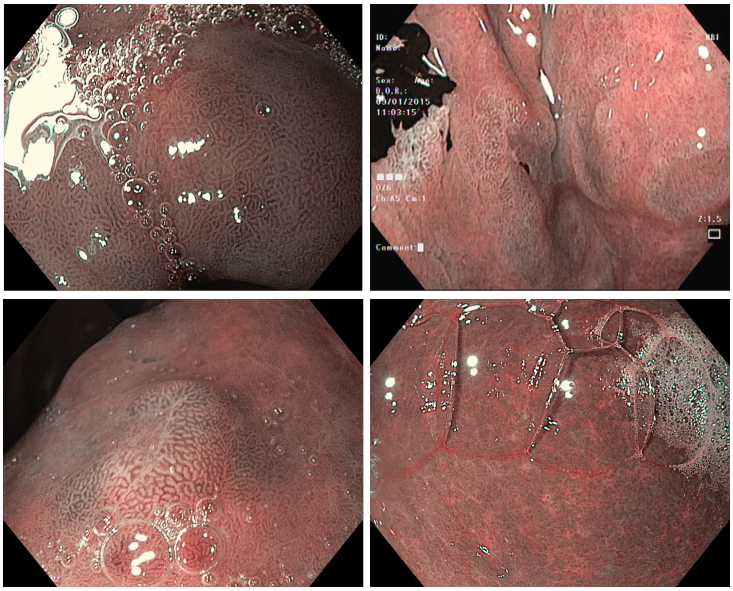


Figure 5.2: Examples of images present in the dataset corrupted with bubbles, blood, foam, and polyps

5.2 Performance Metrics

In order to evaluate the performance of the models, the dataset was divided into three different folds: train, validation, and test. The first one was used to train the model and the second one was used to validate the performance while the model was training. The last fold is composed of unseen data used to compute the evaluation metrics. This divison was made using a method named k-fold cross-validation. In this method, the model is trained in k different folds, and a

different division of the data is made for each. So, in each fold a percentage of the data is used for training and another for testing. This repeats k times, so the same data that in one fold can be used for training in another is used for validating the model. The evaluation metrics presented for each experience are the average of the metrics obtained for each of the considered folds.

The results obtained in a binary classification problem can be represented by a confusion matrix (Fig. 5.3). The quantities represented in this matrix are then used to compute the evaluation metrics. The True Positive (TP) and True Negative (TN) represent the number of correctly classified samples in the positive and negative classes, respectively. False Positive (FP) and False Negative (FN) are the number of samples misclassified that belong to the positive and negative classes.

		True Label	
		+	−
Predicted label	+	TP	FP
	−	FN	TN

Figure 5.3: Confusion matrix of a binary classification problem. The (+) represents the positive class and the (−) the negative class.

A set of 8 representative metrics for this classification task were selected to understand the capability of the methods developed.

- Accuracy - the ratio between the number of samples correctly classified and the total number of samples. If the accuracy is 1 the model classified all the samples correctly.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5.1}$$

- Specificity - the ratio between correctly classified negative samples and all the samples of the negative class. Represents how good the model is at predicting the negative class. It is also named as the True Negative Rate (TNR).

$$Specificity = \frac{TN}{FP + TN} \tag{5.2}$$

- Sensitivity - also known as the True Positive Rate (TPR), is similar to the specificity, but for the positive class. Thus, it is the ratio between the correctly classified positive samples and all the positive samples.

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.3}$$

- Positive Predictive Value (PPV) - is the ratio between the correctly classified positive samples and all the samples classified as positive. If it is 1 all the samples classified as positive were correct ($FP = 0$).

$$PPV = \frac{TP}{TP + FP} \tag{5.4}$$

- Negative Predictive Value (NPV) - similar to the PPV, but for the negative class. It is computed by the ratio between the correctly classified negative samples and all the samples classified as negative.

$$NPV = \frac{TN}{TN + FN} \tag{5.5}$$

- Area Under the Curve (AUC) - this metric is the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve represents the plot of the TPR *vs.* False Positive Rate (FPR) ($1 - \text{TPR} = \text{FPR}$), for a given threshold between 0 and 1. In a binary classification problem a sample belongs to the positive class if the predicted probability is higher than a given threshold, otherwise is classified as negative. Furthermore, if the model classifies as positive with a higher probability, it means that there is high 'confidence' regarding that prediction. Regarding this, the perfect and the worst scenarios for the ROC curve are represented in Fig. 5.4. So, AUC is an approximation of the area under the ROC curve, and the higher the values for the AUC the better.

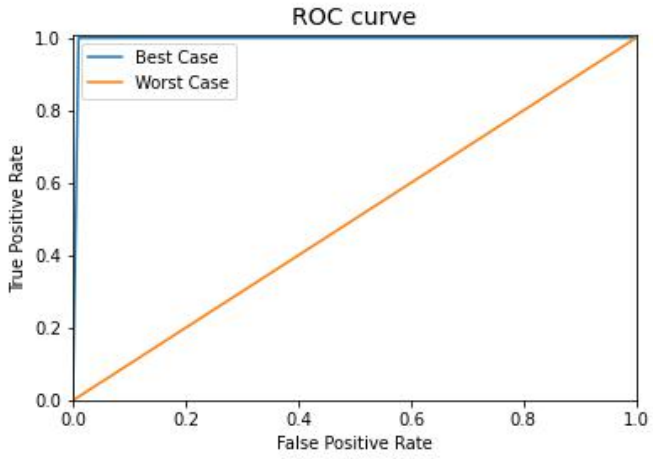


Figure 5.4: Representation of a ROC curve. The orange line represents the worst scenario obtained, while the blue line represents the best.

- Positive Likelihood Ratio (LR+) - this metric refers to the comparison of the probability that someone with the disease has a positive diagnosis with the probability of someone without the disease having the same prediction. Considering $D+$ as having the disease and $D-$ as not having, LR+ can be computed as:

$$LR+ = \frac{P(+|D+)}{P(+|D-)} = \frac{Sensitivity}{1 - Specificity} \tag{5.6}$$

  Regarding this if $LR+ > 1$ the model is more likely to predict a positive label if the patient has the disease, while a $LR+ < 1$ is the opposite [51]. Thus, the higher the $LR+$, the better the model.

- Positive Likelihood Ratio (LR-) - the definition of this metric is similar to the LR+, with the difference of the class considered. Thus, $LR-$ is the ratio between the probability of

someone having the disease and the model prediction being negative and the probability of someone not having the disease and obtaining a negative prediction:

$$LR+ = \frac{P(-|D+)}{P(-|D-)} = \frac{1 - Sensitivity}{Specificity} \tag{5.7}$$

The interpretation of the values is the opposite of the LR+, since the lower the $LR-$, the better the model. If $LR- < 1$ it is more likely that someone who does not have the disease obtains a negative prediction.

5.3 Baseline Models

In order to create the baseline models, state-of-the-art results were analysed to see the best options. Concerning the analysis made in sub-section 3.2.1, there are mainly 4 articles that focus on detecting GIM in endoscopic images. The work of T. Yan *et al.* [68] that used an EfficientNetB4 with some modifications, the work of M. Xu *et al.*[65] that applied a VGG-16 in their proposed system, the work of H. Li *et al.*[34] implemented a ResNet in their novel multi-feature fusion model, and, the work of N. Lin *et al.* [36] that applied a TresNet.

All these models were considered to generate baseline results over the dataset presented in Section 5.1. The DenseNet was also used since it was implemented to detect atrophic gastritis in the work of Y. Zhang *et al.* [71]. Among these, only the TresNet model was not considered since one example per type of architecture seems to be sufficient and the one considered from the 'family of ResNet models' was the ResNet-50.

