UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

AUXÍLIO AO DIAGNÓSTICO AUTOMÁTICO DO ESÔFAGO DE BARRETT UTILIZANDO APRENDIZADO DE MÁQUINA

Computer-assisted diagnosis of Barretts's esophagus using machine learning techniques

Luis Antonio de Souza Júnior

Orientador: João Paulo Papa

Co-orientador: Christoph Palm

São Carlos – SP
Março/2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Luis Antonio de Souza Júnior

AUXÍLIO AO DIAGNÓSTICO AUTOMÁTICO DO ESÔFAGO DE BARRETT UTILIZANDO APRENDIZADO DE MÁQUINA

Computer-assisted diagnosis of Barretts's esophagus using machine learning techniques.

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação, área de concentração: Metodologia e Técnicas de Computação.
Orientador: Prof. Dr. João Paulo Papa.

Beneficiário

Orientador

São Carlos – SP
Março/2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Defesa de Tese de Doutorado do candidato Luis Antonio de Souza Júnior, realizada em 28/03/2022.

Comissão Julgadora:

Prof. Dr. João Paulo Papa (UFSCar)

Prof. Dr. Ricardo Cerri (UFSCar)

Prof. Dr. Alexandre Luis Magalhães Levada (UFSCar)

Prof. Dr. Aparecido Nilceu Marana (UNESP)

Prof. Dr. Lucas Alexandre Ramos (NHL)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.
O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

AGRADECIMENTOS

Agradeço a todos os membros da minha família pelo apoio ímpar e carinho durante o período de pós-graduação, e em especial aos meus avós *Márcia* e *Dito* e a minha mãe *Josoaine*.

Agradeço aos meus amigos do ReMIC que me proporcionaram momentos felizes e inesquecíveis: *Summye*, *Jason* e *Danilo*. Agradeço imensamento ao meu melhor amigo que compartilhou todos os momentos de vida dos quais me recordo, *Lucas Alexandre Ramos*.

Agradeço aos membros do **Recogna**[1] e também do **ReMIC**[2], meus grupos de pesquisa que muito me ajudaram durante as necessidades relacionadas a minha pesquisa.

Por fim, gostaria de registrar um agradecimento especial aos professores que me orientaram, guiaram e inspiraram minha evolução acadêmica e pessoal. Muitíssimo obrigado professores *João Paulo Papa*, *Aparecido Nilceu Marana* e *Christoph Palm*.

[1] http://www.recogna.tech/
[2] https://re-mic.de/

RESUMO

O câncer no esôfago é uma doença de difícil detecção nos estágios iniciais, especialmente na presença do esôfago de Barrett. O desenvolvimento de sistemas automáticos de avaliação de tal doença podem ser muito úteis, auxiliando os especialistas na detecção da região cancerígena. Com o forte crescimento das técnicas de aprendizado de máquina e, visando melhorar a eficácia do diagnóstico médico, seu uso caracteriza um cenário forte a ser explorado para o diagnóstico precoce do adenocarcinoma de esôfago. O esôfago de Barrett como antecessor do adenocarcinoma pode ser explicado por alguns fatores de risco, como obesidade, tabagismo e diagnóstico médico tardío. Este projeto visa o desenvolvimento de novas técnicas de visão computacional e aprendizado de máquina para o auxílio do diagnóstico automático de câncer esofageal baseado na avaliação de dois tipos de características: (i) extraídas a mão (*handcrafted features*), calculadas com base no conhecimento humano usando técnicas de processamento de imagens e (ii) extraídas por aprendizado em profundidade (*deeply-learnable features*), calculadas exclusivamente com base em técnicas de aprendizado em profundidade. Pela extensa aplicação de protocolos globais e locais para os modelos propostos neste trabalho, a descrição de imagens acometidas por câncer e esôfago de Barrett foram generalizadas e profundamente avaliadas utilizando, por exemplo, classificadores como *Support Vector Machines*, ResNet-50 e a combinação de descrições por *handcrafted* e *deeply-learnable features*. Ainda, observou-se o comportamento da definição automática dos pontos de interesse dentro das técnicas avaliadas, algo de suma importância nos dias atuais para garantir transparência e confiabilidade as decisões tomadas por técnicas computacionais. Assim, este projeto contribui com ambas as aŕeas computaonal e médica, introduzindo novos classificadores, abordagens e interpretação do processo de generalização das classes, além de propor formas precisas e rápidas de definir câncer, entregando resultados importantes e de caráter inovador no que permeia a acurada identificação de câncer em amostras acometidas por Barrett, com valores que aproximam-se de taxas de 95% de correta identificação, e dispostos em uma coletânea de trabalhos científicos elaborados pelo autor durante o período de pesquisa e submetidos/publicados até a presente data.

Palavras-chave: Aprendizado de Máquina, Esôfago de Barrett, Aprendizado em profundidade, *Handcrafted features*, *Deeply-learnable features*, Redes Neurais Convolucionais, Interpretabilidade

ABSTRACT

Esophageal adenocarcinoma is an illness that is usually hard to detect at the early stages in the presence of Barrett's esohagus. The development of automatic evaluation systems of such illness may be very useful, thus assisting the experts in the neoplastic region detection. With the strong growth of machine learning techniques aiming to improve the effectivess of medical diagnosis, the use of such approaches characterizes a strong scenario to be explored for the early diagnosis of esophageal adenocarcinoma. Barrett's esophagus as a predecessor of adenocarcinoma can be explained by some risk factors, such as obesity, smoking, and late medical diagnosis. This project proposes the development of new computer vision and machine learning techniques to assist the automatic diagnosis of the esophageal adenocarcinioma based on the evaluation of two kind of features: (i) handcrafted features, calculated by means of human knowledge using some image processing technique and; (ii) deeply-learnable features, calculated exclusively based on deep learning techniques. From the extensive application of global and local protocols for the models proposed in this work, the description of cancer-affected images and Barrett's esophagus-affected samples were generalized and deeply evaluated using, for example, classifiers such as *Support Vector Machines*, ResNet-50 and the combination of descriptions by *handcrafted* and *deeply-learnable features*. Also, the behavior of the automatic definition of key-points within the evaluated techniques was observed, something of a paramount importance nowadays to guarantee transparency and reliability in the decisions made by computational techniques. Thus, this project contributes to both the computational and medical fields, introducing new classifiers, approaches and interpretation of the class generalization process, in addition to proposing fast and precise manners to define cancer, delivering important and novel results concerning the accurate identification of cancer in samples affected by Barrett's esophagus, showing values around 95% of correct identification rates and arranged in a collection of scientific works developed by the author during the research period and submitted/published to date.

Keywords: Machine Learning, Barrett's esophagus, Deep learning, Handcrafted features, Deeply-learnable features, Convolutional Neural Networks, Interpretability.

LIST OF FIGURES

LIST OF TABLES

SUMMARY

Chapter 1

INTRODUCTION

This section introduces the main problem we aim to solve in this thesis, the detection of early cancerous tissues in Barrett's esophagus-diagnosed patients. After defining such a context, we display this thesis' objetives, hypothesis, and finally, its organization.

1.1 Context Defintion

Usually, the trivial "solution" related to the machine learning (ML) workflow process comprises four mains steps: (i) data processing, (ii) feature extraction, (iii) feature selection/transformation, and (iv) pattern recognition. Considering the evolution of the aftermentioned steps in the last decades allied to a new set of techniques based on deep learning (DL) strategies, that provide an approach that mimics the brain-behavior while processing visual information (where the extraction of different kinds of information is performed on distinct layers), the expectation of ML expansions can be considered for the following years. In the last decades, the application of ML techniques in a wide range of areas has grown exponentially, even more, the ones regarding decision-making tasks. Such tasks become of extreme interest in environments that involve large amounts of data, such as laboratory diagnosis, image, and video processing, and data mining, just to cite a few. ML techniques have been largely applied to several research areas, such as remote sensing, signal processing, speech recognition, and medicine, and others. This latter research area has benefited by the constant advances related to computer vision and artificial intelligence since more effective and efficient prognosis have helped physicians to provide a faster and more accurate diagnosis.

A field that has been receiving increased attention is the early detection of cancer cells, given the incidence of cancer-related diseases has significantly grown in recent years. Among the most prominent diseases, we can mention lung, breast, and skin cancer. Another type that

has shown increased incidence and requires attention concerns esophageal and stomach cancer, which can be detected through endoscopy. However, only experienced professionals are able to operate the examination and subsequently detect cancerous cells.

The condition in which columnar cells replace squamous cells in the esophagus mucosa is known as “Barrett’s esophagus" (BE). This condition is recognized as a complication of gastrointestinal reflux, and in some extreme cases, may progress and evolve into esophageal adenocarcinoma (EAC) [Hopkins 2008, Dent 2011]. The early detection of neoplasia based on computer-aided techniques can help gastroenterologists with systems capable of visualizing and alerting about possible dysplasia in the esophagus [van der Sommen et al. 2016]. Some algorithms have been proposed for the detection of abnormal cancers, tumors or patterns in the esophagus region, but the automatic dysplasia detection can only be performed using endoscopic images of high quality and resolution, which are previously subjected to preprocessing techniques [Dent 2011]. The use of dysplastic regions represented by endoscopic imaging should be described based on texture or color criteria concerning classification purposes [Dent 2011].

From works that evaluated cancerous conditions in BE-diagnosed samples, the description based on handcrafted features that employs, for instance, object detection techniques as Speed-Up Robust Features (SURF) [Bay et al. 2008] and de description based on deeply-learnable features obtained from deep learning models [Mendel et al. 2017] highlight the current emphasis on computer-aided diagnostics, in a wide application to improve the automatic identification of early esophageal adenocarcinima.To cite a few, recent works proposed by Souza Jr.*et al.* [Souza Jr. et al. 2018, Souza Jr. et al. 2019, Souza Jr. et al. 2017, Souza Jr. et al. 2018, Souza Jr. et al. 2017] and Mendel [Mendel et al. 2017] can be highlighted, which exemplify the use of injured region description of the esophagus using artificial intelligence techniques for classification purposes. Concerning the works proposed by Souza Jr. *et al.* [Souza Jr. et al. 2018, Souza Jr. et al. 2019, Souza Jr. et al. 2017, Souza Jr. et al. 2018, Souza Jr. et al. 2017], the authors deeply studied the use of image representation techniques to describe and classify the adenocarcinoma in Barrett’s esophagus regions. For such works the use of feature extraction techniques such as SURF and Scale-Invariant Feature Transform (SIFT) [Lowe 2004], and classifiers such as Support Vector Machines (SVM), Optimum-Path Forest (OPF) [Papa, Falcão e Suzuki 2009] were considered in order to provide the class prediction of both injured regions. Mendel *et al.* [Mendel et al. 2017] introduced the use of deep learning techniques for the classification of expert annotated images of the esophagus presenting adenocarcinoma and Barrett’s esophagus, where Convolutional Neural Network (CNN) was adapted to the set of images by a learning transfer approach in a leave-one-patient-out cross-validation (LOPO-CV) protocol, and trending several works to come in which the esophagus regions would be descriped by deep learning

models, methods and architectures to cope the task of early esophageal cancer detection.

One of the major constraints regarding the use of ML for prognosis-assisted systems stands on the cancerous region definition. This step can differ among specialists during the manual delineation of Barrett's esophagus and adenocarcinoma, making such task not straightforward. Considering that the human knowledge may be used for the ML techniques to produce computational knowledge, the correct delineation provided by the computer may be threatened because of this high intra-observer variability. In addition, it is important to understand and provide robust ways of adenocarcinoma definition in BE-diagnosed images, given the bottleneck related to it. Moreover, when dealing with deep learning approaches for CAD systems, it is important not only to observe how high are the classification rates but how human-interpretative they are. It is well-known that deep learning techniques can achieve promising results for a wide range of classification fields, including the medical one. However, dealing with diagnosis, and especially with cancer diagnosis, it is important to show up which areas belonging to the learning process are discriminative for the deep learning architecture. Furthermore, lightening up the learning process attached to the deep learning model generalization provides important insights about discriminative regions that may not be observed by the experts, increasing even more the learning process related to the cancerous definition, making it a human-and-computational hybrid learninig.

1.2 Research Hypotheses

The present work's main hypothesis is: **The definition of early adenocarcinoma in tissues already ill with BE is a hard task to be accomplished, presenting high inter-observer variability. Due to that, the manual delineation of esophageal cancer lacks precision and demands time from experts, still presenting high human-dependent results. With the advances of ML in the medical diagnosis assistance, we believe that the combination of deeply-learnable features, obtained from deep learning models, and handcrafted features, computed based on experts' insights about the cancerous tissues, may not only enhance the correct identification of esophageal cancer but will also carry interpretable information regarding the proper cancer observation, once in our perspective, such representations present complementary information for the context description.** Then, from this main hypothesis, two generic ones can be translated to the sub-areas of description we aim to evaluate, the handcrafted features description and the deeply-learnable features description. Hence, from our main hypothesis, two specific ones are proposed:

- The application of handcrafted features present relevant information in the spatial description of cancerous tissues. Due to that, we believe that handcrafted features' description aggregates the experts' insights of the regional impact in the correct definition of adenocarcinoma in BE images. Also, considering the non-linear nature we believe belongs to the distinction of cancerous and non-cancerous tissues, handcrafted features based on techniques to highlight such a non-linearity behavior may benefit the description we aim to propose.

- The cancer description based on deeply-learnable features presents a powerful tool for enhancing the correct identification of early-esophageal cancer, considering the high generalization that deep models propose. Therefore, we believe that deeply-learnable features, computed from deep models we aim to employ, can bring relevant information that, in proper ways of interpretation, can assist the correct classification of cancerous samples in BE-diagnosed patients. Such a description may encode crucial information in a global representation that possibly enhances the correct classification of cancerous tissues in BE samples, and hence, we aim to employ such a description to benefit the methods we propose.

To fulfill such assumptions, this thesis aims at answering the following question: which strategies could one adopt towards enhancing early adenocarcinoma detection in Barrett's esophagus diagnosed images? Two approaches are proposed to accomplish such task, based on the aforementioned hypothesis: (i) the application of handcrafted features techniques, based on the knowledge provided by the expert physicians, and (ii) the application of deep learning techniques based on fully-automated defined models for such problem, without the use of any human annotation. In addition, it is imperative to define the best association among description and classification techniques, aiming to improve the early detection of esophageal cancer, for assessing our main hypothesis.

Once our objective is the evaluation of handcrafted features, deeply-learnable features, and proper ways of modeling such descriptions for improving correct and interpretable ways of detecting esophageal cancer in BE images, Figure 1.1 illustrates the pipeline adopted for developing this entire research.

1.3 Thesis' Organization

This research study is composed of a collection of works published/submitted by authors during the study period. The works presented in the next sections aim towards the description

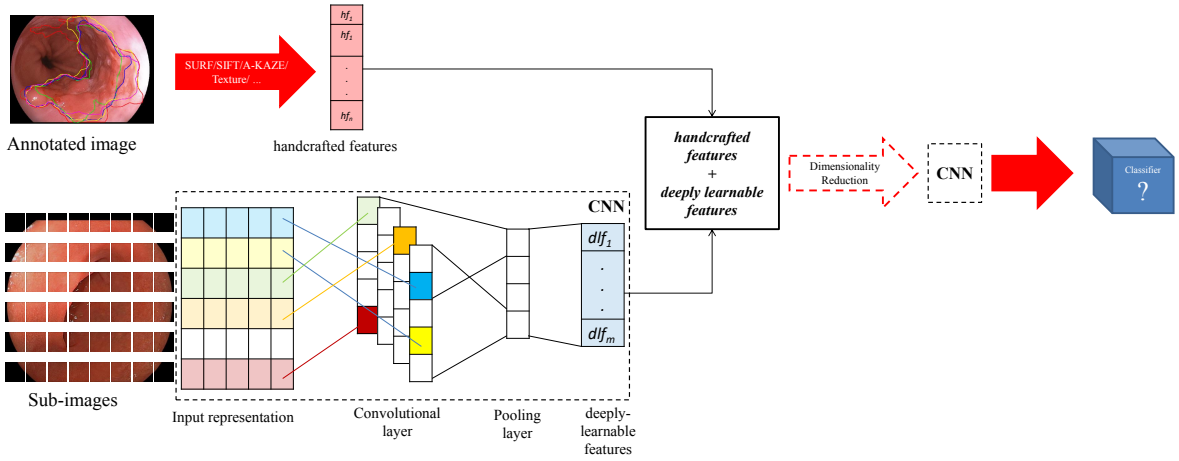


Figure 1.1: Research's overview: the upper part of the pipeline represents the work developed to evaluate the handcrafted feature's application, in Chapters 3 to 7, while the bottom flow is related to the evaluation of deep learning methods to cope with the description, generalization, and interpretation of BE and adenocarcinoma context, detailed in Chapters 8 to 11. Finally, the entire pipeline describes the combination of achievements obtained since the beginning of the research, illustrating the work conducted in Chapter 12.

and classification of adenocarcinoma and BE tissues based on two approaches: the association of several handcrafted features and classifiers and the association of several deep learning techniques. Besides the presented in the following, the author published: (i) Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma [Ebigbo et al. 2019], in which a computer-aided diagnosis method was proposed using deep learning as an instrument to improve endoscopic assessment of BE and early oesophageal adenocarcinoma; (ii) Real-time use of artificial intelligence in the evaluation of cancer in Barrett's oesophagus [Ebigbo et al. 2020], in which a computer-aided diagnosis method was proposed using deep learning as an instrument to improve endoscopic assessment of BE and early esophageal adenocarcinoma in endoscopic video recordings, and (iii) Semi-Supervised Segmentation based on Error-Correcting Supervision [Mendel et al. 2020], that proposed a method to bridge the gap between supervised and unsupervised learning by the application of Generative Adversarial Networks (GAN) concepts [Goodfellow et al. 2014] with deep models for enhancing the semantic segmentation.

Hence, considering this thesis' content, Chapter 2 presents a meticulous referential background, based on a survey, regarding the use of machine learning techniques to assist the evaluation and prognosis of EAC in endoscopic images. The paper presented in Chapter 3 evaluates the problem of EAC and BE differentiation based on SURF features and SVM classifier. The OPF classifier and SIFT description are introduced for the very first time into the BE and EAC context evaluation in the paper presented in Chapter 4, in a (BoVW) approach, while Chapter 5

introduces to the BE and adenocarcinoma context an evaluation based on color co-occurrence matrices. The work validates three approaches of classification based on patches, patients, and images in two datasets (MICCAI 2015 and Augsburg) using the color-and-texture descriptors and OPF, SVM, and Bayesian classifiers.

A continuation of the work presented in Chapter 4 is provided in Chapter 6. The work introduces the unsupervised OPF classifier for learning visual dictionaries in the context of BE and adenocarcinoma automatic diagnosis, validated over MICCAI 2015 and Augsburg datasets. Finally, Chapter 7 presents an application of handcrafted features standalone evaluation employing infinite Restricted Boltzmann Machines (iRBM) for Barret's Esophagus lesions detection.

Evaluating the deeply-learnable features techniques from Chapter 8 on, a further evaluation of GAN concepts [Goodfellow et al. 2014] was performed in the paper for increasing the amount of samples to be classified employing deep models. Chapter 9 presents the introduction of meta-heuristically fine-tuned Generative Adversarial Networks to the context of Barrett's esophagus identification and investigates its feasibility to generate high-quality synthetic images from the esophagus for further assisting the identification of the disease. To provide a quantitative interpretation of CNN learning, Chapter 10 presents a method to interpret such learning based on Explainable Artificial Intelligence (XAI) techniques and compare it with the human knowledge provided by the ground truth's annotation of cancerous samples in BE endoscopic images. Further XAI investigation was conducted in Chapter 11, evaluating the impact of the ResNet-50 deep-convolutional design for Barrett's esophagus and adenocarcinoma classification by proposing a two-step learning technique.

To combine all the achievements obtained in the evaluation of handcrafted features and deeply-learnable features from previous chapters, Chapter 12 propose a novel method, named DeepCraftFuse that encodes spatial and global information provided by the aforementioned descriptions. To close this thesis, Chapter 13 provides the conclusions, as well as contributions of this work.

To close this thesis, Chapter 13 provides the conclusions, as well as contributions of this work.

Chapter 2

A Survey on Barrett's Esphagus Analysis Using Machine Learning

This chapter introduces Barrett's esophagus and adenocarcinoma context by means of the paper published in Computers in Biology and Medicine journal [Souza Jr. et al. 2018]. The scope of the work is to compile some works published at some well-established databases, such as Science Direct, IEEEXplore, PubMed, Plos One, Multidisciplinary Digital Publishing Institute (MDPI), Association for Computing Machinery (ACM), Springer, and Hindawi Publishing Corporation. Each selected work has been analyzed to present its objective, methodology, and results.

2.1 Introduction

The adenocarcinoma appearance in BE diagnosed patients has increased significantly in western populations. This is mainly explained by obesity, a known risk factor [Lagergren e Lagergren 2010, Dent 2011, Lepage, Rachet e Jooste 2008]. As such, the expectation of this disease to rise in the next years must be considered. The bad prognosis for patients suffering from esophageal adenocarcinoma is related to its late diagnosis. However, when detected at the early stages, the dysplastic tissue can be treated with very successful rates of handling the disease, such as 5% of morbidity and 0% of mortality. Additionally, 93% of patients featured a complete remission of the disease after 10 years treatment [Dent 2011, Sharma et al. 2016, Phoa et al. 2016]. Developments in interventional therapies, such as endoscopic resection and ablation techniques (radiofrequency ablation, cryoablation) are promising methods for the management of BE, with the potential of reducing the cancer risk in dysplasia diagnosed patients. However, there are limitations of the currently accepted methods for monitoring and evaluating the dise-

ase state of BE patients, with the benefit from early diagnosis and additional tools to improve the detection of dysplasia [Shaheen et al. 2009, Johnston et al. 2005, Overholt, Panjehpour e Halberg 2003].

