# Fractal Bilinear Deep Neural Network Models for Gastric Intestinal Metaplasia Detection

Maria Pedroso[1], Miguel L. Martins[1], Diogo Libânio[2],
Mário Dinis-Ribeiro[2], Miguel Coimbra[1] and Francesco Renna[1]

*Abstract*— **Gastric Intestinal Metaplasia (GIM) is a precancerous gastric lesion and its early detection facilitates patient followup, thus lowering significantly the risk of death by gastric cancer. However, effective screening of this condition is a very challenging task, resulting low intra and inter-observer concordance. Computer assisted diagnosis systems leveraging deep neural networks (DNNs) have emerged as a way to mitigate these ailments. Notwithstanding, these approaches typically require large datasets in order to learn invariance to the extreme variations typically present in *Esophagogastroduodenoscopy* (EGD) still frames, such as perspective, illumination, and scale. Hence, we propose to combine *a priori* information regarding texture characteristics of GIM with data-driven DNN solutions. In particular, we define two different models that treat pre-trained DNNs as general features extractors, whose pairwise interactions with a collection of highly invariant local texture descriptors grounded on *fractal geometry* are computed by means of an outer product in the embedding space. Our experiments show that these models outperform a baseline DNN by a significant margin over several metrics (*e.g.*, area under the curve (AUC)** 0.792 ***vs*. **0.705) **in a dataset comprised of EGD narrow-band images. Our best model measures double the positive likelihood ratio when compared to a baseline GIM detector.**

*Clinical relevance*— **Better automatic tools for Gastric Intestinal Metaplasia detection can help mitigate human diagnostic errors, which directly impacts gastric cancer mortality rate.**

## I. INTRODUCTION

Gastric cancer (GC) is the fifth most prevalent form of cancer worldwide and is responsible for causing the third-highest number of cancer-related deaths [1]. According to the European Society of Gastrointerninal Endoscopy (ESGE), early diagnosis is crucial since the survival rate is just 24%. However, it is also expected that effective early diagnosis might lower mortality rates by 40% [2]. Improving the diagnosis of Gastric Intestinal Metaplasia (GIM) is crucial for screening effectiveness, since it is a critical precursor to gastric cancer. Patients with GIM suffer from an increased risk of developing gastric cancer by a factor of 10 [3]. GIM can be characterized during *Esophagogastroduodenoscopy* (EGD), a minimally invasive procedure for diagnosing pre-cancerous or early cancerous conditions by detecting aberrant tissue in the gastric mucosa using a Narrow-band Imaging (NBI) modality. GIM screening is challenging since accurate diagnosis depends on fine-grained details that characterize GIM lesions on the gastric mucosa. In fact, the inter-observer concordance among clinicians is low, with reports of up to 11.3% of upper gastrointestinal (UGI) cancer-related lesions being missed during endoscopic screening up to 3 years before diagnosis [4]. This motivates us to find an automatic tool that is unaffected by subjective factors for optical diagnosis, such as the currently emerging deep neural networks (DNN). DNNs have been successfully applied in the domain of UGI endoscopy, such as landmark detection [5], detection of gastric cancer [6], prediction of invasion status [7] and GIM detection [8]. However, the data-driven nature of these approaches implies that good performance is achieved in the presence of a large quantity of high-quality data for training. On the other hand, collecting endoscopic images is an expensive procedure, with very limited public access datasets. Furthermore, the downstream task of GIM detection is very challenging since the views of the mucosa are subjected to considerable scale and perspectives variance, while simultaneously being populated by other phenomena, such as bubbles, undigested food, and blood. In order to efficiently cope with the lack of large, high-quality, annotated dataset, we introduce a stronger inductive bias given knowledge of the importance of texture in GIM detection by leveraging features based on *fractal geometry*. The idea that fractal descriptors combined with deep learning models could result in a robust GIM detector was motivated by the usefulness of fractal dimension in detecting texture patterns, especially in natural images [9][10] and by the discriminative features already obtained in a modality with very close visual characteristics, more specifically, that of polyp characterization in colonoscopy [11].