For each network, the architecture pre-trained in the ImageNet dataset was implemented changing the classifier to Multi-Layer Perceptron (MLP) that was composed of two dense layers with a Rectified Linear Unit (ReLU) activation function, two dropout layers, and an output layer with a single neuron composed of a sigmoid activation function. All the remaining layers were frozen, and Global Average Pooling (GAP) was applied to the results. The procedure is presented in Fig. 5.5.

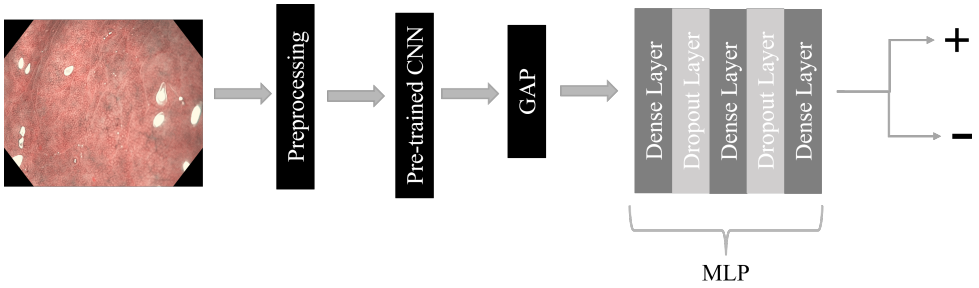


Figure 5.5: Procedure followed to train the baseline models.

Regarding the structure described in Fig. 5.5, some parameters can be defined before training: the number of neurons of the dense layers and the probabilities of the dropout layers. The choice

of each parameter can influence the performance of the model and there is no simple criterion to select the best one for a specific task. Therefore, an architectural grid-search was performed to find the best possible configuration.

Considering the hyper-parameter set $\mathbf{\Psi} = \{\alpha, d_1, d_2, p_1, p_2\}$. Let the learning rate be $\alpha \in \{1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}\}$, the dropout probabilities $p_1, p_2 \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, the number of neurons of the first dense layer $d_1 \in \{128, 256, 512\}$ and the number of neurons in the second dense layer $d_2 \in \{64, 128, 256\}$.

$\mathbf{\Psi}$ was searched for all of the four architectures. The images were pre-processed according to the procedure described in Section 4.1 and divided into a training set with 89 samples $(43+, 46-)$, a validation set with 23 samples $(11+, 12-)$, and a test set containing 13 samples $(6+, 7-)$. The training images were augmented using the procedure described in Section 4.3, and the number of training images increased to 1087 samples $(521+, 566-)$. The models were trained for 50 epochs with early stopping considering the validation loss. The loss function used was the binary cross-entropy and Adam was chosen as the optimiser. The configuration with the best AUC was selected for each architecture.

A stratified 5-fold cross-validation was then applied to assess the performance of the models. For each fold from the 125 samples, 100 samples $(48+, 52-)$ were for the train set, 12 samples $(6+, 6-)$ were for the validation set and the 13 remaining ones $(6+, 7-)$ were for the test set. The images of the train set were augmented, obtaining on average 1214 samples $(582+, 632-)$. The evaluation metrics were computed using the test set, and the best architecture for the proposed task was the ResNet-50 as we can see in Table 6.2.

5.4 Fractal Bilinear DNNs

In this thesis, we introduce two architectural approaches that leverage the Multi-Fractal Spectrum (MFS) as a texture descriptor. The idea is to apply the MFS as a tool to obtain a texture description of the images helping the Convolutional Neural Networks (CNN) improve its performance.

Following Young Xu *et al.* [67] a bilinear pooling module was applied in order to combine the convolutional and multifractal features within the same model. Two separate approaches $\mathcal{B}_1$ and $\mathcal{B}_2$ were proposed.

Firstly, the patch-wise MFS is computed as described in Section 4.1. Specifically, it was defined that $r, \delta \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ and m was set to 26 (number of bins to partition the interval $[1, 4]$).

Given our experimental results in Section 6.1, the VGG-16 and ResNet-50 were selected since they were the two CNN that scored the highest AUC. Using the notation defined in the section 4.5, the model $\mathcal{B}_1$ (Fig. 5.6) was composed by the outputs of the GAP of a pre-trained VGG-16 or ResNet-50 as the feature function f_a and the maximum response over μ_1, μ_2 and

μ_3 as f_b for each of the patch-wise MFS feature vectors. The pooling function used was the sum-pooling. The classification function $\mathcal{C}$ was the MLP defined in the section 5.3 (see Fig. 5.5). Since, GAP is used in the embeddings of the networks the VGG-16 outputs have a shape of 1×512, while with the ResNet-50 the shape is 1×2048. Each patch-wise MFS has shape $1 \times 26 \times 3$. Using the notation defined in Section 4.5, for the VGG-16 approach $f_a : 1 \times I \rightarrow \mathbb{R}^{1 \times 512}$ and the MFS was reshaped to 1×78, so $f_b : 1 \times I \rightarrow \mathbb{R}^{1 \times 78}$. Thus, the shape of $\mathrm{A}^{\mathrm{T}}\mathrm{B}$ was 512×78. Regarding the ResNet-50 approach, the shape of the obtained matrix was 2048×78.

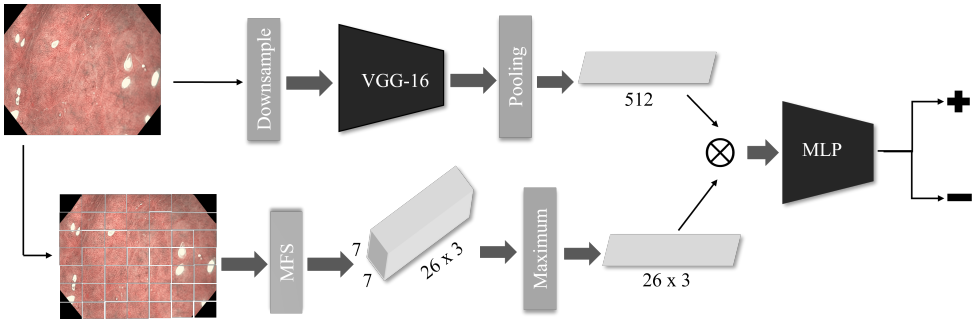


Figure 5.6: Structure of the model $\mathcal{B}_1$ with the VGG-16.

The model $\mathcal{B}_2$ (Fig. 5.7) was different since it considered f_a as the direct embeddings of the pre-trained VGG-16/ResNet-50 without applying a pooling function and f_b as the response over μ_1, μ_2 and μ_3 for the 49 patches. The pooling function used was the sum-pooling. The embeddings of the VGG-16 have a shape of $7 \times 7 \times 512$ and the resulting MFS was $49 \times 26 \times 3$. To compute the outer product the output of f_a was reshaped to 49×512. A similar reshaping process was applied to f_b obtaining a matrix of 49×78. Considering the notation defined in the Section 4.5, $f_a : 1 \times I \rightarrow \mathbb{R}^{49 \times 512}$ and $f_b : 1 \times I \rightarrow \mathbb{R}^{49 \times 78}$. Thus, the shape $\mathrm{A}^{\mathrm{T}}\mathrm{B}$ was 512×78. For the ResNet-50 the procedure was similar, however, the resulted matrix had a shape of 78×2048, since the embeddings of the ResNet-50 have a shape of $7 \times 7 \times 2048$.

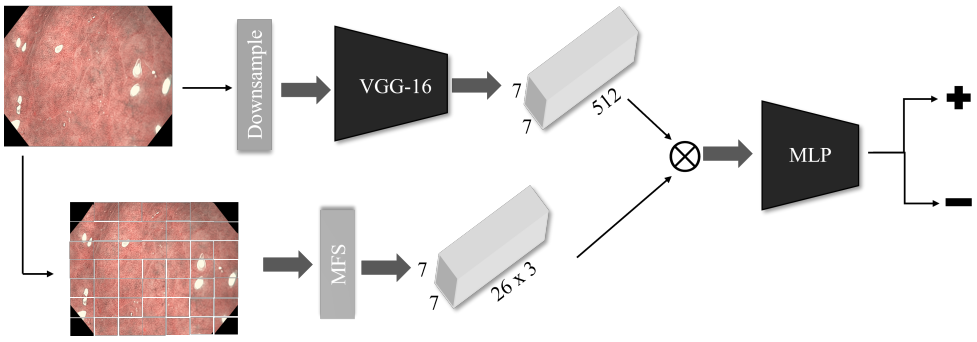


Figure 5.7: Structure of the model $\mathcal{B}_2$ with VGG-16.