Several endoscopic technologies for image enhancement, such as chromoendoscopy, electronic image enhancement, optical coherence tomography, and confocal laser endomicroscopy have been developed for BE evaluation, enabling endoscopists to conduct a more accurate assessment of the dysplasia with an *in vivo* characterization of esophageal histology [Sharma et al. 2015]. This ability could result in improvements concerning the detection of BE (screening), detection of dysplasia based on BE surveillance, characterizing abnormalities within BE (selecting lesions and delineating margins during endoscopic therapy), and detection of recurrent neoplasia in patients who have received endoscopic treatment (post-treatment surveillance) [Sharma et al. 2015].

BE is often misdiagnosed during endoscopy because of: (1) inability to differentiate columnar mucosa of the proximal stomach (cardia) from metaplastic epithelium in the distal esophagus; or (2) lack of goblet cells in biopsies obtained from columnar lined epithelium in the esophagus. Since dysplasia/BE areas are sometimes not readily perceived with standard white-light endoscopy, the Seattle biopsy protocol is usually recommended, where biopsies are taken for every 1 cm of the BE's mucosa. However, this protocol may be susceptible to sampling errors because only a small part of the entire BE mucosa is usually considered for sampling purposes, especially in patients with extensive disease area [Sharma et al. 2015]. Besides, the biopsy protocol can be costly and time-consuming, and thus prone to errors. Consequently, the risk of missed dysplasia or cancer diagnosis rises significantly [Abrams et al. 2009]. Other studies have also considered the classification of different esophagus' lesion types based on color and texture information of the injured tissue area as well [van der Sommen et al. 2014].

Machine learning techniques have benefited from significant improvements in image analysis and artificial intelligence fields. However, related to the automated analysis of BE, we observed one recent work only that attempted to compile relevant articles [Ghatwary, Ahmed e Ye 2017]. This work is a very brief survey to discuss advances in BE Computer-assisted diagnosis (CAD) systems in three endoscopy modalities used for esophageal examination: (i) white light endoscopy (WLE), (ii) high-definition white light endoscopy (HD-WLE), and (iii) narrow band imaging (NBI). Focusing on detection methods lately developed for BE detection, the survey is composed of eight papers about automatic detection and evaluation of the BE, compared by its endoscopy modality, number of images and evaluated classifiers applied to the problem, validation method and results. The authors state the challenges for this detection and

mention some directions for future research.

Our work aims at reviewing and investigating the feasibility and usage of machine learning techniques in the context of BE evaluation, dysplasia description, and treatment, thus providing more details to the previous brief survey. Next sections present the methodology used to evaluate the compiled articles, as well as some medical background related to the disease.

2.2 Theoretical Background

2.2.1 Barrett's Esophagus

The replacement of squamous cells by columnar cells in the esophagus' mucosa is known as BE. This process is recognized as a complication of gastroesophageal reflux disease, and in some critical stage, it can progress and evolve into esophageal cancer. Figure 5.1 illustrates the human esophagus region [Dent 2011, Sharma et al. 2016, Hopkins 2008].

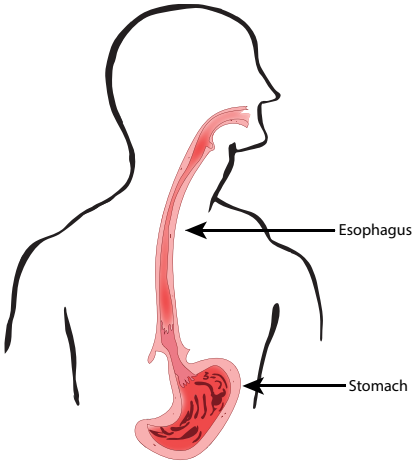


Figure 2.1: Esophagus' location in the human body.

Squamous cells (similar cells to skin or mouth ones) compose the mucosa of the normal esophagus. The normal color of the squamous mucosal surface looks like whitish-pink, while the gastric mucosa goes sharply from salmon-pink to red [Dent 2011,Sharma et al. 2016]. A demarcation line called squamocolumnar junction or "Z-line"defines the normal esophagogastric junction (Figure 2.2), where the squamous mucosa of the esophagus and the columnar mucosa of the stomach meet [Hopkins 2008]. BE's mucosa may extend upward in a continuous pattern, making the entire circumference of the distal esophagus covered by columnar mucosa. A difference is stablished among patients with more than 3 cm of BE ("long-segment BE") and those who feature the so-called "short-segment BE", with refers to BE that figures less than 3 cm, as depicted in Figure 6.4.

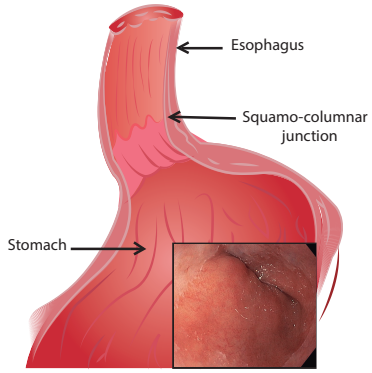


Figure 2.2: Squamo-columnar junction and its respective esophagus endoscopic image.

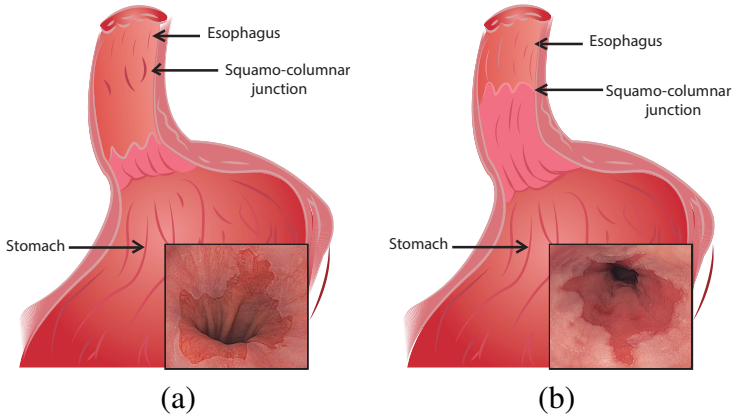


Figure 2.3: Endoscopic views from: (a) BE's short-segment and (b) BE's long-segment.

2.2.2 Machine Learning

Machine learning techniques have been paramount in the last decades mainly due to their capability in handling problems non-linearly by nature. Given a dataset composed of samples, the traditional pipeline used for so many years considers partitioning the data into training and testing sets. The former is used to learn the model (i.e., statistics of the data) meanwhile the testing set is used to assess the efficiency of the method.

Depending on the amount of knowledge we have about the training set, machine learning techniques can be categorized into three main groups: (i) supervised, (ii) semi-supervised, and (iii) unsupervised approaches. Supervised techniques usually achieve the best results since they make use of an entirely labeled training set, thus having more information to cope with. Semi-supervised learning approaches make use of both labeled and unlabeled data since only a fraction of the training data is labeled. Different approaches such as active learning-based or reinforcement learning can be referred to as well. In a nutshell, these approaches employ the user knowledge into the learning process, which can thus refine the results and correct possible errors.

Unsupervised learning or clustering stands for the group of techniques that have no information about the training data, which means they must group the data using some heuristic that can get together (i.e., same cluster) samples that share some information. To evaluate such techniques, we usually make use of measures that take into account the compactness and separability of clusters in the feature space, i.e., it is highly desirable to have well-separated and compact clusters at the end of the clustering process. Figure 9.1 depicts a standard pipeline used in applications that use machine learning for solving problems.

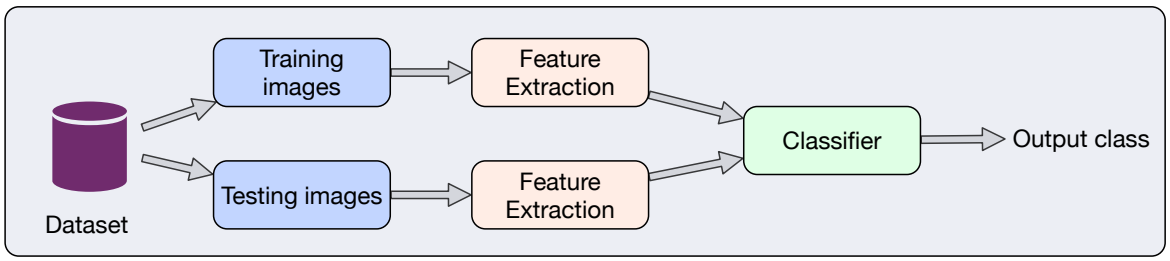


Figure 2.4: Standard pipeline used in machine learning-driven applications.

2.3 Surveyed Works

In this section, we present the works considered for further study and discussion. The next subsections describe in more in-depth details the works selected by their primary classifier employed.

2.3.1 Paper Selection

To select works within the scope addressed in this systematic review, a search in Science Direct, IEEEXplore, PubMed, Plos One, Multidisciplinary Digital Publishing Institute, Association for Computing Machinery, Springer, and Hindawi Publishing Corporation databases was carried out. To this end, only two keywords were considered for searching purposes: (i) “Barrett’s esophagus" and (ii) “Barrett’s esophagus machine learning". The main idea is to provide a fair selection of works and to cover a total of 35 recent works published as follows: before 2011 (5 works), between 2011 and 2014 (6 works), and early 2015-2017 (24 works). Also, the search returned a number of papers not related to machine learning-assisted BE analysis. Therefore, the outcoming of this survey does not consider them all.

2.3.2 Machine Learning Analysis of Barrett's Esophagus

Machine learning is a branch of computational intelligence dedicated to the development of algorithms that enable a computer program to improve its performance based on learned information. The intense research in this field has motivated a number of works that aimed at using machine learning-oriented techniques to aid BE recognition and distinction between adenocarcinoma and healthy tissue.

In this survey, the works are divided according to the classifiers they have employed to cope with BE identification. Considering the number of works that used SVM and neural networks, we decided to have a dedicated section for each them. Additionally, a section presenting the comparison of two or more classifiers is also considered, followed by a description of other techniques applied to BE identification, such as k-nearest neighbors (*k*-NN), k-statistics, and decision trees, among others.

2.3.2.1 Support Vector Machines-based Barrett's Esophagus Recognition

Li and Meng [Li e Meng 2009] presented a new texture-based protocol for ulcer regions using capsule endoscopy (CE) discrimination in endoscopic images. A novel approach based on curvelets and Local Binary Patterns (LBP) was proposed for texture extraction aiming to distinguish ulcer and normal regions. These new features are sensitive to illumination changes, multidirectional features, and feature invariance. Experiments were conducted using two different classifiers on a 4-fold cross-validation procedure: (i) a Multilayer Perceptron Neural Network (MLP) and (ii) Support Vector Machines. The database used for the experiments is private and composed of 100 images from 5 different patients. Regarding the images, $1,800$ patches of normal images and $1,800$ patches of ulcer-diagnosed images were extracted.The authors concluded the proposed textural features were suitable to identify ulcerous regions in CE images, once detection rates of the proposed features with the MLP were 92.37% of accuracy, 91.46% of specificity, and 93.28% of sensitivity.

Rodriguez-Diaz and Singh [Rodriguez-Diaz e Singh 2016] proposed a computer-based approach that employs the NICE criterion for diagnostic purposes, as well as it provides an on-the-fly interpretation of the histology of the polyps represented by near-focus narrow-band imaging (NF-NBI) images. The NICE criterion considers three main characteristics when learning the information that may be useful to deal with BE identification: color, vasculature, and surface pattern. The color information was used to encode the tone of neoplastic regions compared to non-neoplastic polyps, which appears to be brown-colored. Regarding the vessels, the authors

performed an automatic segmentation of the inter-crypt space and compared its color to the remaining tissue of the polyp to distinguish brown vessels from lighter structures around them. Finally, the authors employed the discrete wavelet transform to describe the local spatial distribution of gray levels (green and blue channels) and then characterize the neoplastic and non-neoplastic regions. Individual features were used as inputs for SVM in a leave-one-out cross validation (LOO-CV) protocol. A total of 26 patients and 56 images (16 non-neoplastic and 40 neoplastic polyps) were considered for the database composition. The classification results achieved 86% of sensitivity and specificity.

Nancarrow *et al.* [Nancarrow et al. 2011] performed a comparative study to define convincing differences between BE and EAC in biopsies from selected patients using SVM classification. A database composed of 54 biopsy specimens from 54 different patients were used for validation purposes, and certified by a pathologist (23 annotated as presenting EAC). The results concluded that BE-affected regions figure a tissue containing an enhanced glycoprotein synthesis mechanism designed to provide mucosal defenses. Such mechanism resists to gastro-esophageal reflux, while EAC exhibits the enhanced extracellular remodeling effects expected in an aggressive form of cancer. Also, evidence of reduced expression of genes associated with mucosal and xenobiotic defenses was also perceived. The authors observed eleven genes that are also represented in at least three other profiling studies used to discriminate among squamous epithelium, BE, and EAC, within the two largest cohorts using an SVM-based LOO-CV analysis. The proposed method was considered able to distinguish squamous epithelium and BE reasonably, and it also evidenced that more detailed investigations into profiling changes between BE and EAC are desired. The work mentioned above achieved the following results: sensitivity and specificity higher than 88% concerning the task of discriminating BE from squamous samples, as well as a sensitivity of around 73% when distinguishing EAC (cancer) from BE or squamous (non-cancer).

Veronese *et al.* [Veronese et al. 2013] proposed a computer-assisted approach to distinguish gastric metaplasia (GM), intestinal metaplasia (IM), and neoplasia based on the use of appearance features in confocal laser endomicroscopy images. The database was composed of CLE images obtained from consecutive 29 BE patients undergoing surveillance. In a nutshell, features are extracted based on the division of the image in sub-regions for the further application of LBP for a multiscale evaluation. The evaluation of the method was performed by the comparison of the automatic results with the histological gold standard using SVM-based classifiers. The proposed method identified IM, GM, and neoplasia in confocal images with accuracy close to the human observer. The validation protocol adopted was the LOO-CV, and the overall sensitivity results were: 96% for GM, 95% for IM and 100% for NPL.

Using the technology of High Definition (HD) endoscopy, Muldoon *et al.* [Muldoon et al. 2010] developed a CAD system to help physicians with faster prognosis and decrease the diagnosis miss rate in the context of early-stage cancer detection. The work compared several techniques for texture-oriented feature extraction, including Gabor, co-occurrence matrix features, and LBP. For a better image description, an efficient combination of color and texture features were proposed. A pre-processing step designed for endoscopy images was also considered to improve the classification accuracy. Later on, Principal Component Analysis was used to reduce the number of features for the further usage of SVM. The experimental results were validated by a gastroenterologist and showed a classification accuracy up to 96.48%, in 129 sites calculated in a database composed of 16 HMRE images.

Van der Sommen *et al.* [van der Sommen et al. 2013] presented an approach based on HD endoscopic images for the automatic esophagus irregularity identification. They employed the concept of tile-based image processing, so that the system was able to identify early cancer and also locate it in endoscopic images. The identification process was based on the following steps: (i) pre-processing, (ii) feature extraction with dimensionality reduction, and further (iii) classification. The performance detection was evaluated in RGB, HSI, and YCbCr color spaces using the Color Histogram and Gabor features in a database of HD endoscopic images obtained from 66 patients. Other well-known texture features were also considered for comparison purposes. Concerning the classification step, an SVM configured with different parameters and kernel functions were applied. The proposed approach achieved a classification accuracy of 95.9% considering tiles of tumorous and normal tissue of 50×50 pixels, with area under the curve (AUC) of 99%.

Van der Sommen *et al.* [van der Sommen et al. 2014] proposed a novel algorithm that calculates local texture and color features based on the original and Gabor-filtered images for the automatic detection of early cancer in high-definition endoscopic images. Appropriate filters based on spectral characteristics of the cancerous regions were designed, and post-processing techniques were further applied to annotate the injured regions and the features were extracted and classified by a trained SVM classifier. For seven evaluated patients, the experiments compared 32 annotations performed by the algorithm with the corresponding annotations made by a gastroenterologist expert using a LOO-CV protocol. From 38 lesions highlighted independently by the gastroenterologist, 36 of those lesions with a recall of 95% and precision of 75% were correctly detected by the system.

Hassan and Haque [Hassan e Haque 2015] proposed a real-time and computationally efficient bleeding detection technique using wireless capsule endoscopy (WCE) technology.

The technique was based on the observation of characteristic patterns present in the frequency spectrum of WCE frames. After these patterns have been defined, the authors developed a texture-based feature descriptor that operates on the Normalized Gray Level Co-occurrence Matrix (NGLCM) in the magnitude spectrum of the images. This descriptor was called Difference Average. The proposed algorithm was validated using a WCE database; the SVM training set was composed of 600 bleeding and 600 non-bleeding frames. Additionally, 860 bleeding and 860 non-bleeding images were chosen from the remaining images to compose the test set. The accuracy, sensitivity, and specificity values achieved were 99.19%, 99.41%, and 98.95%, respectively. The proposed method requires a low computational cost, thus making it suitable for real-time implementations.

Souza Jr. *et al.* [Souza Jr. et al. 2017] conducted a study to test the feasibility of adenocarcinoma classification in endoscopic images. The 2016 Endovis Challenge database [MICCAI 2015: 18th International Conference 2015] was used for the further extraction of SURF [Bay et al. 2008], which were employed together with SVM using the LOPO-CV protocol for training and testing purposes. Two classes composed the problem: non-cancerous- and cancerous-annotated images. Two approaches for feature extraction and classification were carried out: using the full images and using the expert-annotated regions of the adenocarcinoma. The results for the "full images approach" were 77% of sensitivity and 82% of specificity. For the "regions approach", the results were 90% of sensitivity and 95% of specificity.

Zhang *et al.* [Zhang et al. 2016] conducted a study using endoscopic ultrasonography (EUS) to calculate textural features in a spectral analysis of pixels to provide a quantification of early esophageal carcinoma tissue. A database composed of $1,210$ EUS examination samples was used from 66 patients with early esophageal cancer and 91 without cancer. The textural features of the EUS images were represented as a graph, in which the pixels are the nodes and the similarity between the gray-level or local features of the images are the edges. The similarity were provided by a high-order graph matching of the texture features. Finally, a 10-fold cross-validation approach was considered, and an SVM classifier was applied to calculate the optimal prediction of the esophageal carcinoma samples represented in the graph. As the primary results, the authors obtained 93% of accuracy in the prediction of early esophageal carcinoma, normal and leiomyoma tissues. Considering only the early esophageal carcinoma prediction, the average results of accuracy, sensitivity, specificity and negative prediction were 89.4%, 94%, 95%, and 97%, respectively.

Klomp [Klomp et al. 2017] explored the feasibility in the use of computer vision techniques to correctly predict the presence of dysplastic tissue in VLE images. Three new features based

on the classic Haralick features were proposed, and the SVM classifier was applied in a dataset composed of 30 dysplastic BE images and 30 non-dysplastic BE images. Using a 10-fold cross-validation protocol, the authors obtained an area under the Receiver Operating Characteristic (ROC) curve as of 0.95 compared to the 0.81 achieved by the clinical prediction model.

2.3.2.2 Neural Network-based Barrett's Esophagus Recognition

Seguí *et al.* [Seguí et al. 2016] introduced a system for small intestine motility identification based on Deep CNN to avoid the time-consuming step of specifying features for each motility event. This study aimed to help physicians with the diagnosis performed by the WCE video screening. Concerning the network training, $100,000$ annotated WCE samples were used, and the remaining $10,000$ samples were employed for testing purposes. The experimental results evidenced the robustness of the new features over others designed using state-of-the-art handcrafted approaches. In particular, the proposed approach obtained a mean accuracy of 96% for six intestinal motility events. Such result allowed the proposed approach to outperform other classifiers trained with classic handcrafted features (a 14% relative performance increase was observed).

Mendel *et al.* [Mendel et al. 2017] carried out a work in which deep learning was applied in specialist-annotated images containing adenocarcinoma and BE's disease. A dataset provided by the 2016 Endovis Challenge [MICCAI 2015: 18th International Conference 2015] was used for the experimental step, and it comprises 100 annotated endoscopic images (50 presenting BE and 50 presenting adenocarcinoma) from 39 patients (17 not presenting adenocarcinoma and 22 presenting adenocarcinoma). The convolutional neural network was adapted to the set of images by a transfer learning approach in a LOPO-CV. With positive results of sensitivity and specificity (94% and 88% respectively), the study demonstrated that it is possible to extend its results to the BE's esophageal segmentation domain itself using deep learning to reach the region affected by adenocarcinoma specifically.

To cope with the problem of time-consuming and cost-inefficient manually ground-truth definition, Georgakopoulos *et al.* [Georgakopoulos et al. 2016] proposed a weakly-supervised learning technique based on CNN that uses only image-level semantic annotations for the training process, instead of using annotations at the pixel's level. The performance of the proposed method was evaluated in the context of CAD system of inflammatory gastrointestinal lesions represented in WCE videos. The results showed the proposed method could be more accurate than the conventional supervised learning with an accuracy of around 90% obtained in a previous proposed data set proposed by [Koulaouzidis e Iakovidis 2015].

Chan *et al.* [Chan et al. 2016] used an e-nose, a device that utilizes chemical-to-electrical interfaces, to measure the volatile organic compounds (VOC) profiles of disease states. When paired with a machine learning platform, an e-nose can be trained as a canine to serve as a tool for noninvasive diagnostic testing. Such approach was used to perform a cross-sectional study evaluating the breath VOCs of a cohort of 112 patients (66 with BE and 56 without BE) with a history of dysplastic BE to differentiate the differences in BE by dysplasia grade. These VOC profiles were introduced into an artificial neural network in a supervised step to identify data classifiers to discriminate differences in subjects stratified by the presence or absence of BE by biopsies. Optimal models were validated using a leave-some-out cross-validation (LSO-CV) approach to generate performance characteristics of BE detection in a CNN classification. The sensitivity and specificity was 82% and 80%, respectively, the accuracy was 81%, and the AUC was 79%. The task of analysis and interpretation of WCE records is complex and require sophisticated CAD systems to assist physicians in the video screening and further diagnosis. Because most of the capsule endoscopy CAD systems share a standard design, each time a new clinical application of WCE appears, a new CAD system has to be structured from scratch.