This work is financed by National Funds through the Portuguese funding agency, FCT—Fundação para a Ciência e a Tecnologia, within project PTDC/EEI-EEE/5557/2020. Co-funded by the European Union (grant number 101095359) and supported by the UK Research and Innovation (grant number 10058099). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. Miguel L. Martins' work is also funded by the individual FCT grant reference number 2021.06503.BD

[1] Miguel L. Martins, Maria Pedroso, Miguel Coimbra and Francesco Renna are with INESC TEC - Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência, Faculty of Science, University of Porto, (emails: `miguel.l.martins@inesctec.pt`, `up201805037@edu.fc.up.pt`, {`mcoimbra`, `francesco.renna`}`@fc.up.pt`).

[2]Diogo Libânio and Mário Dinis-Ribeiro are with CIDES/CINTESIS, Faculty of Medicine, University of Porto, (email: `diogolibanio@med.up.pt`, `mdinisribeiro@gmail.com`).

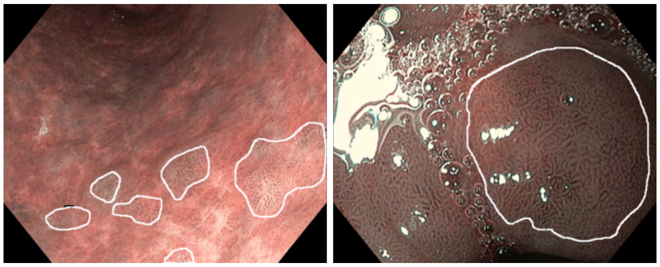


Fig. 1: Examples of two endoscopic images with GIM outlined in white.

### A. Contributions

We propose a new approach that bridges the gap between fine-grained classification and local texture descriptors through a unifying bilinear model that combines the outputs of a general purpose a convolutional neural network (CNN) with patch-wise multi-fractal spectra. We show its applicability in the task of GIM detection, a biomedical application that benefits from the enhanced discriminative properties of this approach since its target variable class depends on subtle local aberrations in the gastric lining across several scales, perspectives, and illumination variations. Our model outperforms a pure CNN baseline throughout area under the curve (AUC), true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV) and negative predictive value (NPV) by almost 10% and exceeds the results reported in a previous work [12].

## II. MATERIALS AND METHODS

### A. The GIM-NBI dataset

For the experimental purposes of this paper, we use a dataset collected at the Gastroenterology department of Instituto Português de Oncologia, Porto (IPO-Porto), spanning 883 high-resolution images: 808 classified as normal, 64 as GIM, 10 as dysplasia or carcinoma, and 1 as atrophic gastritis. These data displays three distinct modalities: White Light Imaging (WLI) and NBI. An endoscopist filtered frames with the incorrect diagnosis of GIM, low resolution, and frames captured in WLI. This results in a total of 125 high-quality NBI images, 65 classified as normal (- class) and 60 as GIM (+ class). If a discernible metaplastic pattern is visible in the mucosa and there are no other pathological findings in the scene, the image is classified as belonging to the + class. The dataset is populated with frames collected from standard clinical practice, so the mucosa is not always captured under ideal conditions. Consequently, foam, bubbles, bile, blood, alongside other pathological findings such as polyps may be captured in the scene.

### B. Methodology

*1) Bilinear model:* We follow [13], and define a bilinear model as a quadruple $\mathscr{B} = (f_a, f_b, \mathscr{P}, \mathscr{C})$, where $f_a : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{k\times A}$ and $f_b : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{k\times B}$ are feature extracting functions, $\mathscr{P}$ is a pooling function and $\mathscr{C}$ is a classification function. A feature extracting function receives an image and a location and outputs a feature vector. The basic idea behind this model is to combine the output of these two feature functions using the outer product at each location of the image. Then, we define the pooling function as the outer product between $f_a$ and $f_b$, specifically $f_a(\mathbf{x})^{\mathrm{T}} f_b(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ and $(\cdot)^{\mathrm{T}}$ stands for the transpose operator. Note that this pooling function is *orderless*, which is proven to be an important property for texture and fine-grained classification [13], [14]. Finally, this image descriptor is given to the classification function to obtain the predicted class. The main advantage of this model is that the outer product captures pairwise correlations between the feature channels and can model part-feature interactions.