The computation of MFS in the two models is based on two different ideas. While in $\mathcal{B}_1$ the maximum response is used to capture the patches where a high response should be associated with GIM in that selection of the image, $\mathcal{B}_2$ includes additional local information concerning the patch-wise pair interaction between the CNN and the MFS feature vectors.

The same folds used for the baseline experiences were used to assess the performance of the proposed approaches. The same procedure used to train the models described in Section 4.5 was applied, but for these models, the computation of the MFS was given as an input to the model combined with the images.

Chapter 6

Results and Discussion

In this chapter, the results obtained with the baseline models (Section 6.1) and the Fractal bilinear Deep Neural Networks (DNN) (Section 6.2) are presented. A comparison and discussion of the results is also given.

6.1 Baseline Models

With the four architectures considered, a grid search was conducted to determine the optimal hyperparameters. The one with the highest value Area Under the Curve (AUC) was selected. In order to obtain the best configuration, a visualization like the one present in Fig. 6.1 was used for each architecture. This was obtained using a TensorFlow visualization toolkit named TensorBoard. The resulting configurations are presented in Table 6.1.

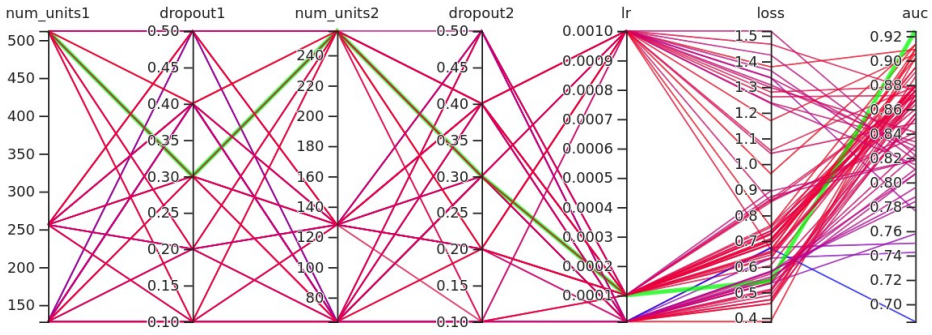


Figure 6.1: Results obtained with ResNet-50. The best configuration is highlighted by the green line.

In order to evaluate the performance of the models a 5-fold cross-validation was implemented to test the performance of each configuration. For the Accuracy, AUC, Sensitivity, Specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) the average for the 5-folds was used to compute 95% confidence intervals with non-parametric Bootstrap (n = 15000) applying the percentile method (see Table. 6.2).

	d_1	p_1	d_2	p_2	α
ResNet-50	512	0.3	256	0.3	1×10^{-4}
VGG-16	256	0.2	128	0.1	1×10^{-5}
DenseNet-121	128	0.3	64	0.5	1×10^{-5}
EfficientNet-b4	512	0.5	256	0.2	1×10^{-5}

Table 6.1: Values of the parameters obtained for the best configuration for each architecture. d_i is the number of neurons in the ith layer, p_i is the dropout probabilities and α is the learning rate.

Given the limited sample size in each of the folds, bootstrap was applied to obtain more robust values of the metrics For each metric, 5 values were obtained regarding the different folds. From this sample, using the bootstrap method, n samples with replacement were gathered and the mean for each was computed. This procedure resulted in a vector with n mean values. Then, the percentile method was applied to obtain the sampling distribution. An interval with 95% confidence was considered, so the significance level α was 0.05. Thus, to obtain the confidence interval $\frac{\alpha}{2} = 0.025$ and $1 - \frac{\alpha}{2} = 0.975$ percentiles were selected from the distribution.

	Accuracy	AUC	Sensitivity	Specificity	PPV	NPV
ResNet-50	**.815 (.723-.908)**	**.812 (.714-.910)**	**.768 (.533-.967)**	**.857 (.743-.971)**	**.854 (.738-.971)**	**.845 (.715-.975)**
VGG-16	.738 (.615-.862)	.738 (.614-.857)	.700 (.600-.800)	.771 (.629-.914)	.741 (.581-.893)	.745 (.645-.843)
DenseNet-121	.738 (.615-.862)	.731 (.607-.855)	.634 (.467-.800)	.830 (.629-.971)	.801 (.633-.967)	.726 (.620-.833)
EfficientNet-b4	.739 (.692-.800)	.731 (.681-.798)	.633 (.467-.767)	.829 (.743-.914)	.794 (.693-.900)	.735 (.675-.800)

Table 6.2: 5-fold cross validation mean metric estimates with 95% confidence interval from a bootstrapped set ($n = 15000$), using the baseline models. The values in bold represent the highest value for each metric.

To assess the inter-fold diagnostic variability for each fold the Positive Likelihood Ratio (LR+) and Positive Likelihood Ratio (LR-) were computed for each fold (see Table. 6.3).

DNN backbone	Likelihood Ratio	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± St.dev
ResNet-50	LR+	83.300	2.332	33.300	3.497	6.993	25.884 ± 30.869
	LR-	.167	.466	.667	0	0	**.260 ± .265**
VGG-16	LR+	5.825	1.555	83.300	1.166	4.664	19.302 ± 32.049
	LR-	.195	.583	.167	.876	.389	.442 ± .264
DenseNet-121	LR+	2.329	1.168	50.000	5.825	83.300	**28.524 ± 32.898**
	LR-	.778	.776	.500	.195	.167	.483 ± .267
EfficientNet-B4	LR+	4.664	2.332	33.300	2.332	5.825	9.691 ± 11.882
	LR-	.389	.466	.667	.466	.195	.436 ± .152

Table 6.3: The LR+ and LR- of each baseline model on each fold.

Regarding the results obtained presented in Table 6.2, it is clear that the ResNet-50 outperforms the other models. On the other hand, the DenseNet-121, EfficientNet-B4, and the VGG-16 showed similar results, with the VGG-16 obtaining the best AUC. The difference between Sensitivity and Specificity shows that all the models seem to be better at predicting the negative class.

Furthermore, the range of the confidence intervals and the higher value for the standard deviation (sometimes higher than the mean) demonstrates that there is a high inter-fold variability in each model. A closer inspection of the cross-validation results reveals that fold 2 was the most difficult since it has the lowest LR+ and the highest LR- (by a factor of around 0.5 for all models). The higher mean of the LR+ was obtained with DenseNet-121 and the lower mean of LR- was observed from the ResNet-50 (see Table 6.3)

In order to obtain a clear conclusion, the Wilcoxon signed-rank test [63] was applied to determine if there exists a statistically significant difference between the baseline models across various metrics. While Student's t-test [28] is commonly used for the comparison of two models, it assumes a Gaussian distribution within the sample data. However, the small sample size of only 5 observations made it impossible to assume that it follows the required distribution. The Wilcoxon signed-rank test is the non-parametric alternative to the t-test, requiring fewer data assumptions, including the Gaussian distribution assumption. Regarding this, the Wilcoxon signed-rank was applied with a significant level $\alpha = 0.05$, and the null hypothesis was if there was a difference between the mean of each sample. This hypothesis was rejected if the p-value was higher than the significance level α. The results are represented in Fig. 6.2.