Yoshida *et al.* [Yoshida et al. 2017] performed a study to evaluate the classification accuracy of gastric biopsy specimens using the *e-Pathologist* image software. A dataset composed of 3062 gastric-biopsy specimen slides were used, being each one evaluated by at least two experts to provide the diagnosis. Finally, the comparison was performed between the experts and the e-Pathologist classification results. A cross-validation protocol was used together with a neural network, which achieved a recognition rate as of 55.6% over a three-class problem: (i) positive for carcinoma, (ii) caution for adenoma, and (iii) negative for a neoplastic lesion. An additional experiment was carried out in a two-class problem: (i) negative for neoplastic regions, and (ii) positive for neoplastic areas. In this analysis, the sensitivity, specificity, and negative predictive value were 89.5%, 50.7%, and 90.6%, respectively, showing a promising direction for the automated classification of injured regions of the intestine.

Hong *et al.* [Hong, Park e Park 2017] developed a CAD system to classify endomicroscopy images between gastric metaplasia, intestinal metaplasia, and neoplasia (these last two are subclasses of BE). A database provided by ISBI 2016 challenge and composed of 155 gastric metaplasia instances, 26 intestinal metaplasia instances and 55 neoplastic samples was considered for experimental purposes together with Convolutional Neural Networks in a cross-validation protocol. The training data were distorted for augmentation purposes as well, providing an accuracy as of 80.77%, thus suggesting that CNN might be useful to this kind of problem.

2.3.2.3 Comparison among Classifiers for Barrett's Esophagus Recognition

Rajan *et al.* [Rajan et al. 2010] performed a comparison experiment using several classifiers, such as SVM, *k*-NN, and Boosting on images from different endoscopy modalities (WLE, NBI, Chromoendoscopy). The datasets (125 WLE images, 122 NBI images, and 150 Chromoendoscopy images) have been classified between four categories: Normal Squamous, Gastric Mucosa, BE, and High-grade dysplasia (adenocarcinoma). The classification step was performed using features (i.e., color and texture) obtained from the injured regions of the endoscopic images in a cross-validation protocol. The accuracy ranged from 36.36% up to 89.17% according to the endoscopy modality images and classifier applied.

Considering the use of vibrational spectroscopy for the diagnosis and staging of cancer, Sattlecker *et al.* [Sattlecker, Stone e Bessant 2014] conducted a study aiming to corroborate the many promising benefits in the current histopathology methods used in the context of BE identification. To correlate complex multivariate spectral and the disease level, the authors applied machine learning methods, such as SVM, linear discriminant analysis (LDA), artificial neural networks (ANN), and Random Forests to recognize spectral patterns. The validation protocols adopted were LOO-CV, bootstrapping, and independent testing. A detailed review of related works was conducted, and the average recognition rates of the surveyed studies were around 90% of sensitivity and specificity, although the majority of the studies used less than 40 samples. The authors concluded that more studies need to be carried out in case we decide to put in practice the combination of spectroscopy and machine learning.

Kandemir *et al.* [Kandemir et al. 2014] performed a study for the diagnosis of BE presented in hematoxylin-eosin stained histopathological biopsies using multiple instance learning (MIL) and Support Vector Machines. Regarding the experiments, the database comprised 214 tissue cores (165 presenting cancer and 69 showing healthy condition) from 97 patients. Rectangular patches of the tissue cores were extracted, and a feature vector was calculated based on a large set of cell- and patch-level features (color features, texture features, and object features such as minimum, maximum, and standard deviation of the cells) for each patch. The tissue core as considered a bag (a group of instances with a unique group-level ground-truth label), while each patch was considered an instance. After many MIL approaches, the authors realized that a graph-based MIL algorithm (mi-Graph [Zhou, Sun e Li 2009]) obtained the best performance explained by its inherent suitability to bags with instances that presents spatial correlation. For patch-level diagnosis, the result was around 82% of accuracy and 89% of AUC using Bayesian logistic regression in the distinction of BE and cancer region patches.

Considering the WCE as a promising technology for gastrointestinal disease examination

in a non-invasive way, Yu *et al.* [Yu et al. 2015] studied the classification problem of the digestive organs for WCE images aiming to save the time of doctors in the image review task. Based on a previous study using CNN, a database composed of 25 real WCE recording samples (approximately 1 million of WCE images) was considered for experimental purposes, and with results nearly to 95% of accuracy. The authors also tried to improve the results by the proposition of a WCE classification system built as a hybrid CNN with an extreme learning machine (ELM). In the new approach, the CNN was designed as a data-driven feature extractor, and the cascaded ELM was designed as a strong classifier instead of the conventionally use of a deep CNN fully-connected classifier. The results showed a performance of around 97% concerning the WCE organ classification accuracy.

Swager *et al.* [Swager et al. 2017] conducted a study to identify VLE features from neoplasia areas regarding BE identification, as well as the authors aimed to develop an approach to predict VLE scores. The work used a VLE image database composed of 52 endoscopic resection specimens from 29 BE patients, which were assigned positive and negative to neoplasia. Features potentially significant for the prediction of early BE neoplasia were identified over twenty-five VLE-histology images. In a learning phase, twenty VLE images presenting or not BE neoplasia were scored by two experts blinded to histology. A prediction score was developed by the use multivariable logistic regression analysis, being validated by scoring forty VLE images (50% neoplastic) using the area under ROC curve analysis. The work identified three main VLE features that can be used to assist BE neoplasia identification: (i) lack of layering, (ii) higher surface than subsurface signal, and (iii) presence of dilated glands/ducts. The sensitivity and specificity values obtained were 83% and 71%, respectively, showing promising accuracy.

Another study conducted by Souza Jr. *et al.* [Souza Jr. et al. 2017] introduced the OPF [Papa, Falcão e Suzuki 2009, Papa et al. 2012] classifier in the adenocarcinoma and BE classification using endoscopic images. The work considered describing endoscopic images (database provided by [MICCAI 2015: 18th International Conference 2015]) using feature extractors based on key point information, such as the SURF and SIFT [Lowe 2004], for further designing a bag-of-visual-words that were used as input to both OPF and SVM classifiers in a cross-validation protocol. The OPF classification outperformed the well-known SVM, presenting better results for both feature extractors, with values lying on 73.2% (SURF) - 73.5% (SIFT) for sensitivity, 78.2% (SURF) - 80.6% (SIFT) for specificity, and 73.8% (SURF) - 73.2% (SIFT) for the accuracy.

Pu *et al.* [Pu et al. 2017] performed a study to extract cost-efficient biomarkers more efficient than the ones currently available, aiming to provide high sensitivity and specificity in the

task of esophageal squamous cell carcinoma (ESCC) diagnosis. The proposed biomarker to be evaluated was the DNA methylation, and 100 samples of ESSC DNA methylation from "The Cancer Genome Atlas"were analyzed along with a particular dataset of 12 samples of the same kind. Candidate CpG sites and their adjacent regions were defined and compared with adjacent normal tissue regions using several machine learning techniques such as Random Forest, SVM, CNN, Logistic Regression, Naive Bayes, LDA and Flexible Discriminant Analysis in a 5-fold cross-validation protocol. The sensitivity, specificity, and area under the curve results of the diagnostic model based on the combination of five genomic regions were 75%, 88%, and 85%, respectively.

Serpa-Andrade *et al.* [Serpa-Andrade et al. 2015] proposed a method in which the esophagitis (a condition of chronic BE stage) was described using Fourier Transform on the Z-line signature (esophageal irregularities) for classification purposes. The proposed descriptors were based on statical features and textural information. A database comprising 10 samples of healthy tissue and 16 samples of ill tissue was used, and a cross-validation protocol applied for classification purposes based on k-NN and Random Forests. The best average results obtained were around 81% of precision, 86% of sensitivity, and 72% of specificity.

2.3.2.4 Additional Works

In this section, we present works that make use of classification techniques other than SVM and neural networks. Zopf *et al.* [Zopf et al. 2009] proposed a study using NBI endoscopy images for the automatic detection of BE using a nearest neighbor classifier. The model extracted features from a proper database of images presenting 326 regions of interest (ROIs) annotated by experts, and classified between three classes: epithelium, cardiac mucosa, and BE. The features applied to the classification were the co-occurrence matrices, summation and difference of histogram, statistical geometric, and Gabor Filters, leading to a high-dimension vector reduced later. The evaluation has been made measuring the performance of each feature and also by combining them all, using the LOO-CV protocol and the Euclidean distance as similarity metric with the following results: accuracy between 85% and 92%. The best BE classification accuracy was around 74%.

In 2016, the BE's International NBI Group (BING) [Sharma et al. 2016] performed a study with the goal of developing a simple and reliable approach to recognize dysplasia as well as EAC in individuals affected by Barrett's esophagus. The research group analyzed 60 NBI images containing nondysplastic BE, high-grade dysplasia, and esophageal adenocarcinoma to find out vascular and mucosal patterns visible by NBI to be used as features and then creating

the BING criteria. Further, patients that were under supervision or endoscopic treatment for BE were recruited at four institutions in the United States and Europe, performing histologic biopsy analysis and composing a high-quality NBI image database. Experts reviewed 50 NBI images to validate the proposed approach and then evaluated 120 additional NBI images (without previous review) to assess its prediction accuracy. The proposed method identified patients with dysplasia with 85% of overall efficiency.

Pech *et al.* [Pech et al. 2008] designed a study to evaluate the potential of endomicroscopy for predicting histology *in vivo* in patients presenting early stage of squamous cells in the esophagus. Twenty-one patients suspected to early squamous cell cancer and recommended for endoscopic therapy were included in this study, being their mucosal areas examined using confocal imaging and resulting in 43 lesion images. Each scanned lesion image was stored and *in vivo* diagnosis was performed during routine endoscopy. Biopsy specimens were extracted from every lesion. The confocal images were reviewed by two personnel blinded to the histology endoscopists. The overall accuracy using the k-statistics was 95%, and the sensitivity and specificity were 100% and 87%, respectively. Intraobserver agreement was close to perfection ($kappa = 0.95$), meanwhile the interobserver agreement was very relevant ($kappa = 0.79$).

Rosenfeld *et al.* [Rosenfeld et al. 2014] aimed at studying how data mining can be applied to assist the diagnosis of high-risk lesion patients affected by BE in HD videos. As the patient information is open to interpretation, the authors demonstrated that composite rules learned from many experts can be more accurate than that of a single expert. Such fact can be explained because even expert physicians can interpret endoscopy images differently, thus potentially making it relevant to aggregate multiple opinions for the precise interpretation of the endoscopic image. Also, the authors demonstrated that decision trees could learn simple rules to assist the dysplasia diagnosis. The authors employed two decision models in a dataset composed of 47 HD endoscopic videos of the esophagus (23 dysplastic and 24 non-dysplastic): one considering the expert decisions about dysplasia and no-dysplasia, and another without the expert decisions. The overall accuracies concerned the aforementioned models were around 79% (with the experts' decision) and 77% (without the experts' decision).

Li *et al.* [Li et al. 2015] proposed a new learning method based on the multiple instance paradigm to recognize tumor invasion of gastric cancer using computed tomography (CT) imaging. The authors extracted bag-level- and instance-level features for processing and classification purposes using a database composed of 26 patient exams. Since there might be ambiguity when assigning labels to some selected patches in instance-level features, the authors proposed an improved Citation k-nearest neighborhood (Citation-kNN) that achieved recognition rates

of around 76.92%.

Curvers and Bergman [Curvers e Bergman 2016] carried out a study with patients undergoing BE neoplastic or with a suspect of such disease. Microscopic images of the esophagus were obtained from regions with neoplasia suspicious and from other locations randomly sampled for further biopsy analysis. The database images were further identified as neoplastic or non-neoplastic by two experts (blinded for histology results) with experience in high-resolution microendoscopic (HRME) image representation of BE. A tool for the visual interpretation of HRME images was designed for the reviewers to classify each image of the dataset. Also, three endoscopists with HRME experience annotated the entire image set using the developed tool. As a result, an analysis of the HRME images was performed based on the relevant image features selected for the classification step. A sequential and automatic image classification approach was developed and trained in a separate learning set. The results of this learning phase were validated in a selected set of images. The experimental results concerning sensitivity and specificity for neoplasia were around 81% and 76%, respectively, presenting a fair interobserver agreement. The results of the quantitative image classification algorithm achieved were: sensitivity of 84% and specificity of 85% for the learning set, and sensitivity of 88% and specificity of 85% for the validation data. These results corroborate that quantitative analysis of HRME images can provide an accurate classification of neoplastic and non-neoplastic BE tissue, which can be compared to the precision showed in the assessment of experienced endoscopists.

Wang *et al.* [Wang et al. 2016] developed an approach for the automatic detection and quantification of subsquamous glandular structures (SGSs) in volumetric laser endomicroscopy (VLE) data sets using automated image processing for the radiofrequency ablation (RFA) in the decrease of the BE extension. There were considered import information of SGSs right before RFA treatment, such as the average number, size, depth, and eccentricity per cross-section, and their correlations with the reduction of maximum BE length at follow-up after RFA were evaluated. After the analysis of nine VLE volumetric datasets from seven patients, there were found strong correlations between the SGS characteristics immediately before RFA, and the change of maximum BE length at follow-up after RFA. The SGS depth and eccentricity characteristics were the most significant for the RFA outcome.

Swager *et al.* [Swager et al. 2017] investigated the effectiveness of a computer-assisted tool to identify early BE neoplasia in sixty *ex vivo* VLE image using VLE features and machine learning methods for classification purpose. The database comprises sixty BE patients (30 nondysplastic BE - NDBE- and 30 high-grade dysplasia/early adenocarcinoma images). VLE features from a clinical VLE prediction score for BE neoplasia were used to feed the proposed approach,

and novel clinically-inspired algorithm features were developed based on signal intensity statistics and grayscale correlations. The comparison was performed with generic image analysis methods for neoplasia detection. For classification purpose, several machine learning methods were evaluated, such as SVM, adaBoost, and *k*NN, allied to an LOO-CV protocol. Three novel clinically-inspired algorithm features were developed as a result of the work, presenting an area under the receiver operating characteristic curve of 95%. Corresponding sensitivity and specificity were 90% and 93%, respectively.

Boschetto *et al.* [Boschetto, Gambaretto e Grisan 2016] presented a CAD system to automate classification of normal and metaplastic endoscopic NBI images. Eight features were extracted from regions defined as clusters of superpixels, which are based on the superpixels of each region: three features are calculated as mean intensities of each color channel, three other features stand for mean intensities of the red-channel with the application of three different morphological filters (top-hat, entropy and range filters), and the last two features are related to the contrast and homogeneity of the superpixels. The classification step was performed using Random Forests in a 10-fold cross-validation approach on a dataset with 116 NBI samples. Following the feature extraction step, the samples were split into training (70% of the instances) and testing (30% of the instances) sets, and the overall accuracy, sensitivity and specificity results were 83.9%, 79.2%, and 87.3%, respectively.

2.4 Discussions and Conclusions

In the last years, the amount of people with BE has increased considerably, mostly in the western countries, turning it in a world's health problem up to date. The use of artificial intelligence and machine learning techniques showed promising results, thus becoming a major aid to cope with BE pattern prognosis.

As it can be noticed in this survey, the application of machine learning has risen in the last years, with high use of SVM, CNN, and other methods for the detection and classification of adenocarcinoma or abnormalities in the esophagus region. In light of that, research in this area becomes very relevant and important for the early, fast, and standardized detection of BE and adenocarcinoma.

In this work, we presented a review concerning BE detection and monitoring using recent studies, being its main contribution to consider very recent works dating from 2011 to 2017 mostly, with the application of machine learning and computer vision for the description and classification between BE and adenocarcinoma. Additionally, the very recent studies for the

detection, treatment, and evaluation of the BE are reviewed in this survey, and Table 2.1 presents a summarization of them all. Currently, based on the works considered in this survey, we can conclude the BE problem assisted by machine learning techniques is not mature yet, and there is a need for even more research to provide solid methods to distinguish the BE and adenocarcinoma regions in endoscopy images and videos.

We have observed the vast majority of works that use machine learning and computer vision for any BE purpose are brand new (between 2015 and 2017), thus highlighting new directions in which the prognosis and treatment of BE will benefit from the technologies to help experts in this task. Also, the definition of a pattern in the BE identification is very important, considering the massive human evaluation of this problem in practice today. We believe the computer learning and classification may help to define markers and identifiers for the best BE description, helping the accurate and fast definition of the injured region in endoscopy images or even in endoscopy videos in a real-time definition way.

The primary challenges related to computer-assisted BE identification are mainly associated with the lack of data since most of the datasets figure a few dozens of patients only. Another problem is related to the absence of public datasets to cope with BE identification, which could foster the research towards more effective approaches to detect early-stage illness from endoscopic images.

Another bottleneck concerns unbalanced data, which may bias the machine learning technique towards the majority class. Data augmentation appears to be an exciting solution together with transfer learning approaches. Also, we believe that combining handcrafted with automatic learned features may be a useful idea, mainly in the context of medical-drive data where the lack of images is of great concern.

Table 2.1: Summarization of the works considered in this survey.

Reference	Classifier	Database	Validation Protocol	Evaluation Method	Results
[Li e Meng 2009]	SVM	100 endoscopic images from 5 patients	4-folder cross-validation	Presented a study for the application of a new texture extraction scheme in ulcer regions for capsule endoscopy discrimination in endoscopic images.	Detection rate of the proposed features with the MLP was 92.37%, 91.46%, and 93.28% in terms of accuracy, specificity and sensitivity, respectively.
[Rodriguez-Diaz e Singh 2016]	SVM	16 non-neoplastic and 40 neoplastic NBI images from 26 patients	LOO-CV	Conducted a study aiming to explore the feasibility in developing a diagnostic computer algorithm based on the NICE criterion for a real-time interpretation of polyp histology from near-focus NBI images.	The classification based on color resulted in a sensitivity of 86% and specificity of 86% with a high-confidence rate of 77.%
[Nancarrow et al. 2011]	SVM	54 images from 54 patients (23 presenting EAC)	LOO-CV	Performed a comparative study to define convincing differences between BE and EAC using SVM classification.	The results were: sensitivity and specificity higher than 0.88 for discriminating BE from squamous, and the sensitivity for determining EAC (cancer) from BE or squamous (non-cancer) was 0.73.
[Veronese et al. 2013]	SVM	CLE images from 29 consecutive BE patients undergoing surveillance	LOO-CV	Presented a computer-based method for the automatic classification of gastric metaplasia, intestinal metaplasia, and neoplasia on the basis of appearance features of confocal images.	The sensitivity overall results were: 96% of gastric metaplasia, 95% of intestinal metaplasia, and 100% of neoplasia.

[Muldoon et al. 2010]	SVM	129 sites from a database of 16 HMRE images	cross-validation	Developed a CAD system to help physicians with faster identification of early cancer using color and texture features extracted from HD endoscopy images.	In an SVM-classification approach, the results reached were up to 96.48%.
[van der Sommen et al. 2013]	SVM	HD endoscopic images from 66 patients	10-fold cross-validation	Proposed an algorithm based on HD endoscopic images for the automatic esophagus irregularity identification and location using color histograms, Gabor features, and SVM-based classification.	The proposed system achieved a classification accuracy of 95.9% with AUC value as of 99%.
[van der Sommen et al. 2014]	SVM	32 HD endosocopic images from 7 patients	LOO-CV	Presented a novel algorithm that computes local color and texture features based on the original and on the Gabor-filtered image for the automatic detection of early cancerous tissue in high defintion endoscopic images.	From 38 lesions, the system detected correctly 36 of those lesions with a recall of 0.95 and a precision of 0.75.
[Hassan e Haque 2015]	SVM	120,000 WCE frames for the training set and 1720 WCE frames for the test set	cross-validation	Proposed a real-time bleeding detection technique based on the observation of characteristic patterns that appear in the frequency spectrum of the WCE.	The accuracy, sensitivity, and specificity values achieved were 99.19%, 99.41% and 98.95%, respectively.

[Seguí et al. 2016]	CNN	100,000 WCE images for training the network and 10,000 for testing	cross-validation	Introduced a system for small intestine motility characterization based on Deep Convolutional Neural Networks.	The proposed approach obtained a mean classification accuracy of 96% for six intestinal motility events.
[Mendel et al. 2017]	CNN	Database provided by [MICCAI 2015: 18th International Conference 2015]	LOPO-CV	Carried out a work in which deep learning was applied in specialist annotated images containing adenocarcinoma and BE's disease.	The sensitivity and specificity values of 94% and 88% respectively.
[Souza Jr. et al. 2017]	SVM	Database provided by [MICCAI 2015: 18th International Conference 2015]	LOPO-CV	Tested the feasibility of adenocarcima classification in endoscopic images using SURF features and SVM classifier.	The mean results for the "full image approach" were 77% of sensitivity and 82% of specificity. For the "region-based" approach, the results were 89.6% of sensitivity and 95.1% of specificity.