*2) Multi-fractal Spectrum:* A fractal point set $\mathbb{F}$ has *fine structure*, *i.e.*, detail at arbitrary scales. Formally, this set is required to have *fractal dimension* ($\beta$) greater than its *topological dimension* [15]. For our purposes, we can only compute estimates of the fractal dimension due to finite resolution. A common empirical prior is assuming that the number of $\delta$-covers $M$ that span $\mathbb{F}$ vary proportionally to the power of $\delta$, as $\delta \to 0$:

$$M(\mathbb{F},\delta) \approx k\delta^{-\beta} \implies \log M(\mathbb{F},\delta) \approx \log k - \beta \log \delta, \tag{1}$$

where $k, \beta \in \mathbb{R}$. Consequently, the *empirical estimation* of the fractal dimension $\beta$ will be:

$$\beta = \lim_{\delta \to 0} \frac{\log M(\mathbb{F},\delta)}{-\log \delta}. \tag{2}$$

This result is of particular computational relevance, since $\beta$ can be approximated by determining the slope of the plot of $\log M(\mathbb{F},\delta)$ *versus* $\log \delta$, for an appropriate finite range of $\delta$. Note that $\beta$ is a global statistic with regards to $\mathbb{F}$, which does not translate realistically for most practical settings. A simple example of its limitations is the case when $\mathbb{F}$ can be partitioned into two or more disjoint sets generated by independent fractal processes whose fractal dimensions are significantly different under (2). Thus, we are also interested in capturing local fractal behavior, *i.e.*, a *spectrum* of fractal dimensions that more accurately describe $\mathbb{F}$. We use the definition of a *Multi-fractal spectrum* (MFS) as proposed by Yong Xu *et al.* [9] in order to collect viewpoint invariant texture feature tensors.

Consider an image $I$ defined over $\Omega \subseteq \mathbb{R}^2$ and let $\mu$ be a measure over $\Omega$ 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 $k \in \mathbb{R}$. Then, the *local density function* or *Hölder exponent* is determined as:

$$\hat{\beta}(I,\mathbf{x}) = \lim_{r \to 0} \frac{\mu\left(B(I(\mathbf{x}), r)\right)}{\log r}, \tag{3}$$

where $B(I(\mathbf{x}), r)$ denotes a closed disk of length $r$ around coordinate $\mathbf{x}$ in $I$. Clearly the density $\hat{\beta}(I,\mathbf{x})$ can be estimated in a very similar way to (2). We partition $\Omega$ using the following level-set categorization:

$$\mathbb{F}_{\bar{\beta}} = \left\{ \mathbf{x} \in \Omega : \hat{\beta}(I,\mathbf{x}) = \bar{\beta} \right\}. \tag{4}$$

The MFS can thus be determined by computing the fractal dimension according to (2) for each possible categorization:

$$\mathrm{MFS}(I) = \left\{ \lim_{\delta \to 0} \frac{\log M(\mathbb{F}_{\bar{\beta}}, \delta)}{-\log \delta} : \bar{\beta} \in \mathbb{R} \right\}. \tag{5}$$

In practice, we define a suitable range $N$, and compute (5) using a uniform partition of $[0, N]$ into $m$ discrete uniformly spaced bins, instead of operating point-wise for all $\mathbb{R}$.

Note that we chose the above-mentioned construction of $\mathrm{MFS}(\cdot)$ since it can be shown that it is invariant under the bi-Lipschitz map [9]. These theoretical guarantees make the descriptor specially suitable for our downstream task, since the region of interest can be subjected to drastic perspective, scale, color, and illumination changes.