The results obtained were not conclusive. For all the metrics, the differences between the models were not significant. However, these results can be influenced by three hypothesis. Firstly, the high inter-fold variability observed in the models indicates their inability to generalise effectively within this dataset (possibly due to the small size of the training set). Secondly, the partitioning of the dataset may not yield identically distributed sets, which results in significant variations in the performance metrics that would not occur with different distributions. Finally, assuming that the sets are identically distributed and the models are capable of estimating their distributions, the reduced sample size of the test sets and the random partition of the data on the cross-validation could introduce a bias that accentuates the limitations of the networks in certain folds.

Given that the ResNet-50 and VGG-16 models achieved the highest values for the AUC, they were chosen as the two baseline models for the remainder of the experiments.

6.2 Fractal Bilinear DNNs

The proposed approaches were evaluated in a similar fashion to the baseline models (see Tables 6.4 and 6.5)

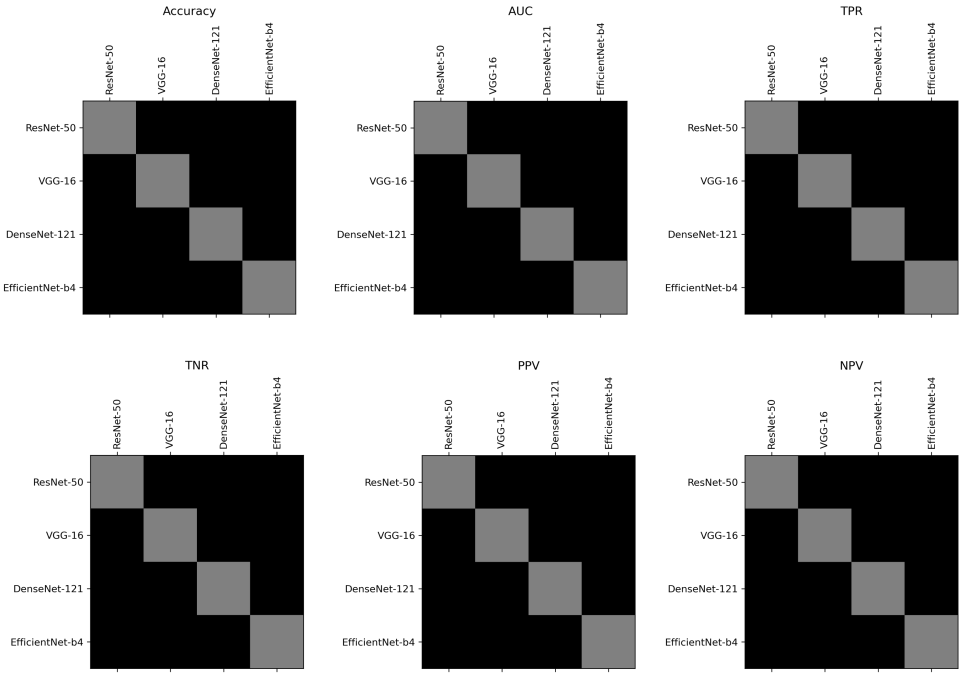


Figure 6.2: The six Wilcoxon signed-rank tests with significance level $\alpha = 0.05$ for each metric obtained for the baseline models. Black cells signify that we can not reject the null hypothesis for the corresponding pair of models and white cells are the opposite. The metrics are identified at the top of each square.

	Accuracy	AUC	Sensitivity	Specificity	PPV	NPV
$\mathcal{B}_1$(ResNet-50)	.785 (.708-.862)	.776 (.693-.860)	.665 (.467-.867)	.885 (.800-.971)	.865 (.757-.971)	.784 (.677-.900)
$\mathcal{B}_2$(ResNet-50)	**.815 (.723-.908)**	.809 (.717-.902)	.735 (.533-.833)	**.886 (.714-1.000)**	**.891 (.742-1.000)**	.809 (.717-.871)
$\mathcal{B}_1$(VGG-16)	**.815 (.692-.923)**	**.815 (.693-.926)**	.800 (.633-.967)	.829 (.686-.943)	.807 (.679-.921)	.842 (.708-.975)
$\mathcal{B}_2$(VGG-16)	.785 (.662-.908)	.789 (.671-.910)	**.867 (.767-.967)**	.713 (.514-.914)	.765 (.616-.914)	**.855 (.756-.956)**

Table 6.4: 5-fold cross validation mean metric estimates with 95% confidence interval from a bootstrapped set ($n = 15000$), using the proposed models. The values in bold represent the highest value for each metric and $\mathcal{B}_i$ (CNN) represents the $\mathcal{B}_i$ approach using the embeddings of the CNN selected.

With these results, it is not so obvious which one is the best model. In Table 6.4, the higher values for each metric were obtained with different models. The best AUC was obtained with $\mathcal{B}_1$ (VGG-16), while the worst was with $\mathcal{B}_1$ (ResNet-50). Regarding the Specificity and PPV, $\mathcal{B}_2$ (ResNet-50) displayed the highest values, whereas $\mathcal{B}_2$ (VGG-16) obtained the worst. The best Sensitivity and NPV were obtained with $\mathcal{B}_2$ (VGG-16), while the worst were observed in $\mathcal{B}_1$ (ResNet-50). The best accuracy was obtained with $\mathcal{B}_2$ (ResNet-50) and $\mathcal{B}_1$ (VGG-16). A very important detail is that $\mathcal{B}_2$ (VGG-16) was the only model that got a Sensitivity higher than the Specificity which means this model is the only one that is better at detecting the positive class

DNN backbone	Likelihood Ratio	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± St.dev
$\mathcal{B}_1$ (ResNet-50)	LR+	66.700	3.497	33.300	2.913	6.993	22.680± 24.720
	LR-	.333	.583	.667	.234	0	.363 ± .241
$\mathcal{B}_2$ (ResNet-50)	LR+	83.300	1.942	33.300	83.300	5.825	**41.533 ± 35.775**
	LR-	.167	.292	.667	.167	.195	.298 ± .190
$\mathcal{B}_1$ (VGG-16)	LR+	6.993	3.497	83.300	1.555	6.993	20.467 ± 31.486
	LR-	0	.583	.167	.583	0	.267 ± .266
$\mathcal{B}_2$ (VGG-16)	LR+	6.993	1.459	66.700	1.459	6.993	16.721± 25.112
	LR-	0	.389	.333	.389	0	**.222 ± .183**

Table 6.5: The LR+ and LR- of each proposed approach on each fold.

than the negative class.

Further the distinction between $\mathcal{B}_1$ and $\mathcal{B}_2$ is unclear. Using ResNet-50, the $\mathcal{B}_2$ yields higher AUC. For the case of VGG-16, $\mathcal{B}_1$ obtained the higher. However, the Sensitivity is higher in the two configurations with $\mathcal{B}_2$. The other metrics are inconclusive. Table 6.5 suggests ResNet-50 benefits more with $\mathcal{B}_2$. In contrast, regarding VGG-16, $\mathcal{B}_1$ seems to be the better option.

For the proposed approaches, the Wilcoxon signed rank test was also implemented to understand if a statistical difference existed between the predcitions of the proposed models. The results presented in Fig. 6.3 and Fig. 6.4 showed that for all the metrics there was no statistical difference between the models.