[Zhang et al. 2016]	SVM	1,210 EUS images from 157 patients (66 with early cancer and 91 without)	10-fold cross-validation	A study was conducted using endoscopic ultrasonography (EUS) to calculate textural features in a spectral analysis of pixels to provide a quantification of early esophageal carcinoma tissue.	In the first classification approach, the overall concordance rate was 55.6% with kappa coeficient of 28%. The early esophageal carcinoma average prediction results of accuracy, sensitivity, specificity, and negative prediction were 89.4%, 94%, 95%, and 97%.
[Klomp et al. 2017]	SVM	30 VLE non-dysplastic images and 30 VLE dysplastic images	10-fold cross-validation	Tested the feasibility in the use of computer vision techniques correctly predict the presence of dysplastic tissue in VLE BE images.	Considering the novel proposed descriptors, the area under ROC curve result was 95%, compared to the 81% of the clinical predicition model.
[Georgakopoulos et al. 2016]	CNN	Database proposed in [Koulaouzidis e Iakovidis 2015]	Patches evaluation (normal or abnormal)	Proposed a weakly-supervised learning method based on CNNs that uses only image-level semantic annotations in the training process for ground truth calculation.	The results achieved by the authors showed the proposed method can be more effective than the conventional supervised learning with an accuracy of around 90%.
[Chan et al. 2016]	CNN	66 VOCs of patients presenting BE and 56 VOCs of patients without BE	LSOCV	Used an e-nose to perform a cross-sectional study evaluating the breath VOCs of a cohort of 112 patients with a history of dysplastic BE to differentiate the differences in BE by dysplasia grade.	The sensitivity result was 82%, the specificity result value was 80%, the accuracy was 81%, and the AUC was 79%.

[Yoshida et al. 2017]	ANN	3062 gastric-biopsy specimen slides	cross-validation	Using textural features, were performed a study aiming to evaluate the classification accuracy of gastric biopsy spicimens using the *e-Pathologist* image software and expert annotations, in two different comparison approaches.	In the first classification approach, the overall concordance rate was 55.6% with kappa coeficient of 28%. In the second approach, the sensitivity, specificity, and negative predictive vale were 89.5%, 50.7%, and 90.6%, respectively.
[Hong, Park e Park 2017]	CNN	155 gastric metaplasia instances, 26 instestinal metaplasia instances and 55 neoplastic instances	cross-validation	Developed a CAD system to classify endomicroscopy images between gastric metaplasia, intestinal metaplasia and neoplasia (these last two are sub-classes of BE) using a public database with 262 samples.	The accuracy result obtained was 80.77%, suggesting that CNN could become a good classifier for the task of BE tissue distinction.
[Rajan et al. 2010]	SVM, *k*-NN, Boosting 1/2	125 WLE images, 122 NBI images, and 150 Chromoendoscopy images	cross-validation	Performed experiments using several classifiers (SVM, *k*-NN, Boosting) in images from different endoscopy modalities (WLE, NBI, and Chromoendoscopy).	The accuracy for detecting BE presented a range of variation from 36.36% up to 89.17% according to the endoscopy modality and classifier.
[Sattlecker, Stone e Bessant 2014]	SVM, LDA, ANN, and RF	-	LOO-CV, Bootstraping and independent testing	Conducted a study aiming to coholborate the many promising benefits over the currently used histopathology methods.	If the combination of spectroscopy and machine learning is mapped into clinical practice, more studies need to be carried out to support the reproducibility.

[Kandemir et al. 2014]	MIL and SVM	214 tissue cores (165 presenting cancer and 69 showing healthy condition) from 97 patients	cross-validation	Performed a study for the diagnosis of BE's cancer from hematoxylin-eosin stained histopathological biopsy images using multiple instance learning and SVM classifiers.	For patch-level diagnosis, the result was around 82% of accuracy and 0.89 of AUC using Bayesian logistic regression.
[Yu et al. 2015]	CNN and SVM	25 real WCE recording samples (approximately 1 million of WCE images)	CNN-features compared to SVM-features	Studied the classification problem of the digestive organs for WCE images.	The results showed performance of around 97.25% concerning the classification accuracy.
[Swager et al. 2017]	SVM, adaBoost, and *k*-NN	52 endoscopic resection specimens from 29 patients	LOO-CV	Investigated the feasibility of a computer algorithm to identify early BE neoplasia in VLE images using VLE features and machine learning methods.	AUC of 0.95, and sensitivity and specificity were 90% and 93%, respectively.
[Souza Jr. et al. 2017]	SVM and OPF	Database provided by [MICCAI 2015: 18th International Conference 2015]	cross-validation	Introduced the OPF classifier in the context of adenocarcinoma and BE classification using SURF and SIFT features combined with a bag-of-visual-words approach for the feature vectors calculation.	The OPF outperformed the SVM, presenting better results for both feature extractors, with values lying on 73.2% (SURF) - 73.5% (SIFT) for sensitivity, 78.2% (SURF) - 80.6% (SIFT) for specificity, and 73.8% (SURF) - 73.2% (SIFT) for the accuracy.

[Pu et al. 2017]	logistic regression, SVM, CNN, LDA, Naive Bayes and flexible discriminant analysis	100 samples of ESSC DNA methylation and a particular dataset of 12 samples of the same kind	5-fold cross-validation	Performed a study to extract more cost-efficient biomarkers (using DNA methylation) than the ones available until now, aiming to provide high sensitivity and specificity in the ESCC diagnosis.	In the SVM classification approach, the best average accuracy result were reached, with value of 0.82%.
[Serpa-Andrade et al. 2015]	k-NN and Random Forests	10 endoscopic images of healthy tissue and 16 images of ill tissue	cross-validation	proposed a method in which the esophagitis (a condition of cronic BE stage) was described using Fourier Transform on the Z-line signature for classification purposes.	The very best average results obtained were 81% of precision, 86% of sensitivity and 72% of specificity.
[Zopf et al. 2009]	euclidean distance	326 ROIs annotated by experts	LOO-CV	Proposed a study using NBI endoscopy images for automatic detection by classification systems with gastroscopy.	Accuracy in the range of 85% and 92% for the feature combination (BE accuracy as of 74%).
[Sharma et al. 2016]	NBI classification criteria	50 NBI images plus 120 additional NBI images	Comparison between BING criteria and expert's annotations	Aimed to develop and validate a narrow-band imaging classification system for the identification of dysplasia and cancer in patients with BE.	The criteria identified patients with dysplasia with 85% of overall accuracy, 80% of sensitivity, 88% of specificity, 81% of positive predictive value, and 88% of negative predictive value.

[Pech et al. 2008]	*k*-statistics	43 lesion images	confocal image review (2 experts)	Performed a study to assess the potential of endomicroscopy for predicting histology in patients with early squamous cell cancer in the esophagus.	The results were: accuracy value as of 95%, and the sensitivity and specificity as of 100% and 87%, respectively.
[Rosenfeld et al. 2014]	Decision Trees	47 HD endoscopic videos (23 dysplastic and 23 non-dysplastic)	–	Studied how data mining can be applied to aid the diagnosis of patients with high-risk lesions within BE.	The overall accuracies concerned the aforementioned models were around 79% (with the experts' decision) and 77% (without the experts' decision).
[Li et al. 2015]	Citation-*k*-NN	26 patients	LOO-CV	Proposed a novel multiple instance learning method for the identification of tumor invasion of gastric cancer with dual-energy computed tomography imaging.	The experimental evaluation was performed using leave-one-out cross validation, obtaining an accuracy of 76.92%.
[Curvers e Bergman 2016]	Automated image classification	–	independent set of images	Carried out a study with patients undergoing BE neoplastic or surveillance underwent standard gastroscopy.	The results of the quantitative image classification algorithm showed a sensitivity of 84% and specificity of 85% in the learning set, and a sensitivity of 88% and specificity of 85% in the validation data.
[Wang et al. 2016]	Automatic image analysis	9 VLE volumetric datasets from 7 patients	–	Developed an algorithm that can automatically detect and quantify subsquamous glandular structures in VLE data and RFA in the decrease of the BE extension.	The subsquamous glandular structures depth and eccentricity characteristics were the most significant for the RFA outcome.

[Swager et al. 2017]	VLE features comparison	30 non-dysplastic VLW images and 30 high-grade dysplasia/ early-adenocarcinoma VLE images	cross-validation	Conducted a study aiming to identify VLE features of BE neoplasia and to develop a VLE prediction score.	The sensitivity and specificity values obtained were 83% and 71%, respectively, showing promising results.
[Boschetto, Gambaretto e Grisan 2016]	Random Forests	116 NBI images	10-fold cross-validation	Presented a CAD system to automate the classification of normal and metaplastic endoscopic NBI images. Eight features were extracted from regions defined as clusters of superpixels.	The overall accuracy, sensitivity - and specificity result values were 83.9%, 79.2% and 87.3%, respectively.
[Ghatwary, Ahmed e Ye 2017]	–	–	–	Presented a brief survey to discuss advances in the development of BE CAD systems WLE, HD-WLE, and NBI endoscopy modalities.	Eight works were listed based on number of images, classifier, validation method, and results.

2.5 Chapter's Considerations

After publishing this manuscript in 2018, several new works and approaches have been proposed to date. Aiming to keep the coverage of this survey, Table 2.1 has been updated since then with related ML works applied to BE and adenocarcinoma context evaluation.

Hence, below, the extra works from 2018 to 2021 can be observed, considering the same searching metrics of Table 2.1. As one can observe, a rising perspective regarding the results was achieved through the new studies, encouraging the introduction of new methods and techniques to evaluate such an important problem.

So far, 15 new works were conducted and listed in the table below, among the evaluation of images and videos of endoscopic examinations of cancerous tissues and BE. It is mandatory to highlight the increase in approaches that apply deep learning techniques, illustrating how important such an application is for solving current medical problems, with promising correctness rates.

From this survey, that guided the entire study we propose, techniques of description based on handcrafted features, extracted using the experts' knowledge, and based on deeply-learnable features, based only on deep learning techniques, were employed in the evaluation of cancer and BE distinction context in the next Chapters to come.

Table 2.2: Summarization of the works from 2018 to 2021.

Reference	Classifier	Database	Validation Protocol	Evaluation Method	Results
[Souza Jr. et al. 2018]	OPF, SVM and naive-bayes	174 endoscopic images from 2 datasets.	20-fold cross-validation and LOPO-CV.	Presented a study for the extraction of color-coocurrence matrices as BE and cancerous tissue features further classified using OPF, SVM and bayesian-classifier.	Accuracy, sensitivity and specificty and f1-score rates of 76.6%, 72.6%, 76.6%, and 74.8% respectively.
[Passos et al. 2019]	iRBM, OPF, SVM and naive-bayes	100 endoscopic images from 39 patients.	20-fold cross-validation.	In this study, the authors employed the inifite Restricted Boltzmann Machine, based on metaheuristics, to describe and classify BE and cancer, also employing the OPF, SVM and bayesian classifiers for comparison purposes.	Accuracy, sensitivity, and specificty rates of 79.0%, 82.0%, and 76.0% respectively.
[Souza Jr. et al. 2019]	OPF, SVM and naive-bayes	174 endoscopic images from 2 datasets.	20-fold cross-validation.	In this study, the authors conducted the evaluation of BE and adenocarcinoma by calculating its features based on an unsupervised version of OPF, standarizing the dimension of the output features to be classified using OPF, SVM and bayesian classifiers.	Accuracy, sensitivity, and specificty rates of 67.3%, 77.3%, and 76.6%, respectively.

[Souza Jr. et al. 2020]	LeNet-5 and AlexNet	174 endoscopic images from 2 datasets.	20-fold cross-validation.	In this study, artificial samples of BE and adenocarcinoma were provided using generative adversarial networks, building new datasets along with original samples to assess the impact the amount of samples could present in classifying cancerous tissues.	Accuracy, sensitivity, and specificty rates of 85.0%, 88.0%, and 82.0%, respectively.
[Souza Jr. et al. 2021]	LeNet-5 and AlexNet	174 endoscopic images from 2 datasets.	20-fold cross-validation.	Artificial samples of BE and adenocarcinoma were generated using optimazation techniques based on metaheuristics and generative adversarial networks, building new datasets along with original samples to assess the impact the artificial samples could present in classifying cancerous tissues.	Accuracy, sensitivity, and specificty rates of 93.4%, 90.3%, and 94.1%, respectively.
[Souza Jr. et al. 2021]	AlexNet, ResNet-50, SqueezeNet, and VGG-16	174 endoscopic images from 2 datasets.	20-fold cross-validation and LOPO-CV.	The learning process related to training and testing deep models for BE and adenocarcinoma was assessed by employing explainable artificial intelligence methods, in a back-propagation-fashion, into the defined models, to compare the regions of interest from deep-architectures and experts.	Accuracy, sensitivity, specificty and correlation rates of 81.6%, 89.9%, 72.5%, and 72.0%, respectively.

[Ebigbo et al. 2019]	ResNet-101	229 endoscopic samples from 2 datasets.	LOPO-CV.	A system that provides the classification of BE and cancerous samples was proposed, using the generalization of a deep version of ResNet model, the ResNet-101 architecture.	Sensitivity, specificity and DICE values of 97.0%, 88.0%, and 72.0%, respectively.
[Ebigbo et al. 2020]	ResNet-backbone and DeepLab-v3+ architecture	Proprietary dataset coposed of 129 endoscopic images.	LOPO-CV.	A real-time deep learning AI system was developed, capturing random images from the real-time camera livestream and providing a global classification and a segmentation of BE and early oesophageal adenocarcinoma.	Accuracy of 89.9%.
[Ghatwary, Zolgharni e Ye 2019]	VGG-16	100 endoscopic images from 39 patients.	5-fold and LOPO-CV.	This study aimed to evaluate the performance of different deep learning methods to automatically identify adenocarcinoma from high-definition white light endoscopy images.	F1-score, sensitivity, and specificty rates of 94.0%, 96.0%, and 92.0%, respectively.
[Ghatwary, Ye e Zolgharni 2019]	DenseNet	1,100 endoscopic images from 2 public datasets.	20-fold cross-validation and LPOP-CV.	The proposed system is based on combining Gabor handcrafted features with the CNN features.	Recall, precision, and mean average precision rates of 95.0%, 91.0%, and 84.0%, respectively.

[van der Putten et al. 2020]	U-Net based architecture	Proprietary dataset composed of 494,355 endoscopic images.	5-fold cross-validation.	The authors have employed a novel U-Net-based semi-supervised learning algorithm to pre-train instances of BE and adenocarcinoma domain,in a hybrid transfer learning strategy.	Patient level accuracy, sensitivity, and specificity rates close to 90%.
[Hou et al. 2021]	ResNet-based models	100 endoscopic images from 39 patients.	10-fold cross-validation and LOPO-CV.	It was proposed an end-to-end neural network architecture with an attentive hierarchical aggregation and a self-distillation mechanisms to address the classification of adenocarcinoma in BE-diagnosed samples.	Area under curve, F1-score, sensitivity, and specificty rates of 96.3%, 92.5%, 91.2%, and 94.0%, respectively.
[Gehrung et al. 2021]	AlexNet, DenseNet, InceptionV3, ResNet-18 and VGG-16	4,662 slides from 2,331 patients.	cross-validation and LOPO-CV.	It was proposed We have presented a triage-driven approach that analyzes samples of the Cytosponge-TFF3 test using deep learning for the detection of BE tissue.	Area under curve, snsitivity, and specificty rates of 88.0%, 73.5%, and 93.1%, respectively.
[Dickson 2019]	Deep learning-based models	1,704 endoscopic images from 669 patients.	cross-validation.	It was developed a piloty video-based CAD system for BE neoplasia that allows endoscopic detection on video-footage.	Accuracy, sensitivity, and specificty rates of 89.0%, 93.0%, and 83.0%, respectively.

[Pan et al. 2021]	Fully convolutional networks	443 endoscopic images from 187 patients.	cross-validation.	The authors proposed to develop a fully automated deep learning system for the accurate segmentation and identification of BE in endoscopic examinations.	Intersection over Union and Dice coeficients of 0.81%, and 0.90%, respectively.

Chapter 3

BARRETT'S ESOPHAGUS ANALYSIS USING SURF FEATURES

The first proposed attempt of adenocarcinoma and Barrett's esophagus distinction is presented in this chapter and presented in the conference "Bildverarbeitung für die Medizin 2017" [Souza Jr. et al. 2017]. Considering the endoscopic image representation based on the Speed-Up Robust Features (SURF), the classification of cancerous and non-cancerous regions was provided using the Support Vector Machines classifier in two different image-approaches.

3.1 Introduction

In the last decades, the incidence of adenocarcinoma in patients with Barrett's esophagus has increased significantly in Western populations. The dismal prognosis of the disease can be largely improved through early identification and surgical treatment of high-grade dysplasia and non-metastatic stages of cancer [Dent 2011, Sharma et al. 2016, Phoa et al. 2016]. Therefore, strong emphasis is being placed on the computer assisted diagnosis of endoscopy images. Some studies have already been carried out to classify lesions of the esophagus, based on conspicuous color and textural anomalies [van der Sommen et al. 2016]. Benefitting from substantial improvements in the field of image analysis and artificial intelligence, methods like SURF [Bay et al. 2008] and Deep Learning [Mendel et al. 2017] are increasingly applied. The aim of the current study was to investigate the feasibility of a SVM to classify dysplastic and cancerous lesions in Barrett's esophagus based on SURF descriptors.

3.2 Materials and Methods

This section demonstrates the steps to develop a computerized system for the detection, delineation and characterization of endoscopic images obtained from individuals with clinically manifest tissue abnormalities in the esophagus. Based on a given set of endoscopic photographs (benchmark database). SURF descriptors are utilized for the training and validation of an SVM classifier.

3.2.1 Image Database

The set of images used as benchmark database was provided at the MICCAI 2015 Endo-Vis Challenge [MICCAI 2015: 18th International Conference 2015]. It is composed of 100 endoscopic pictures of the lower esophagus, captured from 39 individuals, 17 of them being diagnosed with early stage Barrett's, and 22 displaying signs of esophageal adenocarcinoma. From each proband several endoscopic images were available, ranging from one to a maximum of eight. The database contained a total of 50 images displaying cancerous tissue areas (C2 labeled images), plus 50 images showing dysplasia without signs of cancer (C0 labeled images). Suspicious lesions observed in the C2 images had been delineated individually by five endoscopy experts. Some of the expert's demarcations in identical images exhibited substantial regional deviations. Therefore all delineated (masked) areas were combined to employ a gold standard for definitive states of adenocarcinoma.

3.2.2 SURF

The SURF algorithm [Bay et al. 2008] operates on integral images to detect dominant structures and their spatial orientation. To ensure scale and spatial invariance the SURF seeks for maxima of the determinant of Hessian, demarcating specific key-points in the image [Bay et al. 2008], which are further explored in their local neighborhood. These sub-regions are evenly split into square patches while their wavelet responses in horizontal and vertical directions generate the elements of a high-dimensional feature vector of size 64.

3.2.3 Interest Points

Interest point (IP) acquisition was performed with the SURF algorithm provided in MATLAB using the OpenCV interface support package. The assessment of suitable IPs was based on two approaches. The first approach simulated "real life situations"lacking detailed informa-

tion about tissue abnormalities. This analysis worked on the original full images. Two attributes were defined for the SVM training process: Class 0 images (C0, non-cancerous but with possible signs of early dysplasia), and class 2 images (C2, exhibiting cancerous tissue regions). The second approach was based on designated spatial annotations provided by five endoscopy experts, denoted as regions S_1 to S_5. In order to define a secure gold standard despite considerably dissenting delineations (Fig. 3.1), the area of intersection from all demarcated regions in the same image was denoted as class 2 area (C2, "cancerous"). Tissue linings which had been inconsistently marked as cancerous or as non-cancerous were labelled as class 1 (C1, "fuzzy regions"). Epithelium diagnosed by all experts in unison as negative was labelled as class 0 (C0, "non-cancerous"), cf. Equations 1-3.

$$\text{Cancer}: \quad C2 = \bigcap_{i=1,\ldots,5} S_i \tag{3.1}$$

$$\text{Fuzzy}: \quad C1 = \bigcup_{i=1,\ldots,5} S_i \backslash C_2 \tag{3.2}$$

$$\text{Non-cancer}: \quad C0 = 1 - (C_2 \bigcup C_1) \tag{3.3}$$

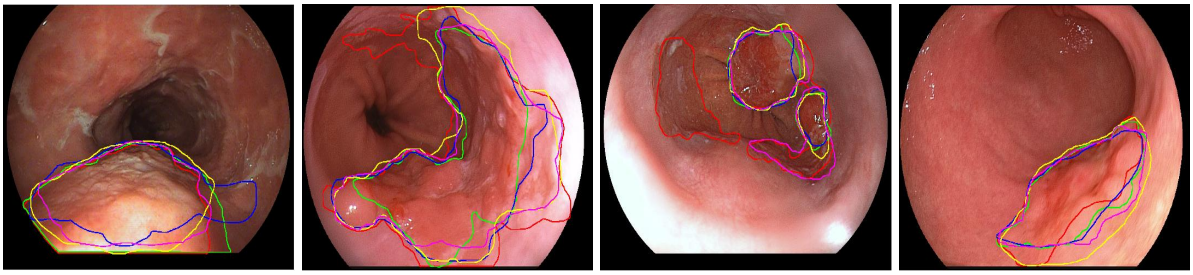


Figure 3.1: Five different experts annotation from four different cancer images.