Concerning the choice of $\mu$, we again follow Yong Xu *et al.* [9], and define three distinct measures. Firstly:

$$\mu_1\big(B(I(\mathbf{x}), r)\big) = \int_{B(I(\mathbf{x}), r)} G_r * I(\mathbf{x})\, d\mathbf{x}, \tag{6}$$

where $G_r$ is a Gaussian blur filter with variance $r$, and '$*$' is the convolution operator. Secondly, assuming that $g_1, g_2, g_3$, and $g_4$ are the differential operator for vertical, horizontal, diagonal, and anti-diagonal directions, respectively:

$$\mu_2\big(B(I(\mathbf{x}), r)\big) = \int_{B(I(\mathbf{x}), r)} \sum_{i=1}^{4} g_i\big(G_r * I(\mathbf{x})\big)\, d\mathbf{x}. \tag{7}$$

Finally, we also calculate the MFS using a measure based on the sum of Laplacians:

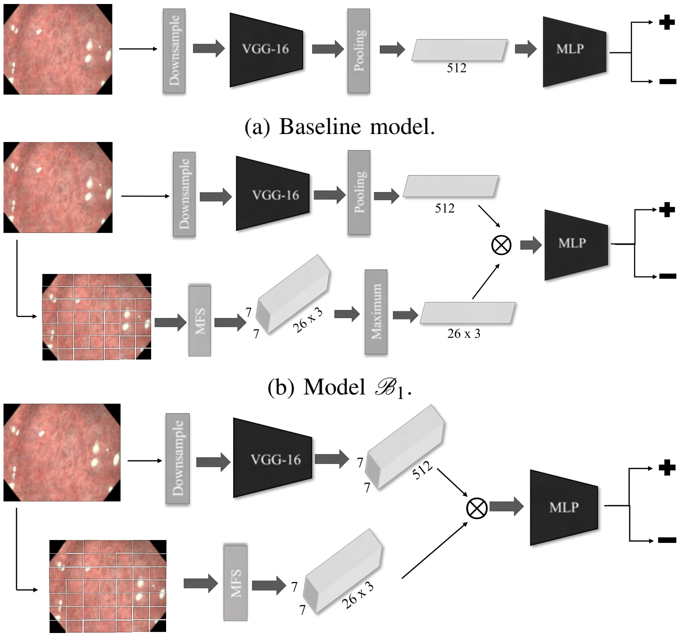
$$\mu_3\big(B(I(\mathbf{x}), r)\big) = \int_{B(I(\mathbf{x}), r)} |\nabla^2\big(G_r * I(\mathbf{x})\big)| d\mathbf{x}. \tag{8}$$

Note that we convolve with gaussian smoothing kernels for each $\mu$. This attenuates the local effect of noise, which can result in imprecise estimates of the Hölder exponents.

The final MFS feature tensor is simply the concatenation of the MFS spectra over $\mu_1$, $\mu_2$, and $\mu_3$.

*3) Pre-processing:* First, we crop excessive black borders and remove the system status information for each frame, resulting in image tensors with the shape $1078 \times 1348 \times 3$. Then, we increased the dimension of the images to $1078 \times 1351 \times 3$ using bilinear interpolation, and divide the images into $7 \times 7$ non-overlapping grayscale patches with the shape of $154 \times 193 \times 1$. This makes the dimension compatible with the embedding of the VGG-16, that is $7 \times 7 \times 512$. Then, the MFS of each patch is computed and the resulting spectra are collected into a tensor of dimension $7 \times 7 \times 78$. The second dimension of the tensor is 78 since we are considering $m = 26$ points for each of the MFS spectra. The MFS vectors computed were normalized using the standardization method ($\mu = 0$ and $\sigma = 1$).

*4) Proposed approaches:* We propose two bilinear models $\mathscr{B}_1$ and $\mathscr{B}_2$, that use the same pooling function defined in section II-B.1, and the same classification function that is a multi-layer perceptron (MLP) composed by two consecutive dense layers with rectified linear unit activation function followed by their respective dropout probabilities and an output layer which is a single neuron with a sigmoid activation function.