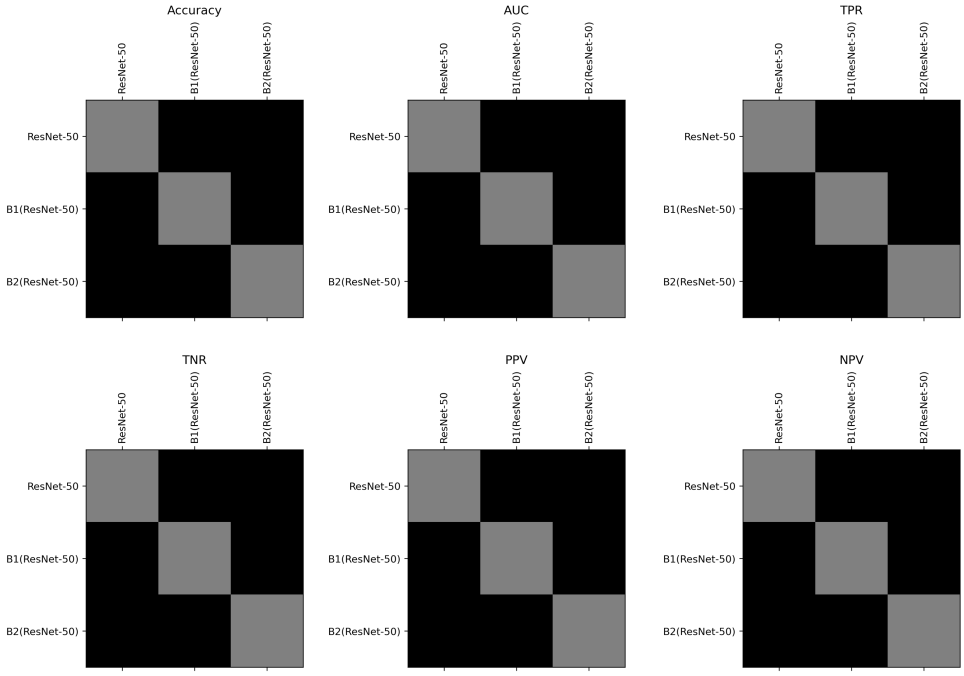


Figure 6.3: The six Wilcoxon signed-rank tests with significance level $\alpha = 0.05$ for each metric obtained for the baseline model ResNet-50 and the two approaches that use this model. Black cells signify that we can not reject the null hypothesis for the corresponding pair of models and white cells are oppossite.

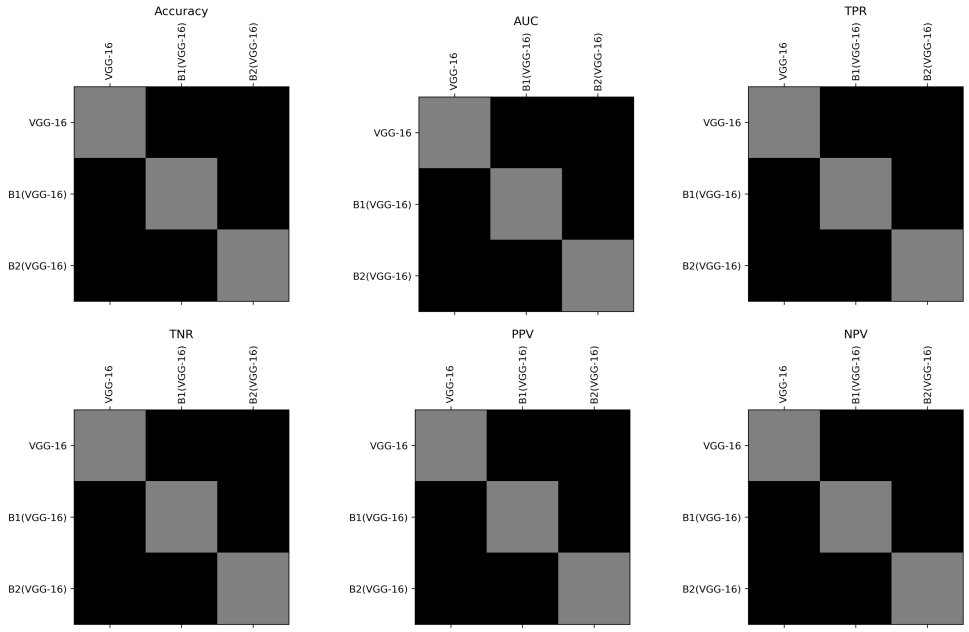


Figure 6.4: The six Wilcoxon signed-rank tests with significance level $\alpha = 0.05$ for each metric obtained for the baseline model VGG-16 and the two approaches that use this model. Black cells signify that we can not reject the null hypothesis for the corresponding pair of models and white cells are the opposite. The metrics are identified at the top of each square.

Regarding the inter-fold variability, it seems that it remains unaltered. With these new approaches the range of the confidence intervals is still high, the values for the likelihood ratios (see Table 6.5) are very different for each fold in all models and the standard deviation is also high. Fold 2 seems again to be the most difficult fold for the models. The model with the higher mean LR+ was $\mathcal{B}_2$ (ResNet-50), while the lower mean LR- was obtained by the $\mathcal{B}_2$ (VGG-16).

Comparing the Tables 6.2 and 6.4, it is concluded that the $\mathcal{B}_1$ (VGG-16) increased the performance of the VGG-16 for all the metrics, and $\mathcal{B}_2$ (VGG-16) also improved the results except for the Specificity. On the other hand, the conclusions for ResNet-50 are different since the proposed approaches were superior only in two metrics, the Specificity and the PPV. This means that VGG-16 improves with the texture description, while ResNet-50 does not.

In order to understand better the difference between the proposed models, an analysis of the misclassified images for each model was conducted. The incorrect images classified by the models that use ResNet-50 are represented in Fig. 6.5. With these examples, it seems that these models are influenced by images affected by external factors (see image C and D in Fig. 6.5) and without zoom (see image A in Fig. 6.5). The models also failed in challenging images which are hard to identify even to a human operator, such as the image B in Fig. 6.5, that seems to have a lesion but it is classified as normal. On one hand, the image failed by the baseline ResNet-50, image E in Fig. 6.5, indicates that the baseline model classifies an image as positive if the correct textural pattern is present. On the other hand, the image failed by the $\mathcal{B}_1$ (ResNet-50) and $\mathcal{B}_2$ (ResNet-50), image F in Fig. 6.5 suggests that these approaches may not work equally well for

all of the required textural patterns.

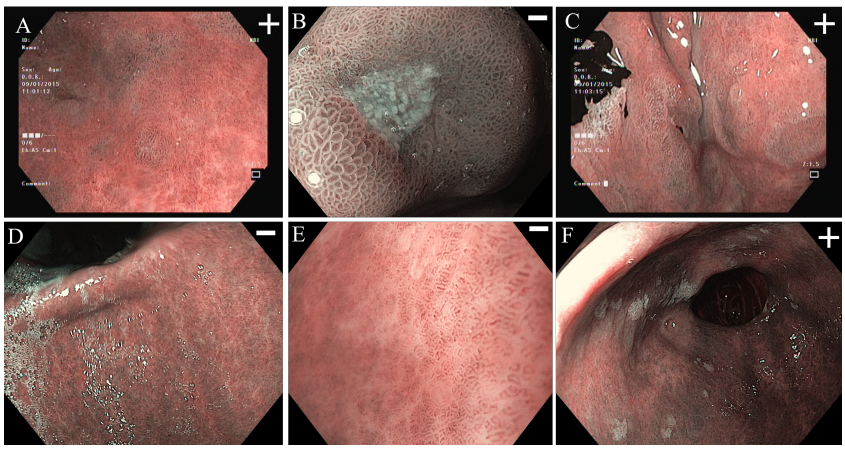


Figure 6.5: Examples of images misclassified by the approaches that use the ResNet-50. The true label is in the right top corner of the image. A, B, C, and D are images misclassified by all the models; E is the image misclassified by baseline ResNet-50 but correctly classified by $\mathcal{B}_1$ (ResNet-50) and $\mathcal{B}_2$ (ResNet-50); F is misclassified by $\mathcal{B}_1$ (ResNet-50) and $\mathcal{B}_2$ (ResNet-50) but correctly classified by baseline ResNet-50.

Observing the images where the models that leveraged VGG-16 failed it can be concluded that the baseline VGG-16 tends to fail in easy positive images, and this appears to be mitigated in $\mathcal{B}_1$ (VGG-16) and $\mathcal{B}_2$ (VGG-16) (see images A and B in Fig. 6.6). Also, the baseline VGG-16 seems to not be able to assign the correct textural pattern to the appropriate classes. For example, it classified images B and C from Fig. 6.6 incorrectly. Furthermore $\mathcal{B}_1$ (VGG-16) and $\mathcal{B}_2$ (VGG-16) seem to be influenced more drastically by the presence of external factors (see images D and F in Fig. 6.6) and by the absence of a texture pattern (see image E in Fig. 6.6).