Full Image Approach: Each image j, $(j = 1, \ldots, 100)$ is mapped to an average vector $\vec{r}(j) \in \mathbb{R}^{64}$ composed of n_{IP} (n_{IP} = number of IP in image j) individual SURF feature vectors $\vec{f}(j,k) \in \mathbb{R}^{64}; k = 1, .., n_{IP}$ (Fig. 3.2, top):

$$\vec{r}(j) = \frac{1}{n_{IP}} . \sum_{k=1}^{n_{IP}} \vec{f}(j,k) \tag{3.4}$$

Masked Image Approach: The process applied to full images is similarly applied to the segmented areas labelled with class codes C0, C1, C2, respectively. Each region is compressed to a mean feature vector, normalized by the number of selected SURF interest points $n_{IP}(i)$ of

the corresponding region. Consequently, each image j ($j = 1,...,50$) belonging to the cancer subset is mapped to three feature vectors $\vec{r}(i,j), (i = 0,1,2)$ (Fig. 3.2, bottom):

$$\vec{r}(j,i) = \frac{1}{n_{IP}(i)} \cdot \sum_{k=1}^{n_{IP}(i)} \vec{f}(j,k,i), i = 0,...,2 \tag{3.5}$$

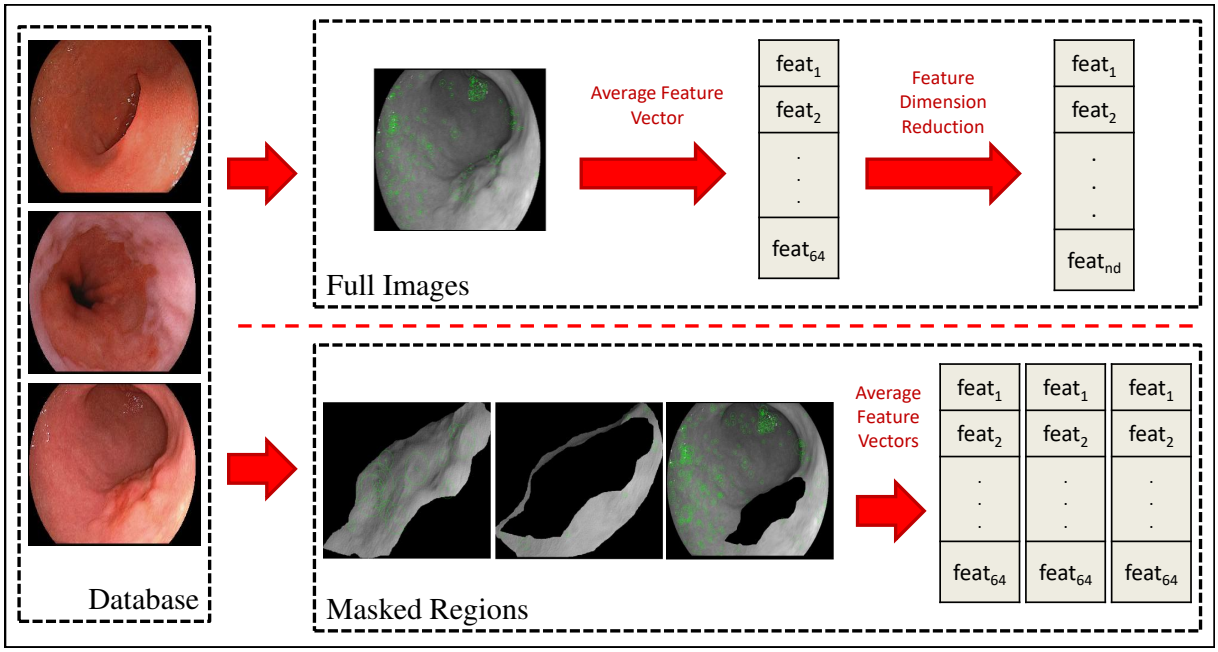


Figure 3.2: Mapping full images and masked areas to feature vectors.

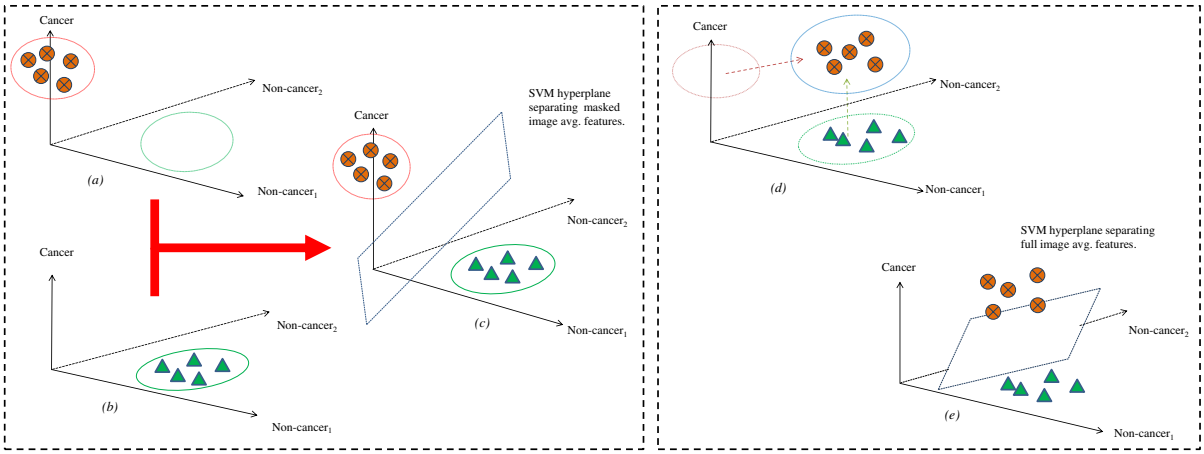
3.2.4 Classification

A SVM classifier was selected to discriminate between C0 and C2 type epithelium. The classification steps were performed in two different ways. In the first approach, SVM training as well as testing was performed using averaged features obtained from the full images. In the second approach, SVM training and test runs worked on average feature vectors of masked image regions. In both cases the conventional LOPO-CV was applied.

3.3 Results

According to the number of patients in the database, 39 computations were performed for each approach. The results are shown in Tab. 3.1. It should be clear that leaving out a certain patient from the training set implies that all available images from this patient are removed from the training set.

Figure 3.3: SVM hyperplanes separating classes C0 and C2 in a fictitious three-dimensional feature space. Points displayed in Figs (a), (b) and (c) indicate average feature vectors extracted from masked C0 and C2 regions, (d) and (e) refer to mean SURF features assessed from full images.



In order to comprehend the results summarized in Tab. 3.1, we consider a feature space spanned by only three variables. Let two features reflect specific properties of the masked non-cancerous regions (C0), and a single feature being indicative for masked cancerous regions (C2). Presuming adequate separability of C0 and C2 points (cf. Fig. 3.3a, b) the SVM hyperplane might be positioned as shown in Fig. 3.3c. A high predictive accuracy of the classifier is expected if the test data are also supplied from masked image regions. SURF feature vectors calculated from full images will of course be located in other regions of the feature space (cf. Fig. 3.3d, e), resulting in an essentially different SVM hyperplane. Moreover, test vectors scatter considerably due to non-distinctive features contained in the full images. Note that hyperplanes generated from masked regions will definitely not cope with test features obtained from full images, and vice versa. The considerations given above are confirmed and documented in Tab. 3.1. Sensitivity and specificity mean value lies above 75% for full image classification, and exceed 90% for masked image classification.

TrainingSets	TestSets	Sensitivity	Specificity
Full Images	Full Images	0.78 ($\sigma = 0.37$)	0.82 ($\sigma = 0.33$)
Masked Regions C0, C1, C2	Masked Regions C0, C1, C2	0.90 ($\sigma = 0.17$)	0.95 ($\sigma = 0.18$)
Masked Regions C0, C1, C2	Masked Regions C2	0.91 ($\sigma = 0.24$)	0
Masked Regions C0, C1, C2	Masked Regions C0	0	0.95 ($\sigma = 0.18$)

Table 3.1: Classification results (mean and standard deviation) referring to full images and masked regions.

3.4 Discussion and Conclusions

The presented CAD system is a promising approach to evaluate various types and stages of epithelial lesions in patients suffering from Barrett's esophagus. The method uses SURF to exploit hidden structural patterns embedded in endoscopy images. This technique was applied to full images as well as to masked image regions which provide a gold standard for the identification of malignant lesions, proposing progress in the application area of the use of machine learning employed in the medical problems solution. Both approaches require a reference database comprising endoscopy images from patients with non-cancerous (C0) and cancerous (C2) linings in the esophagus. In order to develop a classifier with high predictive accuracy, an SVM was trained with a sufficiently large subset of the original endoscopy images. The CAD performance can be improved if the SURF and SVM algorithms operate on delineated (masked) tissue regions. The geometric boundaries separating malignant parts of the mucosa from early dysplastic stages must be provided by clinical experts. Note that the SVM classifier should be trained, validated and tested based on instances from the same set, i.e. either utilizing features from the full image set or from the set of masked regions, respectively. The masked tissue regions were used in order to improve the 'blind' results, once the masked images can be used as a tool in a future totally validated and implemented systems, for the experts help in the lesion area defintion.

The diagnostic accuracy of the designed system expressed by sensitivity and specificity values is high. As expected, the results of the first approach (full images) are somewhat lower than those of the second approach (masked) with 0.78 and 0.82 versus 0.90 and 0.95 for sensitivity and specificity, respectively. It might be argued that using masked regions as test instances is a means to "fake"the SVM performance, because the delineation of malignant tissue regions anticipates the actual SVM classification. However, testing the final SVM with masked area features is twofold legitimate: 1st in order to validate the ultimate classifier on a higher level (as compared with simple LOPO-CV), and 2nd to rate attempts of a yet unexperienced clinician (or even an experienced one) to circumscribe a cancerous tissue area. Motivated by the convincing results of the present study, a further analysis will focus on the fusion of textural features [Palm 2004] and SURF features.

3.5 Chapter's Considerations

The human knowledge concerning cancerous regions within BE, translated into some digital description, presents the trivial observation of how to define, based on interpretation, the

correct injured area. It is imperative to understand if such visual observation presents the same impact in the computational representation of such a context.

Some image processing techniques may be employed to represent human insights, the first way of observing cancer. In this study, the use of SURF to describe experts' knowledge regarding the definition of cancerous tissues has shown promising results and important significance for progressing the study of defining optimal techniques in the automatic detection of cancer BE-diagnosed patients.

First, the regions described by SURF technique provide high-aggregated information from the maxima's seeking performed using Hessian's determinant. This point is corroborated by Table 3.1 and highlights that cancerous samples represented by the experts' delineations could deliver the best results of SURF description and SVM classification.

Then, in another approach conducted by this study, when SURF features were calculated from positive-to-cancer images, a misunderstanding led to the composition of the final descriptor, combining information from cancerous regions (experts' annotations) and non-cancerous regions, the remaining parts of the same image. It is important to point out that SUFT automatically detects its key points, and from this manuscript, both cancerous and only-BE regions present a strong impact to elect key points. Hence, the classification of such descriptors becomes a hard task, considering the inter-combination of features from both classes we aim to differentiate, as Figure 3.3 illustrates.

Finally, not only could we observe a satisfactory performance of SURF for describing the regions, but also interesting achievements in the use of SVM for classifying samples from both BE and adenocarcinoma classes (Figure 3.3). Again, the human knowledge inserted into computation methods can somehow enhance the correct-and-automatic definition of cancer, but its spatial requirements in the automatic key-point definition may present some constraints to be deeply evaluated for further works of the same context. To reinforce the achievements of this work, the next Chapter evaluated more object detector techniques and classifiers, aiming to corroborate the importance of experts' insights in the correct definition of adenocarcinoma in BE patients.

Chapter 4

BARRETT'S ESOPHAGUS IDENTIFICATION USING OPTIMUM-PATH FOREST

This chapter presents the introduction of the OPF classifier to Barrett's esophagus and adenocarcinoma context evaluation published in the 30th Conference on Graphics, Patterns and Images (SIBGRAPI'17) [Souza Jr. et al. 2017]. Considering the very first use of the OPF classifier for such context, the SVM was also used to evaluate visual-word-based features calculated using the SURF and SIFT techniques in an HD-endoscopic database of the low esophagus.

4.1 Introduction

The incidence of adenocarcinoma in patients affected by BE has increased significantly in western populations, explained mainly by the obesity, a well-known risk factor [Lagergren e Lagergren 2010, Dent 2011, Lepage, Rachet e Jooste 2008]. As such, the expectation of this disease to rise up in the next years must be considered. Additionally, the bad prognosis for patients that have esophageal adenocarcinoma is related to its late diagnosis. However, the prognosis of the disease can be largely improved through early identification and surgical treatment, thus achieving very successful rates of handling the disease with 93% of the patients having complete remission after 10 years [Dent 2011, Sharma et al. 2016, Phoa et al. 2016].

Developments in interventional therapies, such as photodynamic therapy, cryotherapy and radio-frequency ablation, showed promising results concerning the treatment of Barrett's esophagus. However, most of such methods are not able to describe properly the disease's level [Shaheen et al. 2009, Johnston et al. 2005, Overholt, Panjehpour e Halberg 2003]. Once the identification of the BE is needed for the further evaluation of its region of interest by any computer or physician, some studies have focused on the definition of the area affected by the disease using the

tissue properties in order to define a pattern to be followed.

Sonmmen *et al.* [van der Sommen et al. 2014] presented a novel algorithm that computes local color and texture features based on the original and on the Gabor-filtered image for the automatic detection of early cancerous tissue in high defintion endoscopic images. Appropriate filters based on spectral characteristics of the cancerous regions were designed, and the extracted features are classified by a trained SVM classifier after which additional post-processing techniques are applied in order to annotate the image region containing early cancer. For seven evaluated patients, the experiments compared thirty two annotations made by the algorithm with the corresponding delineations made by an gastroenterologist expert. From 38 lesions indicated independently by the gastroenterologist, the system detected correctly thirty six of those lesions with a recall of 0.95 and a precision of 0.75.

Souza Jr. *et al.* [Souza Jr. et al. 2017] conducted a study to test the feasibility of adenocarcima classification in endoscopic images. A database composed of 100 expert annotated endoscopic images were used for the SURF features extraction, and the SVM classification using the leave-one patient out protocol for training and test. Two classes composed the problem: C0, from the non-cancerous annotated images, and C2, from the cancerous annotated images. Two approaches for feature extraction and classification were carried out: using the full images and using the expert annotated regions of the adenocarcinoma. The achieved mean results for the full images approach were 0.77 sensitivity and 0.82 specificity. For the regions approach, the results were 0.896 sensivity and 0.951 specificity.

Considering the growth of studies in whitch BE and adenocarcinoma evaluation is performed by means of machine learning techniques, the main contribution of this paper is the application of the OPF classifier for this problem, providing a novel BE and adenocarcinoma assessment and classification, and making possible the comparison between this approach and others already proposed. Once the BE and adenocarcinoma evaluation using CAD systems and computational techniques shows a brand new research area with not too much studies carried out, this work presents an important contribution for the endoscopic image processing and computational evaluation of this context.

Papa *et al.* [Papa et al. 2008] carried out a study for the laryngeal cancer detection using the OPF and SVM classifiers applied to three different public database. the classifiers were evaluated in terms of accuracy and execution time. Using sixteen-sized feature descriptors, a novel supervised classification approach was applied for the problem, once this classifier does not takes account a possible separability of the classes to perform its training and classification phases. The OPF creates an optimum path forest rooted by the prototypes, where the decision

boundaries are obtained by the influence zones of the most representative samples of the training set (prototypes). The samples that fall in the influence region of a prototype will be classified with its label. OPF is three or four times faster than SVM, and in this case, OPF was much faster than SVM, considering that parameters optimization was applied for this last classifier. The OPF classifier outperformed SVM in two databases with respect to the accuracy rates and were faster in all experiments realized. The execution time of a classifier must be considered for this kind of application, due to the large amount of exams that need to be performed in the hospitals and clinics.

4.2 Barrett's Esophagus

The condition in which columnar cells replace the usual squamous cell in the mucosa of the esophagus is known as Barrett's esophagus. This condition is recognized as a complication of gastroesophageal reflux disease, and in some critical stage, it can progress and evolve into esophageal cancer [Hopkins 2008, Sharma et al. 2016, Dent 2011].

Squamous cells similar to those of the skin or mouth compose the mucosa of the normal esophagus. The normal squamous mucosal surface appears whitish-pink in color, while the gastric mucosa appearance goes sharply from salmon pink to red, composed of columnar cells [Sharma et al. 2016, Dent 2011]. A demarcation line, the squamocolumnar (SC) junction or "Z-line", represents the normal esophagogastric junction where the squamous mucosa of the esophagus and columnar mucosa of the stomach meet [Hopkins 2008]. Barrett's mucosa may extend upward in a continuous pattern in which the entire circumference of the distal esophagus is covered by columnar mucosa. A distinction is drawn among patients with more than 3 cm of Barrett's esophagus ("long-segment Barrett's esophagus") and those with less than 3 cm of Barrett's esophagus ("short-segment Barrett's esophagus" [Hopkins 2008]. The esophageal mucosa of patients with suspected BE is carefully examined for the presence of any visible lesions, which are then characterized by the Paris classification [Fujishiro et al. 2006]. Three Targeted biopsies are obtained from the areas of visible lesions.

Multiple endoscopic image enhancement technologies, such as chromoendoscopy, electronic image enhancement (narrow band imaging, flexible spectral imaging color enhancement, i-Scan), confocal laser endomicroscopy, and optical coherence tomography, have been developed for BE use, which may enable endoscopists to conduct a more accurate endoscopic assessment of the diplasia with *in vivo* characterization of esophageal histology. This ability could result in improvements in detection of BE (screening), detection of dysplasia based on BE surveillance,

characterizing abnormalities within BE (selecting lesions and delineating margins during endoscopic therapy), and detection of recurrent neoplasia in patients who have received endoscopic therapy (post-treatment surveillance) [Sharma et al. 2015].

BE is often misdiagnosed during endoscopy and this can be attributed to 1 of 2 reasons: (1) inability to differentiate columnar mucosa of the proximal stomach (cardia) from metaplastic epithelium in the distal esophagus or (2) lack of goblet cells in biopsies obtained from columnar lined epithelium in the esophagus. Areas of dysplasia or early cancer in BE are sometimes not visible with standard white-light endoscopy. Hence, the Seattle biopsy protocol is recommended in which 4-quadrant biopsies are taken every 1 cm of the Barrett's mucosa. However, this biopsy protocol is prone to sampling error because only a small fraction of the entire BE mucosa is sampled (especially in patients with extensive disease) [Sharma et al. 2015]. In addition, the biopsy protocol can be costly and time-consuming (because of the number of biopsies submitted to pathology) making endoscopists usually do not follow it for extensive biopsies, and when it is not followed, there is a significant rise in the risk of missed dysplasia or cancer [Abrams et al. 2009].

4.3 Optimum-Path Forest

In this section, we explain the OPF working mechanism. Although we have different versions in the literature, we considered the first proposed one [Papa, Falcão e Suzuki 2009, Papa et al. 2012]. Roughly speaking, the OPF classifier models the problem of pattern recognition as a graph partition in a given feature space. The nodes are represented by the feature vectors and the edges connect all pairs of them, defining a full connectedness graph. The partition of the graph is performed through a competition process among some key samples (prototypes), which offer optimum paths to the remaining nodes of the graph. Each prototype sample defines its own optimum-path tree (OPT), and the collection of all OPTs defines an optimum-path forest, which gives the name to the classifier.

Let $\mathscr{Z} = \mathscr{Z}_1 \cup \mathscr{Z}_2$ be a dataset labeled with a function λ, in which $\mathscr{Z}_1$ and $\mathscr{Z}_2$ stand for the training and test sets, respectively, such that $\mathscr{Z}_1$ is used to train a given classifier and $\mathscr{Z}_2$ is used to assess its accuracy. Let $\mathscr{S} \subseteq \mathscr{Z}_1$ a set of prototype samples. Essentially, the OPF classifier creates a discrete optimal partition of the feature space such that any sample $\mathbf{s} \in \mathscr{Z}_2$ can be classified according to this partition.

The OPF algorithm may be used with any *smooth* path-cost function which can group samples with similar properties [Falcão, Stolfi e Lotufo 2004]. Papa et al. [Papa, Falcão e Suzuki

2009, Papa et al. 2012] employed the path-cost function f_{max}, which is computed as follows:

$$\begin{aligned} f_{max}(\langle \mathbf{s} \rangle) &= \begin{cases} 0 & \text{if } \mathbf{s} \in \mathscr{S}, \\ +\infty & \text{otherwise} \end{cases} \\ f_{max}(\pi \cdot \langle \mathbf{s}, \mathbf{t} \rangle) &= \max\{f_{max}(\pi), d(\mathbf{s}, \mathbf{t})\}, \end{aligned} \tag{4.1}$$

in which $d(\mathbf{s}, \mathbf{t})$ denotes the distance between samples $\mathbf{s}$ and $\mathbf{t}$, and a path π is defined as a sequence of adjacent samples. As such, we have that $f_{max}(\pi)$ computes the maximum distance among adjacent samples in π, when π is not a trivial path.

The OPF algorithm assigns one optimum path $P^*(\mathbf{s})$ from $\mathscr{S}$ to every sample $\mathbf{s} \in \mathscr{Z}_1$, forming an optimum path forest P (a function with no cycles that assigns to each $s \in \mathscr{Z}_1 \backslash \mathscr{S}$ its predecessor $P(\mathbf{s})$ in $P^*(\mathbf{s})$ or a marker *nil* when $\mathbf{s} \in \mathscr{S}$. Let $R(\mathbf{s}) \in \mathscr{S}$ be the root of $P^*(\mathbf{s})$ that can be reached from $P(\mathbf{s})$. The OPF algorithm computes for each $\mathbf{s} \in \mathscr{Z}_1$, the cost $C(\mathbf{s})$ of $P^*(\mathbf{s})$, the label $L(\mathbf{s}) = \lambda(R(\mathbf{s}))$, and the predecessor $P(\mathbf{s})$.

4.3.1 Training

In the training phase, the OPF algorithm aims to find the set $\mathscr{S}^*$, that is the optimum set of prototypes, by minimizing the classification errors for every $\mathbf{s} \in \mathscr{Z}_1$ through the exploitation of the theoretical relation between minimum-spanning tree (MST) and optimum-path tree (OPT) for f_{max} [Allène et al. 2010]. The training essentially consists in finding $\mathscr{S}^*$ from $\mathscr{Z}_1$ and an OPF classifier rooted at $\mathscr{S}^*$.