(a) Baseline model.

(b) Model $\mathscr{B}_1$.

(c) Model $\mathscr{B}_2$.

Fig. 2: Structure of the baseline and the proposed approaches.

For $\mathscr{B}_1$, $f_a$ is the output of the global average pooling of a pre-trained VGG-16 and $f_b$ the maximum response over $\mu_1$, $\mu_2$, and $\mu_3$. The idea of using the maximum response is to take the maximum along patches, so that, if there is a patch with a high response that should be interpreted as a portion of the image with GIM (see Fig. 2b).

Concerning $\mathscr{B}_2$, $f_a$ we directly use the embeddings of the VGG-16 and $f_b$ the response over $\mu_1$, $\mu_2$, and $\mu_3$ for the 49 patches (see Fig. 2c). In contrast with $\mathscr{B}_1$, $\mathscr{B}_2$ retains information about patch location before applying the outer product, thus enabling explicit spatial representation of pairwise interactions.

*5) Data augmentation:* In order not to compromise the visibility of a lesion in an image in the $(+)$ class, we decided to limit our data augmentation to three linear transformations: random horizontal and vertical flips, and adding Principal Component Gaussian noise to the color channels [16].

## III. Experimental Methodology

We implement a VGG-16 pre-trained in the ImageNet dataset as a baseline (see Fig. 2a) since it achieved the best performance among different pre-trained architectures for GIM detection tested over our dataset in a previous work [12], and we use the same hyperparameters for the present study. In each of the three experiences, we used stratified 5-fold cross-validation, where each fold was split into 100 samples for the train set and 25 for the test set. For each train set, the online data augmentation yielded an average of 1214 samples, 632 of the negative class and 582 of the positive one. The same VGG-16 backbone is used as a feature function in our proposed models. In order to estimate

TABLE I: The positive (LR+) and negative (LR-) likelihood ratios of each configuration on each fold.

| Models | Likelihood ratio | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Mean ± St. dev. |
|---|---|---|---|---|---|---|---|
| Baseline | LR+ | 9.750 | 1.083 | 2.167 | 1.950 | 4.333 | 3.857± 3.135 |
| | LR- | .271 | .813 | .789 | .406 | 0 | .455 ± .310 |
| $\mathscr{B}_1$ | LR+ | 11.917 | 1.625 | 8.667 | 2.167 | 10.833 | 7.042 ± 4.333 |
| | LR- | .090 | .464 | .361 | .481 | .181 | .316 ± .155 |
| $\mathscr{B}_2$ | LR+ | $> 20$ | 1.733 | 4.333 | 1.986 | 10.833 | **7.777** ± **6.938** |
| | LR- | .083 | .542 | .394 | .155 | .181 | **.271** ± **.170** |

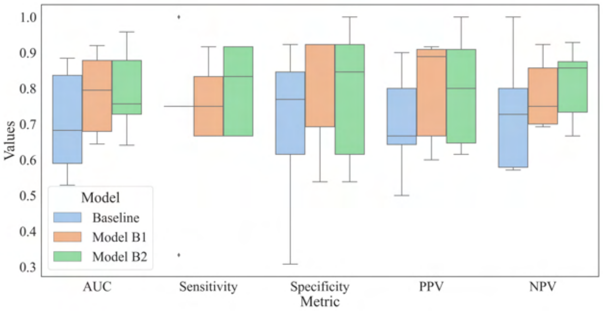


Fig. 3: Boxplots with the metrics computed for each fold obtained in 5-fold cross-validation for the three different models.

the MFS, we set $r, \delta \in \{1,2,3,4,5,6,7,8\}$ and we defined $m = 26$ (number of bins to partition the interval $[0,N]$) . To assess the performance of the models we selected 5 evaluation metrics: positive predictive value (PPV), negative predictive value (NPV), Sensitivity, Specificity, and AUC (Fig. 3). We also calculated Positive (LR+) and Negative (LR-) likelihood ratios:

$$LR+ = \frac{Sensitivity}{1 - Specificity}, \tag{9}$$

$$LR- = \frac{1 - Sensitivity}{Specificity}. \tag{10}$$

We display the LR+ and LR- in Table I to evaluate the inter-fold diagnostic variability for each model. Notice that when computing the LR+, if the specificity is 1, the value is not defined. Thus, we define a maximum value of 20 when computing the mean and standard deviation and denote these cases by '$>$ 20' in Table I.