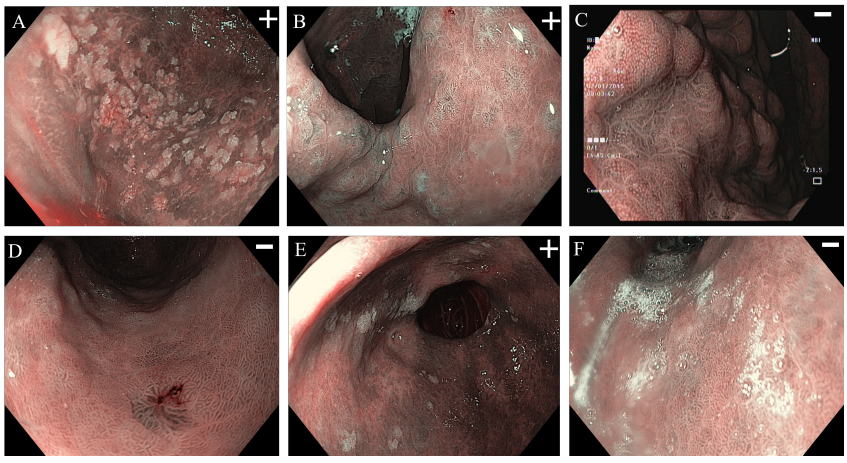


Figure 6.6: Examples of images misclassified by the approaches that use the VGG-16. The true label is in the right top corner of the image. A, B, and C are images misclassified by the baseline VGG-16, but correctly classified by $\mathcal{B}_1$ (VGG-16) and $\mathcal{B}_2$ (VGG-16); D; E, and F are the images misclassified by $\mathcal{B}_1$ (VGG-16) and $\mathcal{B}_2$ (VGG-16), but correctly classified by the baseline VGG-16.

Chapter 7

Conclusion

In this work, two novel Deep Neural Networks (DNN) models combined with a texture descriptor based on fractal geometry were proposed to identify Gastric Intestinal Metaplasia (GIM) in endoscopic images. In this chapter, the main outcomes of the work developed are presented (Section 7.1) and the ideas for the future work are outlined (Section 7.2).

7.1 Main Outcomes

The prognosis for patients diagnosed with advanced gastric cancer is characterised by a low survival rate. Nevertheless, an early diagnosis of this disease can significantly enhance the life expectancy of the patients. The condition that precedes Gastric Cancer (GC) is GIM, and patients with this lesion are very likely to develop advanced gastric cancer. The detection of GIM is very challenging due to its fine-grained details and a solution that has obtained very promising results are the DNN. However, for enhanced generalisation capability, a greater quantity of high-quality data is necessary. Regrettably, endoscopic images are expensive to collect and there are few high-quality datasets available. In order to mitigate this, two bilinear models based on DNN were proposed with the introduction of a fractal texture descriptor to compensate for the lack of high-quality data.

The texture descriptor used was the Multi-Fractal Spectrum (MFS) since it was proved in [66] that this descriptor was (theoretically) invariant under bi-Lipschtz transformation, which includes perspective and illumination transformations that are frequently present in endoscopic images. The proposed approaches combine the outputs of a Convolutional Neural Networks (CNN) with the description obtained by this descriptor. The main difference between the two approaches is the way MFS is interpreted. In the first approach, a high response of this descriptor was considered a discriminative feature for distinguishing between the two classes, while in the second the local description was not filtered in order to leverage local information in its entirety information could be helpful for the model. With the proposed approaches, the highest Positive Predictive Value (PPV) and Sensitivity were obtained which can be a good indicator of the

quality of the models in detecting the positive class.

We conclude that the idea of combining a texture descriptor with a DNN showed promising results. Due to its early prediction capability, we think that it can be an effective tool to help endoscopists in detecting GIM in endoscopic images in an actual clinical setting. For this, the proposed approaches could be integrated into a prototype that could be used while the images are being captured during the *Esophagogastroduodenoscopy* (EGD).

7.2 Future Work

The work developed opened a range of new ideas to improve the quality of the proposed approaches and explore new ways of applying a texture descriptor to enhance the diagnostic of GIM in endoscopic images. The proposed approaches have six limitations that can be solved to improve the performance of the models. First, the choice of the size and level of overlap of the patches used for the computation of the MFS were defined to be compatible with the size of the embeddings of the CNN. Second, a per-patient analysis was not conducted since we do not have information about the patients in the dataset. Also, the images were only annotated by one endoscopist. These two limitations could be tackled by finding a dataset with more information annotated by more than one endoscopist. In addition, it is possible that the selected CNNs backbones may not be optimally suited for this specific task, or that the dataset lacks the necessary quality to effectively address this challenging problem. Finally, the results can be influenced by a potential bias within the test set giving the reduced size of the test set and the random partition conducted in the 5-fold cross-validation.

Furthermore, future work could also further explore the idea of using descriptors based on fractal dimension to describe GIM in an end-to-end fashion. Our first proposal is to implement a network similar to the one proposed by Xu *et al.* [67] using fractal encoders, which are blocks that can be combined with a CNN where a soft and differential version of the MFS is applied. The second one is applying the Cross-Layer Aggregation of Statistical Self-similarity Network (CLASSNet) proposed by Chen *et al.* [9] in our dataset. This last proposal is based on the idea of estimating the local fractal dimension using another method in the output of each residual block of a ResNet-50, and combine its outputs with the embeddings of the CNN. These two approaches aim to summarise in a more nuance manner the embeddings of the CNNs in contrast with Global Average Pooling (GAP).

Bibliography

[1] Jaffer A. Ajani, David J. Bentrem, Stephen Besh, Thomas A. D'Amico, Prajnan Das, Crystal Denlinger, Marwan G. Fakih, Charles S. Fuchs, Hans Gerdes, Robert E. Glasgow, and et al. Gastric cancer, version 2.2013. *Journal of the National Comprehensive Cancer Network*, 11 (5):531–546, 2013. doi:10.6004/jnccn.2013.0070.

[2] Julia Arribas, Giulio Antonelli, Leonardo Frazzoni, Lorenzo Fuccio, Alanna Ebigbo, Fons van der Sommen, Noha Ghatwary, Christoph Palm, Miguel Coimbra, Francesco Renna, et al. Standalone performance of artificial intelligence for upper GI neoplasia: a meta-analysis. *Gut*, 70(8):1458–1468, 2021.

[3] Michael F. Barnsley, Robert L. Devaney, Benoit B. Mandelbrot, Heinz-Otto Peitgen, Dietmar Saupe, and Richard F. Voss. *The Science of Fractal Images*. Springer-Verlag, 1988.

[4] Hanna Borgli, Vajira Thambawita, Pia H. Smedsrud, Steven Hicks, Debesh Jha, Sigrun L. Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, and et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Nature News*, Aug 2020.

[5] Nataliya Boyko, Maxim Kuba, Lesia Mochurad, and Sergio Montenegro. Fractal distribution of medical data in neural network. In *IDDM*, pages 307–318, 2019.

[6] Lisette G. Capelle, Jelle Haringsma, Annemarie C. de Vries, Ewout W. Steyerberg, Katharina Biermann, Herman van Dekken, and Ernst J. Kuipers. Narrow band imaging for the detection of gastric intestinal metaplasia and dysplasia during surveillance endoscopy. *Digestive Diseases and Sciences*, 55(12):3442–3448, 2010. doi:10.1007/s10620-010-1189-2.

[7] M. Chantler, M. Petrou, A. Penirsche, M. Schmidt, and G. MGunnigle. Classifying surface texture while simultaneously estimating illumination direction. *International Journal of Computer Vision*, 62:83–96, 04 2005. doi:10.1023/B:VISI.0000046590.98379.19.