By computing an MST, we obtain a connected acyclic graph whose nodes are all samples of $\mathscr{Z}_1$ and the arcs are undirected and weighted by the distances d between adjacent samples. The spanning tree is optimum in the sense that the sum of its arc weights is minimum as compared to any other spanning tree in the complete graph. In the MST, every pair of samples is connected by a single path which is optimum according to f_{max}. That is, the minimum-spanning tree contains one optimum-path tree for any selected root node.

The optimum prototypes are the closest elements of the MST with different labels in $\mathscr{Z}_1$ (i.e., elements that fall in the frontier of the classes). After finding prototypes, we run the competition process in order to build the optimum-path forest.

4.3.2 Testing

For any sample $\mathbf{t} \in \mathscr{Z}_2$, we consider all arcs connecting $\mathbf{t}$ with samples $\mathbf{s} \in \mathscr{Z}_1$, as though $\mathbf{t}$ were part of the training graph. Considering all possible paths from $\mathscr{S}^*$ to $\mathbf{t}$, we find the

optimum path $P^*(\mathbf{t})$ from $\mathscr{S}^*$ and label $\mathbf{t}$ with the class $\lambda(R(\mathbf{t}))$ of its most strongly connected prototype $R(\mathbf{t}) \in \mathscr{S}^*$. This path can be identified incrementally by evaluating the optimum cost $C(\mathbf{t})$ as

$$C(\mathbf{t}) = \min\{\max\{C(\mathbf{s}), d(\mathbf{s},\mathbf{t})\}\},\ \forall \mathbf{s} \in \mathscr{Z}_1. \tag{4.2}$$

Let the node $\mathbf{s}^* \in \mathscr{Z}_1$ be the one that satisfies Equation 4.2 (i.e., the predecessor $P(\mathbf{t})$ in the optimum path $P^*(\mathbf{t})$). Given that $L(\mathbf{s}^*) = \lambda(R(\mathbf{t}))$, the classification simply assigns $L(\mathbf{s}^*)$ as the class of $\mathbf{t}$. An error occurs when $L(\mathbf{s}^*) \neq \lambda(\mathbf{t})$.

4.4 Materials and Methods

This section demonstrates the steps to develop a computerized system for the detection, delineation and characterization of endoscopic images obtained from individuals with clinically manifest tissue abnormalities in the esophagus. Based on a given set of endoscopic photographs (benchmark database), SURF and SIFT descriptors are utilized for a Bag of Visual Words (BoVW) construction of descriptors used in the training and validation of SVM and OPF classifiers.

4.4.1 Image Database

The set of images used as benchmark database was provided at the MICCAI 2015 Endo-Vis Challenge [MICCAI 2015: 18th International Conference 2015]. It is composed of 100 endoscopic pictures of the lower esophagus, captured from 39 individuals, 17 of them being diagnosed with early stage Barrett's, and 22 displaying signs of esophageal adenocarcinoma. From each proband several endoscopic images were available, ranging from one to a maximum of eight. The database contained a total of 50 images displaying cancerous tissue areas (C2 labeled images), plus 50 images showing dysplasia without signs of cancer (C0 labeled images). Suspicious lesions observed in the C2 images had been delineated individually by five endoscopy experts. Some of the expert's demarcations in identical images exhibited substantial regional deviations.

4.4.2 SURF

The SURF algorithm [Bay et al. 2008] operates on integral images to detect dominant structures and their spatial orientation. To ensure scale and spatial invariance the SURF seeks for maxima of the determinant of Hessian, demarcating specific key-points in the image [Bay et

al. 2008], which are further explored in their local neighborhood. These sub-regions are evenly split into square patches while their wavelet responses in horizontal and vertical directions generate the elements of a high-dimensional feature vector of size 64.

4.4.3 SIFT

The SIFT algorithm [Lowe 2004] operates on image local regions aiming to calculate features that are invariant to image scaling and rotation, and partially invariant to change in illumination and 3D camera viewpoint. First, the algorithm seeks for the scale-space extrema detection, evaluating all the image scales and regions using difference-of-Gaussian function to provide the potential image regions that are invariant to scale and orientation. The keypoint localization is performed based on the candidate regions previously defined and measures of its stability. The final steps of SIFT algorithm are related to the orientation definition of the keypoints by the calculation of the gradient directions, and the keypoint descriptor calculation based on gradients measurement at the selected scale in the region around the keypoint [Lowe 2004]. These local descriptors are transformed into a global high-dimensional feature vector of size 128 that allows invariability for significant levels of local shape distortion and change in illumination.

4.4.4 Interest Points

Interest point (IP) acquisition was performed with the SURF and SIFT algorithms using the OpenCV support package. The assessment of suitable IPs was based on two major approaches, one using SURF features and other using SIFT features. Both approaches simulated "real life situations"lacking detailed information about tissue abnormalities. The analysis worked on the original full images. Two attributes were defined for the SVM and OPF training process: Class 0 images (C0, non-cancerous but with possible signs of early dysplasia), and class 2 images (C2, exhibiting cancerous tissue regions).

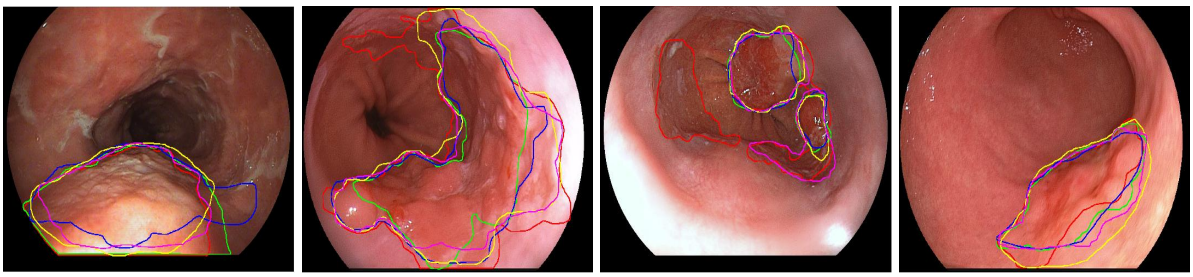


Figure 4.1: Five different experts annotation from four different cancer images.

4.4.5 Bag of Visual Words

BoVW constitutes a robust representation approach in which each image is treated as a collection of regions. For this representation, the only information cared about is the appearance of each region [4]. The objective when visual dictionary is created is to learn, from a training set of examples, the generative model that selects the k more representative regions for a given problem. These regions create a k-dimensional Hilbert space $\mathscr{H}$, in which each region is now represented by a visual word [Afonso et al. 2012]. For that, the original input image regions are mapped from the original space ϕ to a Hilbert space $\mathscr{H}$ represented by the calculated visual words. Therefore, BoVW uses the IPs from a set of reference images in order to generate a visual dictionary that is employed in the training and testing phases.

The reference descriptors for this work were extracted by using the SURF and SIFT methods, explained in the previous subsections. It is important to register that the number of IPs varies for each image and just a few of the total are selected to generate the visual dictionary. The IPs selections, for this work, is performed using Random selection and K-means. Once the visual dictionary is generated, a feature vector is created for each image by computing the frequency of each visual word in the image, and the feature vectors of all the images have the same dimension [Afonso et al. 2012].

In other words, at the end of the feature vector construction phase, will be obtained six different descriptors for both methods, SURF and SIFT: one of 100 words calculated with and one of 100 calculated using Random selection, one of 500 words calculated with K-means and one of 500 words calculated with Random selection, and finally, one of 1000 words calculated with K-menns and one of 1000 words calculated using Random selection. Thereby, twelve cases will compose the experimentation of this work. The IPs and BoVW descriptors acquisition are illustrated in Figure 4.2.

4.4.6 Classification

Two different classifiers were selected to perform experiments aiming to discriminate between C0 and C2 type epithelium: SVM and OPF. So for each classification case, the steps were composed by the use of the BoVW descriptors, selecting a percentage of instances for training and other for testing, in a cross validation protocol. All the 100, 500 and 1000 word descriptors, from SURF and SIFT IPs, were used in separated experiments.

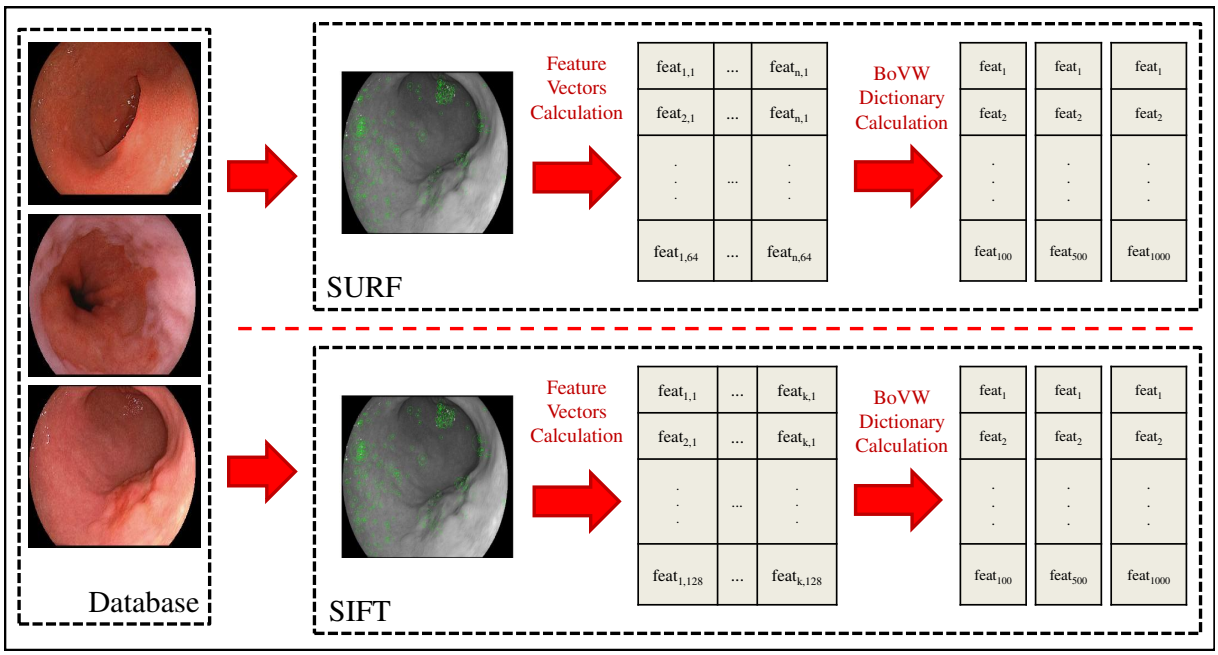


Figure 4.2: BoVW Descriptors calculations based on SURF and SIFT IPs features.

4.5 Results

The results of both approaches are shown in Tables 4.1 and 4.2. The data were generated from 20 computations performed for each experiment approach, using randomly selected training and test candidates. For each experiment, considering a BoVW descriptor and a classifier, were considered for the training and test sets: 80% and 20% of instances respectively. Black values mean the Wilcoxon positive correlation between the best accuracy value and the other accuracy values for each dictionary size (Wilcoxon comparisons performed for accuracy values from the same feature extraction approach).

For the OPF classification approach using SURF features, sensitivity and specificity values lied about 0.71 and 0.75 in the K-means BoVW descriptor and 0.67 and 0.71 in the Random selection BoVW descriptor. For SIFT features, sensitivity and specificity values lied above 0.72 and 0.78 using K-means words, and 0.66 and 0.0.69 in the random selection descriptor approach.

For the SVM RBF classification approach using SURF features, sensitivity and specificity values lied about 0.64 and 0.68 in the K-means BoVW descriptor and 0.63 and 0.66 in the Random selection BoVW descriptor. For SIFT features, sensitivity and specificity values lied above 0.65 and 0.68 using K-means words, and 0.65 and 0.68 in the random selection descriptor approach.

For the SVM Linear classification approach using SURF features, sensitivity and specificity

Table 4.1: Sensitivity (SE), Specificity (SP) and Accuracy (AC) results using SURF Features and 100, 500 and 1000 words

Dictionary		100			500			1000		
		SE	SP	AC	SE	SP	AC	SE	SP	AC
K-means	OPF	0.698	0.711	**0.700**	0.732	0.7823	**0.738**	0.714	0.777	**0.7361**
	RBF	0.644	0.704	0.636	0.657	0.692	0.648	0.639	0.665	0.626
	Linear	0.614	0.672	0.620	0.582	0.593	0.586	0.608	0.627	0.628
Random	OPF	0.688	0.739	**0.697**	0.692	0.718	**0.702**	0.664	0.695	**0.661**
	RBF	0.615	0.656	0.610	0.651	0.678	0.634	0.629	0.657	0.621
	Linear	0.548	0.590	0.517	0.591	0.631	0.576	0.586	0.627	0.565

values lied about 0.61 and 0.64 in the K-means BoVW descriptor and 0.0.58 and 0.0.62 in the Random selection BoVW descriptor. For SIFT features, sensitivity and specificity values lied above 0.0.57 and 0.56 using K-means words, and 0.54 and 0.56 in the random selection descriptor approach.

Table 4.2: Sensitivity (SE), Specificity (SP) and Accuracy (AC) results using SIFT Features and 100, 500 and 1000 words

Dictionary		100			500			1000		
		SE	SP	AC	SE	SP	AC	SE	SP	AC
K-means	OPF	0.705	0.773	**0.683**	0.735	0.806	**0.723**	0.727	0.761	**0.714**
	RBF	0.645	0.658	0.641	0.672	0.706	0.655	0.642	0.686	0.641
	Linear	0.552	0.572	0.552	0.577	0.631	0.568	0.583	0.617	0.673
Random	OPF	0.644	0.663	**0.664**	0.673	0.713	**0.707**	0.642	0.662	**0.712**
	RBF	0.651	0.672	0.621	0.666	0.694	0.656	0.649	0.678	0.637
	Linear	0.545	0.566	0.532	0.548	0.572	0.545	0.544	0.553	0.527

4.6 Conclusion

The presented CAD system is a promising approach to evaluate various types and stages of dysplasia in patients suffering from Barrett's esophagus. The method uses SURF and SIFT to exploit hidden structural patterns in endoscopy images. This technique was applied to full images annotated by five experts, providing a gold standard for the identification of malignant lesions, proposing progress in the application area of the use of machine learning employed in the medical problems solution. The approach require a reference database comprising endoscopy images from patients with non-cancerous (C0) and cancerous (C2) linings in the esophagus. In order to develop classifiers with high predictive accuracy, SVM and OPF were trained

with a sufficiently large subset of the original endoscopy images, and the results of both were compared. The CAD performance showed better results using the OPF classifier for both SURF and SIFT features. The geometric boundaries separating malignant parts of the mucosa from early dysplastic stages must be provided by clinical experts.

Analysing the achieved sensitivity, specificity and accuracy results, can be concluded that the OPF classifier operating on the BoVW approach can provide more efficient results, improving the results achieved using the SVM classifier for this problem. The experiments can be extended using the annotated adenocarcinoma regions made by experts on the images for the BoVW dictionary calculation. This approach can present some other vision of the problem, in which the features generalizing power can be tested.

4.7 Chapter's Considerations

The introduction of new object detector techniques and classifiers was the main contribution this paper aimed to deliver. In addition to SURF features evaluated in Chapter 3, such a study introduced a new technique based on the evaluation of maxima through the space-scale called SIFT, combined not only with SVM but also with OPF classifier.

Considering from Chapter 3 that human knowledge might be strongly correlated to the correct classification of cancerous tissues, in this study, a BoVW technique was employed to describe the features as region representations of key points using unsupervised methods of classification. Then, from all the possible features during the training step, the most representative ones are selected, here called prototypes, to describe the feature positioning behavior. Further, a feature vector is calculated for each sample based on its relative position to the most discriminative features, the key points. This way, the number of representative features for each image could be standardized, also ensuring that spatial localization will impact the feature selection.

From the results, we could observe improvements in the evaluation of full images, as previously cited, aggregates the problem of cancerous and non-cancerous regions at the same image. SURF method was also evaluated using OPF classifier, and so far, such description technique and such classifier, combined, showed up to be the best design in the correct classification of cancerous tissue in BE full-images. This highlights the powerful maxima seeking SURF performs, but the scale operation performed by SIFT deserves attention, providing a high-dimensional feature vector that could perform really close to the ones obtained using SURF descriptors. Probably, due to this high-dimensional nature (double of SURF descriptor size), more information could be represented, but not being specifically discriminative for such a

context evaluation when combined with K-means or Random BoVWs. Finally, OPF classifier shows important generalization for BE context, probably due to its graph representation of features and classes that allied to unsupervised feature representation of BoVW, could strongly organize the classes compared to SVM and Bayes classifiers assessed during the experimental step.

Moreover, there are many object detection techniques to be evaluated, as SIFT was in this manuscript. However, classifiers or descriptors must be evaluated and their clear relation to the experts' delineation of cancer in BE-diagnosed samples. Hence, to progress with such a task, in the next Chapters, we started to evaluate not only new description techniques but also the relation of key-point position and experts' insights of cancer and non-cancer, to establish a clear meaning for the automatic information definition in the classification process our classifiers perform.

Chapter 5

BARRETT'S ESOPHAGUS ANALYSIS USING COLOR CO-OCCURRENCE MATRICES

The idea developed in Chapter 5 is related to the application of color co-occurrence matrices to the evaluation and more specific distinction between adenocarcinoma and Barrett's esophagus tissues. It was used for evaluation two different endoscopic databases of the lower esophagus in three approaches of image preprocessing. This work was published in the 31st Conference on Graphics, Patterns, and Images [Souza Jr. et al. 2018].

5.1 Introduction

The incidence of BE and Barrett's adenocarcinoma in the west of the globe have risen significantly in the past decade. Because of their close association with the metabolic syndrome, this trend is expected to continue rising in the next years [Lagergren e Lagergren 2010, Dent 2011, Lepage, Rachet e Jooste 2008]. The early diagnosis of EAC is critical for the diseases' remission and justifies the necessity of efficient surveillance, detection and characterization. However, the detection of dysplastic regions and their characterization of abnormalities within BE-diagnosed patients can be challenging, especially for endoscopists presenting lack of experience for the evaluation. Even considering the dangerousness of the disease, when detected at the early stages, the injured tissue can be treated with very high rates of remission (93% after 10 years of treatment) [Dent 2011, Sharma et al. 2016, Phoa et al. 2016].

The computer-aided analysis of BE may be one powerful instrument and has been subject of intensive research in the past years [Souza Jr. et al. 2018]. Up to now, mainly handcrafted features of endoscopic images based on texture and color were extracted and classified subsequently. For instance, Van der Sommen [van der Sommen et al. 2016] designed a system for the

automatic extraction of features for detecting and delineating early neoplastic tissue regions in patients diagnosed with BE, followed by some other relevant studies in the same field [Hassan e Haque 2015, Souza Jr. et al. 2017, Mendel et al. 2017] that aimed to assess the feasibility of adenocarcinoma classification in endoscopic images of BE diagnosed patients. Souza et al. [Souza Jr. et al. 2017] also conducted a study introducing two approaches to distinguish between BE and adenocarcinoma: (i) the Optimum-Path Forest (OPF) [Papa, Falcão e Suzuki 2009, Papa et al. 2012] classifier; and (ii) the use of Bag-of-Visual-Words [Csurka et al. 2004, Peng et al. 2016] using points-of-interest extracted from endoscopic images using Speed-Up Robust Features [Bay et al. 2008] and Scale Invariant Feature Transform [Lowe 2004] techniques for the feature vector calculation [Souza Jr. et al. 2017].

There are, in addition, some image processing techniques that can describe the image in different ways, providing feature vectors based on color or texture of the injured region. One of these techniques is the Co-occurrence Matrix (CM), which usually employs gray-scale images to encode texture information. However, there are some new approaches considering the influence of both color and texture for the CM calculation that can provide different descriptions for the BE and adenocarcinoma context [Palm 2004].

Considering the growth of studies in which BE and adenocarcinoma evaluation is performed by means of machine learning and image processing techniques, the main contribution of this paper is the evaluation of Color Co-occurrence Matrices for the description of the dysplastic tissue in BE diagnosed patients. Such assessment provides a novel BE and adenocarcinoma identification approach in which color and texture information can be combined to improve the classification results. Some previous works already evaluated the impact of color and texture information independently for the BE and adenocarcinoma description [van der Sommen et al. 2016, Almond e Barr 2012, Souza Jr. et al. 2017], showing promising results. However, the use of both phenomena in a single descriptor has been poorly studied in this context.

The remainder of this paper is organized as follows: Section II presents a brief background about color and texture combination using Co-occurrence Matrices. Section III discusses the methodology employed in this work, and Section IV presents the experimental results. Finally, Section V states discussions and future works.

5.2 Theoretical Background

5.2.1 Color and texture combination

The “parallel concept" [Palm 2004] for color and texture analysis considers both phenomena for data description separated. While color is measured globally by means of histogram calculation, the texture is characterized by the relationship of the intensities of neighboring pixels ignoring their color. The processing of both information, i.e., color and texture, is performed independently, being combined subsequently to compose a final feature vector (Figure 5.1, (a)). The parallel approach can present advantages; however, the view on texture as a structure purely based on intensity is simplified.