### A. Discussion

Regarding the results represented in Table I and in the boxplots of Fig. 3, we noticed that there is a clear difference between the proposed bilinear MFS models and the baseline. The values for the AUC, Specificity, and PPV obtained for the proposed models exceed substantially the baseline. The range of the values showed in the boxplots for each metric reveals that the baseline has higher variability throughout each fold, specially for the Specificity.

Regarding $\mathscr{B}_1$ $\mathscr{B}_2$ we can see that the boxplots are similar for all the metrics, but the values for the likelihood ratio displayed in Table I are better for $\mathscr{B}_2$, except for the fold 3 and the LR- in fold 2.

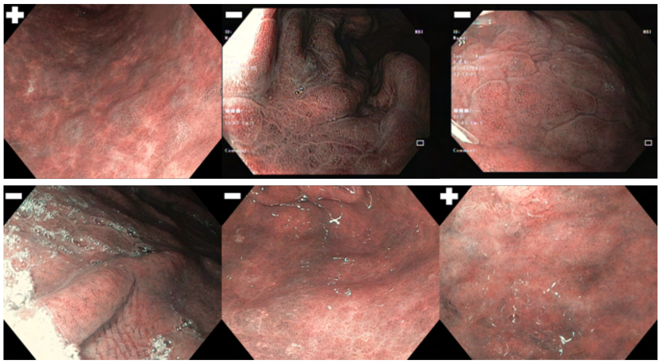


Fig. 4: Examples of images failed by the baseline and not by the other two models (top row) and the opposite (bottom row). The true label of each image is represented in the top left corner with a + if it has intestinal metaplasia and with − if it's normal.

In order to understand better the behavior of our approaches, we verified which images our models fail and the baseline does not, and vice-versa. An example of this is represented in Fig. 4. We observe that the baseline tends to classify as positive several images containing irregular texture patterns, even when those do not represent intestinal metaplasia (see the two negative images in top row of Fig. 4), and fails in images without zoom (positive image in top row of Fig. 4). Regarding the proposed approaches we noticed that they fail in images influenced by a challenging perspective (second and third image of the bottom row of Fig. 4) that are hard to identify even to a human operator, and normal images in which the presence of noise induce the model to identify as intestinal metaplasia (first image of the bottom row of Fig. 4). We hypothesize that if we used a finer grid the results would be better since MFS will be more precise in characterizing the lesion.

This study has three main limitations. Firstly, the dimensions and level of overlap of the patches extracted from the images to calculate the MFS spectra have not been optimized for this problem. The choice of these parameters was dictated by the dimensions of the embeddings obtained with the VGG-16 network to compute outer products. Secondly, the images in the GIM-NBI dataset were only annotated by a single expert. Finally, this dataset has no patient information, hence a per-patient analysis was not possible.

## IV. CONCLUSIONS

In this work, a hybrid deep learning network including explicit fractal descriptors is applied to detect GIM for endoscopic imaging data. The superior results obtained with the proposed approach showed that this inductive bias based into a DNN achieves reliable GIM detection, even with a reduced amount of training samples.

Future work will focus on optimizing trade-off between complexity and resolution regulated by the patch size adopted in the model. Moreover, end-to-end training of fractal encoders similar to those proposed by Xu *et al.* [17] will be considered.

## References

[1] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray, "Global cancer observatory: cancer today," *Lyon, France: international agency for research on cancer*, vol. 3, no. 20, p. 2019, 2018.