[8] Di Chen, Lianlian Wu, Yanxia Li, Jun Zhang, Jun Liu, Li Huang, Xiaoda Jiang, Xu Huang, Ganggang Mu, Shan Hu, et al. Comparing blind spots of unsedated ultrafine, sedated, and unsedated conventional gastroscopy with and without artificial intelligence: a prospective, single-blind, 3-parallel-group, randomized, single-center trial. *Gastrointestinal endoscopy*, 91 (2):332–339, 2020.

[9] Zhile Chen, Feng Li, Yuhui Quan, Yong Xu, and Hui Ji. Deep texture recognition via exploiting cross-layer statistical self-similarity. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. doi:10.1109/cvpr46437.2021.00519.

[10] Pelayo Correa and M Blanca Piazuelo. The gastric precancerous cascade. *Journal of Digestive Diseases*, 13(1):2–9, 2011. doi:10.1111/j.1751-2980.2011.00550.x.

[11] Simon S. Cross, Jonathan P. Bury, Paul B. Silcocks, Timothy J. Stephenson, and Dennis W. Cotton. Fractal geometric analysis of colorectal polyps. *The Journal of Pathology*, 172(4): 317–323, 1994. doi:10.1002/path.1711720406.

[12] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

[13] Mark Feldman, Lawrence S. Friedman, and John S. Fordtran. *Adenocarcinoma of the Stomach and Other Gastric Tumors*, volume 1, page 820. Elsevier, 11 edition, 2021.

[14] Carlos A. González, Maria Luisa Pardo, Juan Maria Ruiz Liso, Pablo Alonso, Catalina Bonet, Raul M. Garcia, Núria Sala, Gabriel Capella, and José Miguel Sanz-Anquela. Gastric cancer occurrence in preneoplastic lesions: A long-term follow-up in a high-risk area in Spain. *International Journal of Cancer*, 127(11):2654–2660, 2010. doi:10.1002/ijc.25273.

[15] Pedro Guimarães, Andreas Keller, Tobias Fehlmann, Frank Lammert, and Markus Casper. Deep-learning based detection of gastric precancerous conditions. *Gut*, 69(1):4–6, 2020. doi:10.1136/gutjnl-2019-319347.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[17] Toshiaki Hirasawa, Kazuharu Aoyama, Tetsuya Tanimoto, Soichiro Ishihara, Satoki Shichijo, Tsuyoshi Ozawa, Tatsuya Ohnishi, Mitsuhiro Fujishiro, Keigo Matsuo, Junko Fujisaki, and et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric cancer : official journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association*, Jan 2018.

[18] Yusuke Horiuchi, Kazuharu Aoyama, Yoshitaka Tokai, Toshiaki Hirasawa, Shoichi Yoshimizu, Akiyoshi Ishiyama, Toshiyuki Yoshio, Tomohiro Tsuchida, Junko Fujisaki, Tomohiro Tada, and et al. Convolutional neural network for differentiating gastric cancer from gastritis using magnified endoscopy with narrow band imaging. *Digestive diseases and sciences*, May 2020.

[19] Hao Hu, Lixin Gong, Di Dong, Liang Zhu, Min Wang, Jie He, Lei Shu, Yiling Cai, Shilun Cai, Wei Su, Yunshi Zhong, Cong Li, Yongbei Zhu, Mengjie Fang, Lianzhen Zhong, Xin Yang, Pinghong Zhou, and Jie Tian. Identifying early gastric cancer under magnifying narrow-band images with deep learning: a multicenter study. *Gastrointestinal Endoscopy*, 93(6):1333–1341.e3, 2021. doi:https://doi.org/10.1016/j.gie.2020.11.014.

[20] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.

[21] Michael Häfner, Toru Tamaki, Shinji Tanaka, Andreas Uhl, Georg Wimmer, and Shigeto Yoshida. Local fractal dimension based approaches for colonic polyp classification. *Medical Image Analysis*, 26(1):92–107, 2015. doi:https://doi.org/10.1016/j.media.2015.08.007.

[22] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[23] Jahidul Islam, Sajjad Bhuiyan, Arafat Hossain, Amit Shaha Surja, and Md. Shahid Iqbal. Classification of gastric precancerous diseases using hybrid CNN-SVM. In *2021 3rd International Conference on Electrical Electronic Engineering (ICEEE)*, pages 137–140, 2021. doi:10.1109/ICEEE54059.2021.9718790.

[24] Yoh Isobe, Atsushi Nashimoto, Kohei Akazawa, Ichiro Oda, Kenichi Hayashi, Isao Miyashiro, Hitoshi Katai, Shunichi Tsujitani, Yasuhiro Kodera, Yasuyuki Seto, and et al. Gastric cancer treatment in Japan: 2008 annual report of the jgca nationwide registry. *Gastric Cancer*, 14 (4):301–316, 2011. doi:10.1007/s10120-011-0085-6.

[25] V. S. Ivanova, I. J. Bunin, and V. I. Nosenko. Fractal material science: A new direction in materials science. *JOM*, 50(1):52–54, 1998. doi:10.1007/s11837-998-0068-1.

[26] David S Jencks, Jason D Adam, Marie L Borum, Joyce M Koh, Sindu Stephen, and David B Doman. Overview of current concepts in gastric intestinal metaplasia and gastric cancer, Feb 2018.

[27] Takashi Kanesaka, Tsung-Chun Lee, Noriya Uedo, Kun-Pei Lin, Huai-Zhe Chen, Ji-Yuh Lee, Hsiu-Po Wang, and Hsuan-Ting Chang. Computer-aided diagnosis for identifying and delineating early gastric cancers in magnifying narrow-band imaging. *Gastrointestinal Endoscopy*, 87(5):1339–1344, 2018. doi:https://doi.org/10.1016/j.gie.2017.11.029.

[28] Tae Kyun Kim. T-test as a parametric statistic. *Korean Journal of Anesthesiology*, 68(6): 540, 2015. doi:10.4097/kjae.2015.68.6.540.

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017. ISSN: 0001-0782. doi:10.1145/3065386.

[32] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, 2005. doi:10.1109/TPAMI.2005.151.

[33] W. K. Leung and J. J. Sung. Intestinal metaplasia and gastric carcinogenesis. *Alimentary Pharmacology amp;amp; Therapeutics*, 16(7):1209–1216, 2002. doi:10.1046/j.1365-2036.2002.01300.x.

[34] Hongyan Li, Chi Man Vong, Pak Kin Wong, Weng Fai Ip, Tao Yan, I. Cheong Choi, and Hon Ho Yu. A multi-feature fusion method for image recognition of gastrointestinal metaplasia (GIM). *Biomedical Signal Processing and Control*, 69:102909, 2021. ISSN: 1746-8094. doi:https://doi.org/10.1016/j.bspc.2021.102909.

[35] Lan Li, Yishu Chen, Zhe Shen, Xuequn Zhang, Jianzhong Sang, Yong Ding, Xiaoyun Yang, Jun Li, Ming Chen, Chaohui Jin, and et al. Convolutional neural network for the diagnosis of early gastric cancer based on magnifying narrow band imaging. *Gastric cancer : official journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association*, Jan 2020.

[36] Ne Lin, Tao Yu, Wenfang Zheng, Huiyi Hu, Lijuan Xiang, Guoliang Ye, Xingwei Zhong, Bin Ye Ye, Rong Wang, JingJing Li, and et al. Simultaneous recognition of atrophic gastritis and intestinal metaplasia on white light endoscopic images based on convolutional neural networks: A multicenter study. *Clinical and translational gastroenterology*, Aug 2021.

[37] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015. doi:10.1109/iccv.2015.170.

[38] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear models for fine-grained visual recognition. In *International Conference on Computer Vision (ICCV)*, 2015. [PowerPoint Presentation].