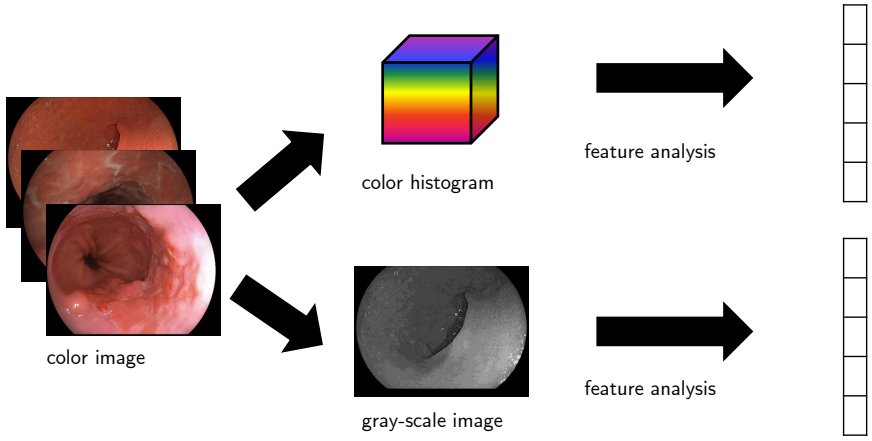
The “sequential concept" [Palm 2004] uses color analysis as a first means, in which the pattern is composed of segmented color primitives obtained by clustering the color histogram. Some previous works showed how useful the sequential approach could be for some tasks, such as industrial quality control and defect detection in granite images [Hauta-Kasari et al. 1999, Song, Kittler e Petrou 1996]. However, the concept of colored texture primitives may not provide generalization support (Figure 5.1, (b)).

In the “integrative concept" [Palm 2004], the information dependency between both color and texture is taken into account for feature extraction purposes. There are two strategies for the integrative color-texture combination: single- and multi-channel (Figure 5.1, (c)). The single-channel method analyses the gray-scale texture on each color channel separately, providing a subtle use for color information restricting the intensity pattern to the wavelength interval associated with that color channel [Palm 2004]. The single channel approach is suitable for methods based on the gray-scale domain. These concepts have been proposed for well-known textural feature description, such as Wavelet-based [van de Wouwer et al. 1999], Gabor filters [Jain e Healey 1998, Palm e Lehmann 2002, Paschos 2001] and Markov Random Fields [Suen e Healey 1999], showing very promising results through the years.

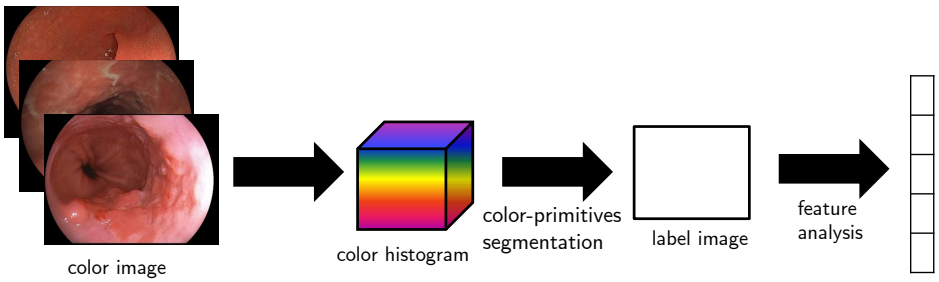
5.2.2 Gray-scale and Single-Channel Co-occurrence Matrices

Co-occurrence Matrices (CM) are defined as the relationship between the values of a central pixel p and its neighboring $\eta(p)$ [Haralick, Shanmugam e Dinstein 1973]. Given a gray-scale image I, a pixel p contains two information: its value $I(p) \in [0,255]$ and its position $p = (m,n)$, such that $m,n \in \mathbb{N}$.

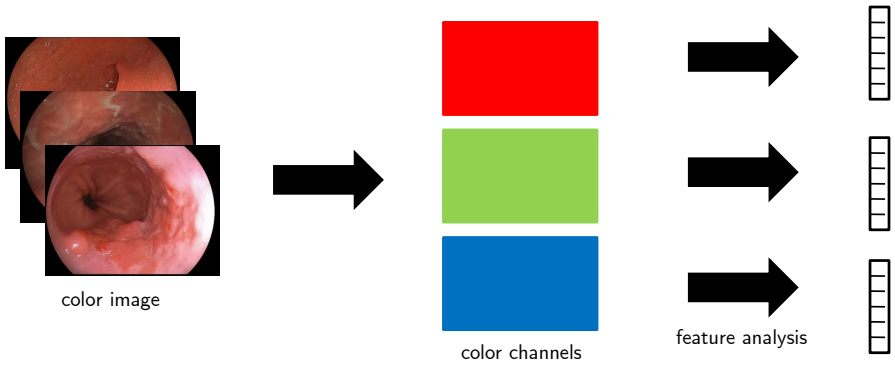
Let η_p be the neighborhood of p such that $p^* \in \eta_p$ when $d(p,p^*) \leq \mathbf{D}$, in which $d(p,p^*)$



(a)



(b)



(c)

Figure 5.1: Color and texture concepts: (a) parallel concept for color texture analysis; (b) sequential concept for color texture analysis and (c) integrative single-channel color texture analysis (adapted from [Palm 2004]).

stands for the polar distance between p and $p^{\star}$. Let C^d be a co-occurrence matrix defined over distance d such that each element is computed as follows:

$$C^d_{i,j} = P(I(p) = i \wedge I(p^*) = j), \tag{5.1}$$

such that $p^* \in \eta_p$. In other words, $C^d_{i,j}$ encodes the probability P of transition between brightness values from i to j. Additionally, it is well known that one must compute one co-occurrence matrix for each orientation angle.

Since C^d is symmetric for each orientation angle according to d, Palm [Palm 2004] proposed to combine the different co-occurrence matrices into a single one that encodes distinct orientation angles. Also, eight Haralick features [Haralick, Shanmugam e Dinstein 1973] were extracted (homogeneity, contrast, correlation, variance, inverse difference moment, entropy, correlation I, correlation II) and distributed over the four feature groups proposed by Gotlieb and Kreyszig [Gotlieb e Kreyszig 1990].

Such approach allows the use of large values for **D**, which is basically the radius of a discrete circle. Computing CMs for four different angles and constant radius **D**, one can obtain 8×4 rotationally-dependent features. In order to be rotationally independent, we compute the mean and variance of each Haralick feature, thus ending up with an 8×2 dimensional gray-scale CM feature (GCF) space.

The Single-Channel Co-occurrence Matrices (SCMs) [Palm 2004] stand for the successively use of gray-scale CMs in each k color channel separately (for RGB system, k = 1,2,3). Such matrices are computed using the very same Equation 5.1, but applied to each color channel. Thus, the corresponding rotational invariant single-channel Co-occurrence features (SCFs) consist of K feature vectors SCF^k presenting analogous behavior of GCF according to k. Therefore, the evaluation becomes a k-dimensional problem, once each k color-channel will provide a different descriptor to be analyzed. The advantage comes with the possibility of evaluation of each color channel independently, analyzing its impact on the texture information composition in a combined color and intensity texture information. The information profit by analyzing intensity independent color textures is quite high, being a brand new way of evaluation for color- and texture-based problems.

5.3 Methodology

In this section, we describe the datasets, pre-processing and feature extraction procedures, classification techniques and approaches employed in this work.

5.3.1 Datasets

5.3.1.1 MICCAI Dataset

The experiments were conducted using the dataset of BE-and adenocarcinoma-diagnosed patients provided at the “MICCAI 2015 EndoVis Challenge"[1]. Such dataset is composed of 100 endoscopic images of the lower esophagus from 39 individuals, in which 22 present BE and 17 present early-stage adenocarcinoma. For each patient, a different number of samples was available (ranging from one to eight), with a total of 50 samples showing BE and cancerous tissue areas and 50 images showing only BE without cancer. The injured tissue in cancerous images has been delineated by five different endoscopy experts. Figure 5.2 shows some samples and their respective delineation performed by the experts.

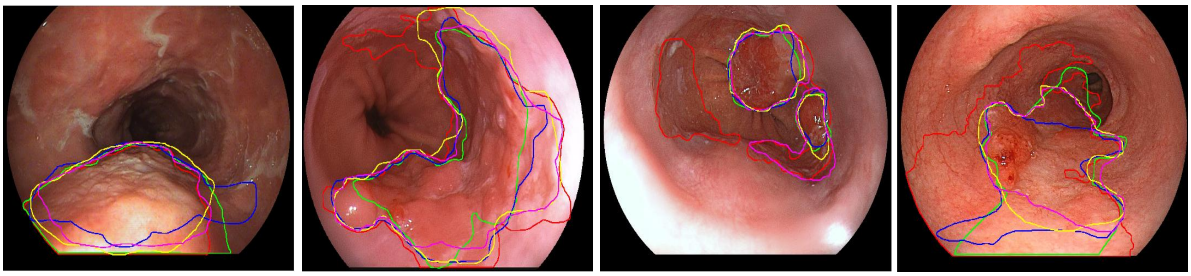


Figure 5.2: Four MICCAI database samples with their respective delineations provided by five different experts.

5.3.1.2 Augsburg Dataset

A dataset provided by the Augsburg Klinikum, Medizinische Klinik III, was also used for the experiments. Such dataset is composed of 76 endoscopic images (esophagus) obtained from different patients with adenocarcinoma (34 samples) and BE (42 samples). The images were annotated (manual segmentation of the adenocarcinoma’s and Barrett’s areas) by an expert from the Augsburg Klinikum. The ground-truth diagnosis was validated by biopsy process. Some Augsburg dataset samples can be observed in Figure 5.3.

5.3.2 Pre-processing

Concerning the pre-processing step, the images were split into patches to be used in different approaches for classification purposes. Considering that databases present different image resolutions, the patch size was chosen in order to cover the entire image without overlapping.

[1] https://endovissub-barrett.grand-challenge.org/

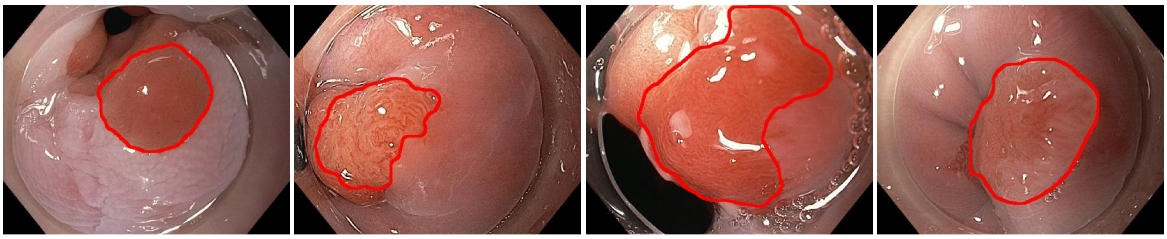


Figure 5.3: Four Augsburg database samples with their respective delineations provided by the expert.

Regarding the MICCAI database, each image was split into 48 patches of 200×200 pixels, resulting in $4,800$ non-overlapped patches for the whole database. The Augsburg database images were split into 20 patches per image, with sizes of 270×270 and resulting in $1,520$ non-overlapped patches.

Additionally, it is important to notice that for the patch labeling process, the experts' annotations were considered for both datasets. Concerning the MICCAI database, the intersection area of the five experts' delineations was considered the correct adenocarcinoma region including the fuzzy delineation area (area of confusion among the delineations). Concerning patches that cross this region, the numbers of cancerous and non-cancerous pixels were compared, being the final label defined by the majority of pixels inside the patch. With respect to the Augsburg data, we used the only delineated area available. Notice that an analogous procedure for labeling patches was employed as well.

5.3.3 Feature Extraction

In order to consider a color-and-texture evaluation protocol for the automatic identification of BE and adenocarcinoma, the integrative single-channel co-occurrence matrix was applied. For each color channel, SCMs and Haralick features were computed and further used for learning purposes. Notice that the same set of features were extracted from the gray-scale images, which were obtained using the mean pixel values of each channel.

The experimental protocol was composed of three distinct evaluations: (i) first, we considered color and texture information from each channel separately, (ii) then the same set of features (i.e., from each channel) were concatenated to produce a single feature vector, and (iii) the color and texture extraction techniques used previously were also considered for the gray-scale images.

Since each color channel and gray-scale feature vector consist of 16 elements, the composite descriptors (RGB) comprise 48 features (i.e., 3×16). Further, the Principal Component

Analysis (PCA) was applied to reduce the number of features to a 16-dimensional space using the Single Value Decomposition and the covariance matrix approach for the largest eigenvalues definition, thus ending up with the same single-channel descriptor dimension. For all approaches, the SCMs were calculated with three different radii (i.e., 1, 5 and 10) to assess the impact on their representation concerning the classification results.

5.3.4 Classification

After the feature extraction using the integrative single-channel co-occurrence matrices, the descriptors from the databases were used as input to the following supervised learning techniques:

- OPF: supervised classifier with complete graph proposed by Papa et al. [Papa, Falcão e Suzuki 2009, Papa et al. 2012];
- SVM-RBF: Support Vector Machines with Radial Basis Function kernel and parameters optimized by cross-validation [Chang e Lin 2011];
- Bayes: standard Bayesian classifier.

Regarding the OPF and SVM-RBF classifiers, we used the open-source libraries LibOPF [Papa e Falcão] and LibSVM [Chang e Lin 2011], respectively. With respect to the Bayesian classifier, we employed our own implementation.

5.3.5 Approaches

This work employs three different approaches for the database pre-processing and classification: patch-based, patient patch-based, and image-based approach. Regarding the patch-based approach, 80% of all patches were randomly selected for training, while 20% of the remaining ones were used for testing purposes, being such partitioning process employed for 20 runs for both databases. Therefore, the patch-based classification step was conducted to discriminate between patches from BE and adenocarcinoma classes without taking into account information about the patients. Concerning MICCAI dataset, $3,840$ patches were randomly selected for the training step and 960 patches were used for the testing set. With respect to the Augsburg database, the training set was composed of $1,216$ random patches, and the test set was composed of the 304 remaining patches.

Concerning the patient patch-based approach, the patient information was used for the patch selection protocol. The available number of patches for this approach was the same available for the previous one (patch-based approach), but the difference here was related to the protocol applied to the definition of training and testing sets. In this experiment, we used the well-known “leave-one-patient-out cross-validation" (LOPO-CV), i.e., $n-1$ patients are used for training and the remaining one is used to evaluate the model, where n stands for the number of patients. This procedure is repeated until all patients have been evaluated.

Finally, the last experiment, i.e., image-based approach, uses the same 20-fold cross-validation protocol applied to the first approach (i.e., patch-based) with 80% for training and 20% for testing purposes. However, the descriptors were now extracted from the full images. Notice that the same protocol was applied to the Augsburg database. Figure 6.4 illustrates the approaches mentioned above.

5.4 Experiments

In this section, we present the experiments used to evaluate the three proposed approaches. The discrimination between positive and negative samples to adenocarcinoma was performed using OPF, SVM-RBF and Bayesian classifier (hereinafter called “Bayes"). The results are presented and discussed for each approach and database. All experiments were conducted on an 8Gb-memory computer equipped with an Intel® Core i5 - 2.30 GHz processor. Additionally, we employed our implementation of the SCM approach in C++ language.

In this work, we adopted the following sensitivity (S), specificity (P), accuracy (A), and $F1$ Score ($F1$) measures:

$$S = \frac{TP}{TP+FN} * 100, \tag{5.2}$$

$$P = \frac{TN}{TN+FP} * 100, \tag{5.3}$$

$$A = \frac{TP+TN}{TP+TN+FP+FN} * 100, \tag{5.4}$$

and

$$F1 = \frac{2 \cdot S \cdot P}{S+P}, \tag{5.5}$$

where TP and TN stand for the true positives and true negatives, respectively, and FN and FP denote the false negatives and false positives, respectively.

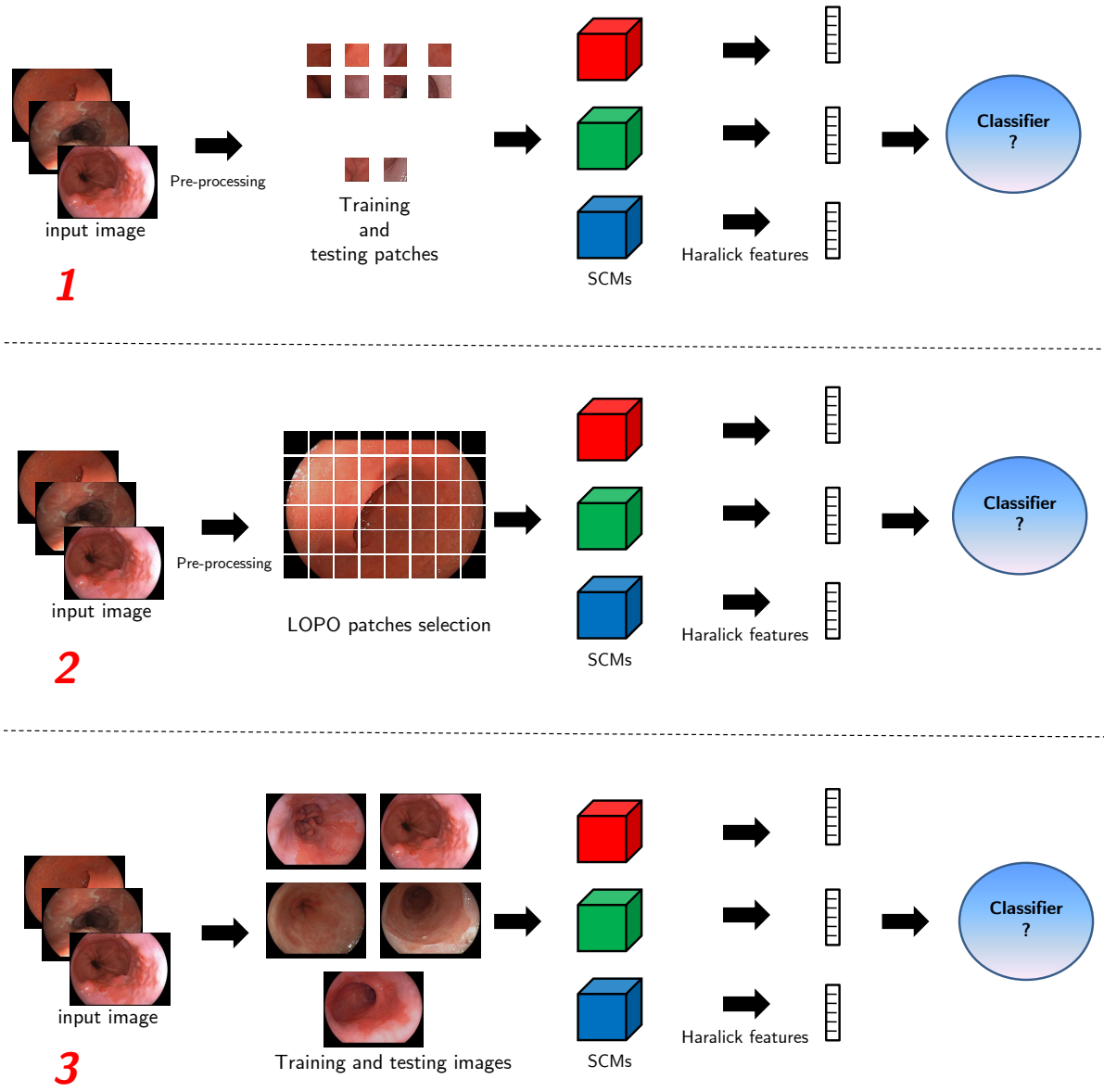


Figure 5.4: Approaches used in the experiments: (1) patch-based, (2) patient patch-based, and (3) image-based approach.

Table 5.1 presents the average results regarding sensitivity, specificity, accuracy, and $F1$ score concerning the patch-based approach. Since we considered different values for the radius used in SCM technique, column “Radius" contains the values that lead to the best results. With respect to the Augsburg database, the best results were obtained using the OPF classifier in the blue channel, with sensitivity, specificity, accuracy and $F1$ values of 66.6%, 75.6%, 70.2%, and 70.8%, respectively. Concerning the SVM-RBF classifier, the best results were obtained on the red channel, with values of 58.6% of sensitivity, 87.5% of specificity, 81.4% of accuracy, and 70.2% of $F1$ score. The Bayesian classifier provided the best results in the red channel features as well, with values of 58.1% for sensitivity, 83.5% of specificity, 75.3% of accuracy,

and 68.5% of $F1$.

Since the $F1$ score presents a trade-off response between the sensitivity and specificity values, the values in bold stand for the best approaches with respect to such measure. The accuracy can be very dependent on the sensitivity/specificity values when we have a unbalanced database. We observed that specificity always influenced considerably the accuracy values, once the number of samples positive to cancer was usually lower when compared to the negative examples.

Concerning the MICCAI database, OPF classifier provided the best results in the blue channel once again, with values of sensitivity, specificity, accuracy, and $F1$ equal to 72.6%, 77.2%, 76.6%, and 74.8%, respectively. The SVM-RBF classifier showed the best results in the blue channel as well, with values of 63.9%, 76.7%, 70.1%, and 69.7% for sensitivity, specificity, accuracy, and $F1$ score, respectively. The Bayes classifier obtained the best results in the PCA-RGB features, with sensitivity, specificity, accuracy, and $F1$ values of 58.3%, 80.3%, 75.6%, and 67.6%, respectively.

Table 5.2 presents the average results of sensitivity, specificity, and accuracy concerning the patient patch-based approach. Regarding the Augsburg database, the best results considering the $F1$ values were obtained using the OPF classifier in the blue channel, with sensitivity, specificity, accuracy and $F1$ values of 63.0%, 77.4%, 73.5%, and 69.5%, respectively. With respect to the SVM-RBF classifier, the best results were obtained in the green channel, with values of 63.6% for sensitivity, 74.4% for specificity, 70.9% for accuracy, and 68.6% for $F1$. The Bayes classifier provided reasonable results in the green channel as well, with values of 65.8% for sensitivity, 70.5% for specificity, 69.1% for accuracy, and 68.1% for the $F1$ score.

Regarding MICCAI dataset, the OPF classifier provided the best results in the blue channel, with values of sensitivity, specificity, accuracy and $F1$ Score equal to 71.6%, 72.9%, 72.3%, and 72.2%, respectively. The SVM-RBF classifier obtained the best results using the PCA-RGB features, with values of 61.9%, 79.5%, 75.4%, and 69.6% of sensitivity, specificity, accuracy and $F1$, respectively. The Bayes classifier with the PCA-RGB features provided the best results as well, with values of sensitivity, specificity, accuracy and $F1$ of 60.1%, 82.4%, 78.6%, and 69.1%, respectively.