[2] R. Bisschops, M. Areia, E. Coron, D. Dobru, B. Kaskas, R. Kuvaev, O. Pech, K. Ragunath, B. Weusten, P. Familiari, *et al.*, "Performance measures for upper gastrointestinal endoscopy: a european society of gastrointestinal endoscopy (ESGE) quality improvement initiative," *Endoscopy*, vol. 48, no. 09, pp. 843–864, 2016.

[3] P. Pimentel-Nunes, D. Libânio, R. Marcos-Pinto, M. Areia, M. Leja, G. Esposito, M. Garrido, I. Kikuste, F. Megraud, T. Matysiak-Budnik, *et al.*, "Management of epithelial precancerous conditions and lesions in the stomach (maps ii): European society of gastrointestinal endoscopy (ESGE), european helicobacter and microbiota study group (EHMSG), european society of pathology (ESP), and sociedade portuguesa de endoscopia digestiva (SPED) guideline update 2019," *Endoscopy*, vol. 51, no. 04, pp. 365–388, 2019.

[4] S. Menon and N. Trudgill, "How commonly is upper gastrointestinal cancer missed at endoscopy? a meta-analysis," *Endoscopy International Open*, vol. 02, no. 02, 2014.

[5] D. Chen, L. Wu, Y. Li, J. Zhang, J. Liu, L. Huang, X. Jiang, X. Huang, G. Mu, S. 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*, vol. 91, no. 2, pp. 332–339, 2020.

[6] J. Arribas, G. Antonelli, L. Frazzoni, L. Fuccio, A. Ebigbo, F. van der Sommen, N. Ghatwary, C. Palm, M. Coimbra, F. Renna, *et al.*, "Standalone performance of artificial intelligence for upper gi neoplasia: a meta-analysis," *Gut*, vol. 70, no. 8, pp. 1458–1468, 2021.

[7] L. Wu, J. Wang, X. He, Y. Zhu, X. Jiang, Y. Chen, Y. Wang, L. Huang, R. Shang, Z. 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*, vol. 95, no. 1, pp. 92–104, 2022.

[8] T. Yan, P. K. Wong, I. C. Choi, C.-M. Vong, and H. H. Yu, "Intelligent diagnosis of gastric intestinal metaplasia based on convolutional neural network and limited number of endoscopic images," *Computers in biology and medicine*, vol. 126, p. 104026, 2020.

[9] Y. Xu, H. Ji, and C. Fermüller, "Viewpoint invariant texture description using fractal analysis," *International Journal of Computer Vision*, vol. 83, no. 1, p. 85–100, 2009.

[10] A. P. Pentland, "Fractal-based description of natural scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, p. 661–674, 1984.

[11] M. Häfner, T. Tamaki, S. Tanaka, A. Uhl, G. Wimmer, and S. Yoshida, "Local fractal dimension based approaches for colonic polyp classification," *Medical Image Analysis*, vol. 26, no. 1, p. 92–107, 2015.

[12] M. Martins, M. Pedroso, D. Libânio, M. Dinis-Ribiero, M. Coimbra, and F. Renna, "Diagnostic performance of deep learning models for gastric intestinal metaplasia detection in narrow-band images," 2023. Available online: https://www.techrxiv.org/articles/preprint/Diagnostic_Performance_of_Deep_Learning_Models_for_Gastric_Intestinal_Metaplasia_Detection_in_Narrow-band_Images/22770146.

[13] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1449–1457, 2015.

[14] M. Xu, W. Zhou, L. Wu, J. Zhang, J. Wang, G. Mu, X. Huang, Y. Li, J. Yuan, Z. Zeng, *et al.*, "Artificial intelligence in the diagnosis of gastric precancerous conditions by image-enhanced endoscopy: a multicenter, diagnostic study (with video)," *Gastrointestinal Endoscopy*, vol. 94, no. 3, pp. 540–548, 2021.

[15] K. J. Falconer, *Fractal geometry: Mathematical foundations and applications*. Wiley, 2014.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[17] Y. Xu, F. Li, Z. Chen, J. Liang, and Y. Quan, "Encoding spatial distribution of convolutional features for texture representation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22732–22744, 2021.