[39] Xiaoqi Liu, Chengliang Wang, Yao Hu, Zhuo Zeng, Jianying Bai, and Guobin Liao. Transfer learning with convolutional neural network for early gastric cancer classification on magnifiying narrow-band imaging images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1388–1392, 2018. doi:10.1109/ICIP.2018.8451067.

[40] Xiaoqi Liu, Chengliang Wang, Jianying Bai, and Guobin Liao. Fine-tuning pre-trained convolutional neural networks for gastric precancerous disease classification on magnification narrow-band imaging images. *Neurocomputing*, 392:253–267, 2020. doi:https://doi.org/10.1016/j.neucom.2018.10.100.

[41] R. Lopes and N. Betrouni. Fractal and multifractal analysis: A review. *Medical Image Analysis*, 13(4):634–649, 2009. doi:https://doi.org/10.1016/j.media.2009.05.003.

[42] Gabriele Angelo Losa. The fractal geometry of life. *Rivista di biologia*, 102 1:29–59, 2009.

[43] Lingyu Ma, Xiufeng Su, Liyong Ma, Xiaozhong Gao, and Mingjian Sun. Deep learning for classification and localization of early gastric cancer in endoscopic images. *Biomedical Signal Processing and Control*, 79:104200, 2023. doi:https://doi.org/10.1016/j.bspc.2022.104200.

[44] B.B. Mandelbrot. *The Fractal Geometry of Nature*. Einaudi paperbacks. Henry Holt and Company, 1983. ISBN: 9780716711865.

[45] Shyam Menon and Nigel Trudgill. How commonly is upper gastrointestinal cancer missed at Endoscopy? A meta-analysis. *Endoscopy International Open*, 02(02), 2014. doi:10.1055/s-0034-1365524.

[46] Syrene A. Miller and Jane L. Forrest. Enhancing your practice through evidence-based decision making: PICO, learning how to ask good questions. *Journal of Evidence Based Dental Practice*, 1(2):136–141, 2001. doi:https://doi.org/10.1016/S1532-3382(01)70024-3.

[47] SEER; Training Modules. Cancer classification.

[48] Joon Yeul Nam, Hyung Jin Chung, Kyu Sung Choi, Hyuk Lee, Tae Jun Kim, Hosim Soh, Eun Ae Kang, nbsp;Soo-Jeong Cho, Jong Chul Ye, Jong Pil Im, and et al. Deep learning model for diagnosing gastric mucosal lesions using endoscopic images: Development, validation, and method comparison. *Gastrointestinal endoscopy*, Feb 2022.

[49] Alex P. Pentland. Fractal-based description of natural scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):661–674, 1984. doi:10.1109/tpami.1984.4767591.

[50] Benjamin Planche and Eliot Andres. *Hands-on Computer Vision with tensorflow 2: Leverage deep learning to create powerful image processing apps with tensorflow 2.0 and keras*. Packt Publishing, 2019.

[51] Priya Ranganathan and Rakesh Aggarwal. Understanding the properties of diagnostic tests – part 2: Likelihood ratios. *Perspectives in Clinical Research*, 9(2):99, 2018. doi:10.4103/picr.picr_41_18.

[52] Guilherme Freire Roberto, Alessandra Lumini, Leandro Alves Neves, and Marcelo Zanchetta do Nascimento. Fractal neural network: A new ensemble of fractal geometry and convolutional neural networks for the classification of histology images. *Expert Systems with Applications*, 166:114103, 2021. doi:https://doi.org/10.1016/j.eswa.2020.114103.

[53] T Rokkas, M I Filipe, and G E Sladen. Detection of an increased incidence of early gastric cancer in patients with intestinal metaplasia type III who are closely followed up. *Gut*, 32 (10):1110–1113, 1991. doi:10.1136/gut.32.10.1110.

[54] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi:10.1007/s11263-015-0816-y.

[55] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. doi:10.1109/CVPR.2018.00474.

[56] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[58] T.G. Smith, G.D. Lange, and W.B. Marks. Fractal methods and results in cellular morphology — dimensions, lacunarity and multifractals. *Journal of Neuroscience Methods*, 69(2):123–136, 1996. doi:https://doi.org/10.1016/S0165-0270(96)00080-5.

[59] Mingjun Song and Tiing Leong Ang. Early detection of early gastric cancer using image-enhanced endoscopy: Current trends. *Gastrointestinal Intervention*, 3(1):1–7, 2014. doi:10.1016/j.gii.2014.02.005.

[60] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021. doi:10.3322/caac.21660.

[61] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks, 2020.

[62] Jonathan R. White and Matthew Banks. Identifying the pre-malignant stomach: From guidelines to practice. *Translational Gastroenterology and Hepatology*, 7:8–8, 2022. doi:10.21037/tgh.2020.03.03.

[63] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6): 80–83, 1945.

[64] Lianlian Wu, Jing Wang, Xinqi He, Yijie Zhu, Xiaoda Jiang, Yiyun Chen, Yonggui Wang, Li Huang, Renduo Shang, Zehua Dong, et al. Deep learning system compared with expert endoscopists in predicting early gastric cancer and its invasion depth and differentiation status (with videos). *Gastrointestinal Endoscopy*, 95(1):92–104, 2022.

[65] Ming Xu, Wei Zhou, Lianlian Wu, Jun Zhang, Jing Wang, Ganggang Mu, Xu Huang, Yanxia Li, Jingping Yuan, Zhi Zeng, Yonggui Wang, Li Huang, Jun Liu, and Honggang Yu. Artificial intelligence in the diagnosis of gastric precancerous conditions by image-enhanced endoscopy: a multicenter, diagnostic study (with video). *Gastrointestinal Endoscopy*, 94(3): 540–548.e4, 2021. doi:https://doi.org/10.1016/j.gie.2021.03.013.

[66] Yong Xu, Hui Ji, and Cornelia Fermüller. Viewpoint invariant texture description using fractal analysis. *SpringerLink*, Feb 2009.

[67] Yong Xu, Feng Li, Zhile Chen, Jinxiu Liang, and Yuhui Quan. Encoding spatial distribution of convolutional features for texture representation. *Advances in Neural Information Processing Systems*, 34:22732–22744, 2021.

[68] Tao Yan, Pak Kin Wong, I. Cheong Choi, Chi Man Vong, and Hon Ho Yu. Intelligent diagnosis of gastric intestinal metaplasia based on convolutional neural network and limited number of endoscopic images. *Computers in Biology and Medicine*, 126:104026, 2020. doi:https://doi.org/10.1016/j.compbiomed.2020.104026.

[69] Naohiro Yoshida, Hisashi Doyama, Tomonori Yano, Takahiro Horimatsu, Noriya Uedo, Yoshinobu Yamamoto, Naomi Kakushima, Hiromitsu Kanzaki, Shinichiro Hori, Kenshi Yao, and et al. Early gastric cancer detection in high-risk patients: A multicentre randomised controlled trial on the effect of second-generation narrow band imaging. *Gut*, 70(1):67–75, 2020. doi:10.1136/gutjnl-2019-319631.

[70] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.

[71] Yaqiong Zhang, Fengxia Li, Fuqiang Yuan, Kai Zhang, Lijuan Huo, Zichen Dong, Yiming Lang, Yapeng Zhang, Meihong Wang, Zenghui Gao, Zhenzhen Qin, and Leixue Shen. Diagnosing chronic atrophic gastritis by gastroscopy using artificial intelligence. *Digestive and Liver Disease*, 52(5):566–572, 2020. doi:https://doi.org/10.1016/j.dld.2019.12.146.

[72] Fan Zhou, Liucheng Wu, Mingwei Huang, Qinwen Jin, Yuzhou Qin, and Jiansi Chen. The accuracy of magnifying narrow band imaging (ME-NBI) in distinguishing between cancerous and noncancerous gastric lesions. *Medicine*, 97(9), 2018. doi:10.1097/md.0000000000009780.