Table 5.3 presents the average results of sensitivity, specificity, accuracy and $F1$ score concerning the classification in the image-based protocol. Concerning the Augsburg dataset, the best results were obtained using the OPF classifier in the red channel, with sensitivity, specificity, accuracy and $F1$ values of 63.5%, 66.2%, 64.5%, and 64.82%, respectively. Using the SVM-RBF classifier, the best results were obtained using the PCA-RGB features, with values

of 37.7% for sensitivity, 83.4% for specificity, 66.9% for accuracy and 51.93% for the $F1$ score. The Bayes classifier obtained its best results with the PCA-RGB features as well, with values of 55.5% for sensitivity, 77.6% for specificity, 62.7% for accuracy, and 64.72% for $F1$.

With respect to the MICCAI database, OPF classifier obtained the best results in the red channel, with values of sensitivity, specificity, accuracy and $F1$ equal to 60.7%, 70.8%, 68.4%, and 65.36%, respectively. The SVM-RBF classifier achieved the best results using the PCA-RGB features, with values of 49.5%, 78.2%, 71.1%, and 60.62% for sensitivity, specificity, accuracy and $F1$ values, respectively. The Bayes classifier obtained the best results in the red channel, with values of sensitivity, specificity, accuracy and $F1$ equal to 61.7%, 73.5%, 67.0%, and 67.1%, respectively.

Table 5.1: Mean values concerning the patch-based approach.

		OPF					SVM-RBF					Bayes				
Database	Channel	S	P	A	$F1$	Radius	S	P	A	$F1$	Radius	S	P	A	$F1$	Radius
Augsburg	gray	53.7	87.5	80.2	66.6	10	59.9	82.7	77.2	69.5	10	55.5	75.7	70.9	64.4	5
	red	54.4	82.7	77.3	65.7	10	58.6	87.5	81.4	**70.2**	10	58.1	83.5	75.3	**68.5**	10
	green	43.4	93.5	87.8	59.3	5	47.4	90.8	87.2	62.3	10	39.5	89.7	86.8	54.9	10
	blue	66.6	75.6	72.0	**70.8**	5	60.2	77.8	71.0	67.9	10	53.5	80.2	75.3	64.2	10
	PCA-RGB	57.7	83.2	73.3	68.1	5	55.7	82.3	76.7	66.4	10	56.8	81.4	76.8	66.9	10
MICCAI	gray	64.3	85.5	83.4	73.4	10	66.7	79.5	79.1	68.8	5	47.8	78.4	73.4	59.4	10
	red	51.2	90.4	87.5	65.4	10	47.9	91.1	86.9	62.8	10	54.7	85.3	80.0	66.7	10
	green	49.7	90.4	82.7	65.3	5	51.4	90.8	89.1	65.6	10	41.5	87.3	81.5	56.3	5
	blue	72.6	77.2	76.6	**74.8**	10	63.9	76.7	70.1	**69.7**	10	50.0	86.7	81.3	63.4	10
	PCA-RGB	65.1	82.1	77.5	72.6	10	57.5	80.4	73.2	67.0	10	58.3	80.3	75.6	**67.6**	10

Table 5.2: Mean values concerning the patient patch-based.

		OPF					SVM-RBF					Bayes				
Database	Channel	S	P	A	$F1$	Radius	S	P	A	$F1$	Radius	S	P	A	$F1$	Radius
Augsburg	gray	52.3	71.0	59.4	60.2	10	54.0	78.2	72.2	63.9	5	50.5	73.4	64.9	59.8	10
	red	52.6	67.4	68.3	59.1	10	56.1	70.9	64.7	62.6	10	54.8	63.9	60.1	59.0	10
	green	70.2	67.2	68.1	68.7	10	63.6	74.4	70.9	**68.6**	10	65.8	70.5	69.1	**68.1**	5
	blue	63.0	77.4	73.5	**69.5**	10	59.7	78.7	72.9	67.9	10	52.0	77.6	70.4	62.3	10
	PCA-RGB	56.2	83.4	76.3	67.2	10	61.4	72.3	66.8	66.4	10	55.2	76.8	68.3	64.2	10
MICCAI	gray	54.3	82.8	73.6	65.6	10	60.9	76.7	67.8	67.9	10	47.0	74.1	63.2	57.5	5
	red	55.3	81.0	64.6	65.7	5	50.2	81.9	62.6	62.2	5	55.1	71.7	60.8	62.3	5
	green	50.9	84.5	71.1	63.5	10	50.0	85.2	73.9	63.0	5	49.9	81.4	69.5	61.9	10
	blue	71.6	72.9	72.3	**72.2**	10	63.8	75.5	69.1	69.1	10	59.6	77.2	69.1	67.3	10
	PCA-RGB	71.1	71.7	71.5	71.4	10	61.9	79.5	75.4	**69.6**	10	60.1	82.4	78.6	**69.1**	5

5.5 Discussion and Conclusions

In this paper, we dealt with the problem of computer-assisted Barrett's esophagus and esophageal adenocarcinoma evaluation using endoscopy images. BE stands for an illness that

Table 5.3: Mean result values concerning the image-based approach.

		OPF					SVM-RBF					Bayes				
Database	Channel	S	P	A	$F1$	Radius	S	P	A	$F1$	Radius	S	P	A	$F1$	Radius
Augsburg	gray	55.5	60.1	57.5	57.7	10	33.7	84.5	63.8	48.2	5	30.7	64.5	51.3	41.6	5
	red	63.5	66.2	64.5	**64.8**	10	29.7	89.5	67.6	44.6	10	57.1	55.6	56.3	56.3	5
	green	52.8	64.2	59.0	57.9	10	30.3	85.5	64.7	44.7	10	46.8	71.8	59.9	56.7	5
	blue	44.9	69.7	58.8	54.6	10	33.9	85.9	65.6	48.6	10	50.3	53.6	50.1	51.9	5
	PCA-RGB	52.0	64.2	57.8	57.5	10	37.7	83.4	66.9	**51.9**	10	55.5	77.6	62.7	**64.7**	5
MICCAI	gray	59.7	52.7	55.5	56.0	10	50.3	55.4	50.7	52.7	10	47.5	64.8	56.3	54.8	5
	red	60.7	70.8	68.4	**65.4**	10	38.9	86.6	79.2	53.7	10	61.7	73.5	67.0	**67.1**	10
	green	54.9	69.7	64.8	61.4	10	37.8	83.9	73.2	52.1	10	44.4	84.6	72.3	58.2	10
	blue	57.8	63.1	59.6	60.3	10	35.8	86.5	76.4	50.6	5	55.5	60.4	58.4	57.9	10
	PCA-RGB	60.0	66.3	65.2	63.0	10	49.5	78.2	71.1	**60.6**	10	54.9	79.4	72.5	64.9	5

is visually confused with adenocarcinoma, requiring more precise ways for its early detection and prevention.

We observed that only a very few works attempted at coping with the problem of automatic BE identification using image processing and machine learning techniques to date. In this work, we fostered the research towards such area and introduced the use of single channel Color Co-occurrence Matrices in the feature extraction step for automatic BE recognition, as well as we showed how each RGB channel could perform compared to the gray-scale image evaluation. The experimental results were considered over two databases: (i) MICCAI 2015 and (ii) Augsburg. For both scenarios, we evaluated three different approaches and supervised learning techniques for classification purposes.

As one can observe in the previous section, each approach presents a particular and interesting result that deserves attention. Considering the patch-based approach (Table 5.1), the results over Augsburg data highlighted that cancerous patches are harder to be identified than non-cancerous ones, thus explaining low values of sensitivity and higher values of specificity. However, the use of blue-channel SCMs associated with the OPF classifier provided the higher sensitivity and $F1$ values. With respect to MICCAI database, the results presented a similar behavior to those obtained over Augsburg ones. The blue-channel SCMs combined with the OPF classifier achieved the higher $F1$ scores when compared to the other classifiers, thus suggesting it can be a strong learning technique for color-and-texture feature classification. For both databases, the number of non-cancer patches was considerably higher than the cancerous ones, thus explaining the higher specificity values. Although SVM-RBF and Bayes obtained satisfactory results, they were outperformed by OPF.

Concerning the patient patch-based approach (Table 5.2), the best results over Augsburg data were obtained with the OPF classifier and blue channel SCMs. The MICCAI dataset results showed a better performance, with the best values achieved by OPF and blue-channel SCMs as

well. Both SVM-RBF and Bayes classifiers were outperformed by OPF. Once again, SVM-RBF provided better sensitivity values in some experiments, but not the overall best results.

Since the image-based approach (Table 5.3) makes use of descriptors obtained from the entire image, the tendency is to achieve the worst results due to the presence of both cancerous and non-cancerous regions in the very same image labeled as cancer. Surprisingly, the results presented satisfactory classification rates for both databases. Concerning the Augsburg and MICCAI databases, the best results were achieved with the OPF classifier and red-channel SCMs. The SVM-RBF showed the worst performance in this approach, with low sensitivity but high specificity values. The image-based approach demonstrates the generalization strength of the SCMs over the entire images. Even with different regions in cancer-labeled images, the obtained feature vector can provide good generalization for the classification step.

Considering all approaches, it is relevant to point out that the results using color-channel features outperformed the gray-scale ones in all experiments, thus corroborating the importance of the color-texture analysis. The blue channel results obtained for the patch and patient-based approaches suggest that, for local evaluation, the blue channel present a more accurate and robust way of description, while the red channel may provide a better global evaluation of the BE and adenocarcinoma problem, according to the global results provided by the image-based approach. The blue results corroborate the ones obtained by Ilgner et al. [Ilgner et al. 2003] in which laryngoscopy images presenting or not diseased tissue were classified using colored-texture descriptors, being blue the color channel that provided the best classification rate (81.4%). PCA-RGB features showed a very well performance, achieving the best results between the channels for some classifiers and approaches. With respect to the SCMs, feature vectors calculated with large distances (i.e., higher values of **D**) showed better results when compared to short distance ones (in this case, **D** = 5 or **D** = 10 always presented better results than the ones obtained using **D** = 1). Such premise is also relevant, reinforcing the importance of the neighboring information during the CMs calculation, suggesting that higher values of **D** may provide better generalization abilities for classification purposes. The experiments pointed out that SCMs are suitable to handle BE automatic identification, and there must be a trade-off between the sensitivity and specificity values to compose a cohesive diagnosis result for the BE and adenocarcinoma distinction.

Concerning the previous results obtained for the BE and adenocarcinoma classification in the literature, this one, in particular, proposes a new protocol of image evaluation, in which the images are split into patches, so the labeling problem is changed. It is well-known that once we work with full images, cancerous and non-cancerous regions receive the same label,

but when the problem is extended to the patches, the labeling problem becomes less critical. With the patches labeling, even with images that present BE and adenocarcinoma, the misclassification of regions do not exist, once each patch will be labeled according to the previous annotation provided by the experts. Regarding the classification results, handling with patches may improve the results because of this accurate labeling definition that does not happen in the full-image approach. The comparison with previous works can be performed with the third approach (image-based approach) for the MICCAI 2015 database. Souza Jr. et al. conducted two different works with a similar image-based evaluation in such database, but using SURF [Souza Jr. et al. 2017] and SIFT [Souza Jr. et al. 2017] descriptors associated to a large number of classifiers. The accuracy, sensitivity and specificity results obtained in this work could not outperform the ones using SURF and SIFT features for the full-images approach. However, this work provides two very important contributions: (i) the introduction of the single-channel co-occurrence matrices technique for the BE and adenocarcinoma description and (ii) the evaluation based on patch-based approaches in which very promising results could be provided, suggesting that the proposed descriptor, associated with a local representation of the problem, can become a strong resource for the BE and adenocarcinoma context evaluation.

In regard to future works, we aim at considering four major new tasks: (i) the multi-channel implementation of co-occurrence matrices [Palm 2004] instead of the single-channel approach used in this work; (ii) the reduction of the feature vector dimensionality using feature selection techniques, and (iii) the use of the methodology used in this work as an end-to-end approach to aid physicians during the diagnosis process and; (iv) a scale-evaluation of the BE and adenocarcinoma context using the proposed color co-occurrence matrice descriptor for many levels of image scale, providing descriptors based on a scale-space approach.

5.6 Chapter's Considerations

The evaluation of cancerous tissues in BE samples based on color and texture shows a first logical way of conducting such a task, and has been performed with remarkable results [van der Sommen et al. 2016, Almond e Barr 2012] for a long time. However, until the proposal of this study, experiments combining both color and texture information in one encoding technique had never been conducted in the evaluation of cancer-and-BE diseases. Then, to cope with such a task, we proposed the use of color-cooccurrence matrices [Palm 2004] as a new way of describing and differentiating BE and adenocarcinoma in endoscopic samples.

In the conducted experiments, coocurrence-descriptors based on each RGB channel were

calculated to be used in the classification task. With that, we could observe the impact of each color channel in the description of cancerous tissues and its relation to texture representation. Besides, we also aimed to assess a new feature vector composed of the combination of individual features of each RGB color channel, trying to enhance the correct classification of cancer once again.

Regarding the key points' spatial condition, as attested in previous Chapters, here we conducted experiments over full-images and patch-based images. Patches mean a small part of full-images in which less misunderstanding about the tissue condition could happen once the label of each patch is defined from experts' annotation, and so, the features can be better defined for each class to be further classified. Hence, three classification approaches were proposed in this work, based on patches, patients, and full-images. As such, the color-cooccurrence description could be deeply evaluated in all three possible scenarios.

As expected, patch-based approaches could achieve better results, considering the smaller misunderstandind and misrepresentation of cancerous tissues. For the descriptor itself, we observed that the BLUE channel provided the best description, delivering the best overall results (as shown by [Ilgner et al. 2003]). Moreover, the description based on color and texture has shown promising results but could not outperform the ones obtained using object detector techniques. The OPF classifier showed once again a high generalization capability, as the best classifier so far.

Moreover, a new validation dataset was introduced for the first time in this manuscript, ensuring the generalization potential of the proposed technique, the Augsburg dataset, composed of 76 samples of different patients showing BE and BE and adenocarcinoma. For continuing our main study, more ways of describing handcrafted features would be proposed in more sophisticated ways of correlating its natures and spatial meanings.

Chapter 6

Learning Visual Representations with Optimum-Path Forest and its Applications to Barrett's Esophagus and Adenocarcinoma Diagnosis

Chapter 6 is continuity of the work started in Chapter 4. Here the introduction of the unsupervised Optimum-Path Forest classifier for learning visual dictionaries in the context of automatic Barrett's esophagus and adenocarcinoma diagnosis was proposed. One can observe, the behavior of BE and adenocarcinoma features based on bag-of-visual-words representation was evaluated. The work was published in the Neural Computing and Applications journal [Souza Jr. et al. 2019].

6.1 Introduction

Pattern classification has been paramount in the last decades, mainly due to the increasing number of applications that require some intelligent-decision-making mechanism. The standard pipeline adopted for so many years follows a robust but straightforward workflow: (i) feature extraction, (ii) model learning, and (iii) classification outcomes. The former step can be performed using handcrafted features or information learned through deep learning approaches. In this latter case, one may not know what kind of information the model is learning, since the set of outcome values that minimizes some loss function is the one employed in the model learning step. Handcrafted features require a more knowledgeable personnel, which is usually in charge of selecting and extracting features that matter when performing pattern classification.

Describing images using their most important information, the so-called “points of interest"(PoIs) or key points, has been an active area of interest by many researchers worldwide. Notable approaches have been proposed in the literature to compute those points, which somehow aim at capturing subtle information that is less variant to geometric transformations such as rotation, and translation, among others. Scale-Invariant Feature Transform [Lowe 2004], Speeded-Up-Robust-Features [Bay et al. 2008], and Accelerated-KAZE features (A-KAZE) [Alcantarilla, Nuevo e Bartoli 2013] are some examples.

However, the main problem related to the mentioned approaches concern the final feature vector. Since the number of PoIs may vary from one image to another, the feature vectors used to represent the images shall have different dimensions. To overcome this issue, an additional step called “quantization" is required (some works refer to this step as the “codebook generation" [Fei-Fei e Perona 2005]). In a nutshell, given a training set composed of PoIs extracted from all training images, we can build a “bag"(i.e., a visual dictionary) with the most representative PoIs (from now on called “visual words"). Further, for each training image and each of its PoIs, we can find the “closest" visual word in the bag and build up a histogram that stores the number of times a visual word is nearest to each PoI from the training images. Therefore, the final feature vector of each training image will be that histogram with a dimensionality that corresponds to the number of visual words (i.e., the size of the dictionary or bag). Essentially, that is the main reason such approaches are usually referred to as “bag-of-visual-words" [Csurka et al. 2004].

Bag-of-visual-words have been widely used in the literature for a number of different purposes, such as video-based action recognition [Peng et al. 2016], retinal health diagnosis [Koh et al. 2018], and perivascular spaces categorization in brain data [González-Castro et al. 2016], among others. Nonetheless, one still has two problems to face regarding the BoVW approach: (i) how to find out the most representative visual words, and (ii) how to establish a proper bag size, i.e., the number of visual words. Notice that both issues are pretty much crucial since they are in charge of the feature vector composition and dimensionality.

To cope with the first issue, i.e., finding out the most representative visual words, two approaches are commonly used: (i) random sampling and (ii) clustering. The former randomly selects a given number of visual words to compose the bag. On the other hand, clustering-based approaches make use of some unsupervised learning algorithm (usually k-means) to group the visual words, and the most representative ones (i.e., centroids) are elected to compose the dictionary [Afonso et al. 2012]. However, randomly choosing visual words does not lead to good results, and the usage of certain unsupervised learning algorithms turns out to be a problem

since most of them require the number of clusters (i.e., the bag size) beforehand.

Therefore, clustering techniques that do not require a priori information about the data are usually preferred. Among several techniques that have been proposed in the literature, one is gaining attention daily due to its effectiveness and efficiency in different research areas. The OPF is a framework for the design of pattern classifiers based on graph partition. In short, OPF-based classifiers work on a reward-competition process, in which previously selected samples called “prototypes" try to conquer other samples by offering them optimum-path costs. Once a sample is conquered by another one, it receives its label and a “mark"(i.e., a predecessor map) that reveals its conqueror. The Optimum-Path Forest framework comprises supervised [Papa, Falcão e Suzuki 2009, Papa et al. 2012, Papa, Fernandes e Falcão 2017], unsupervised [Rocha, Cappabianco e Falcão 2009], and semi-supervised [Amorim et al. 2016] versions that have been widely employed in a number of applications, from remote sensing [Pisani et al. 2014, Nakamura et al. 2014] to human intestinal parasites identification [Suzuki et al. 2013], just to cite a few.

One particular strength of unsupervised OPF concerns the fact it does not require the number of clusters beforehand, i.e., it finds clusters on-the-fly. Such feature is quite interesting in the context of BoVW generation since we skip the problem of choosing suitable bag sizes. As far as we are concerned, only two works attempted at using OPF in the context of BoVW: (i) Papa and Rocha [Papa e Rocha 2011] evaluated the supervised OPF for image categorization using visual words, and further (ii) Afonso et al. [Afonso et al. 2012] studied the impact of using unsupervised OPF for learning proper visual dictionaries.

We are particularly interested in the application of such technique for the recognition of Barrett’s esophagus (BE), which happens to be a side effect of some reflux diseases. BE comprises a very severe and growing disease in the last decades, and since BE is often not identified properly at the early stages, it may evolve to a more aggressive version, and even to cancer. However, the early diagnosis of dysplastic tissue in BE diagnosed patients may provide very high rates of remission after the treatment [Dent 2011, Sharma et al. 2016, Phoa et al. 2016]. There are several endoscopic techniques for the BE diagnosis and detection, such as chromoendoscopy and narrow-band imaging, but the human screening for the injured region definition is still often misclassified by endoscopists, once the region does not present enough goblet cells in biopsy or the experts refuse to use the recommended procedure for extensive biopsies [Sharma et al. 2015]. Moreover, computer-assisted diagnosis may bring precision and accuracy to the BE screening and evaluation, once this task can be very influenced by the human factor [Souza Jr. et al. 2017, Souza Jr. et al. 2018, Souza Jr. et al. 2017]. To the best of our knowledge,

only one very recent work coped with BE identification using OPF. Souza et al. [Souza Jr. et al. 2017] introduced the supervised OPF for Barrett's esophagus automatic identification using features based on BoVW. The authors considered both random- and k-means-based sampling strategies to build the visual dictionaries and then used OPF for classification purposes.

Some works that dealt with endoscopic image analysis can be referred to as well, but that is still an emerging area of research [Souza Jr. et al. 2018]. Seibel et al. [Seibel et al. 2008] developed a low-cost but high-performance technology to assist the diagnosis of BE and esophageal cancer. However, their primary contributions rely on hardware advances rather than software ones. The work presented by van der Sommen [van der Sommen et al. 2016] aimed at using machine learning techniques to detect early neoplasia in Barrett's esophagus, and Swager et al. [Swager et al. 2016] addressed the very same context mentioned above but using volumetric laser endoscopy images.

Klomp et al. [Klomp et al. 2017] proposed new features for computer-aided Barrett's esophagus identification, and Hassan and Haque [Hassan e Haque 2015] used endoscopy videos obtained from wireless capsules to assess gastrointestinal hemorrhages. Later on, Seguí et al. [Seguí et al. 2016] used the same source of images (i.e., wireless capsules) together with Deep Convolutional Neural Networks for intestine motility characterization. Mendel et al. [Mendel et al. 2017] started the study of deep learning application to the BE and adenocarcinoma evaluation problem.

The major problems around the computer-assisted systems developed for the BE and adenocarcinoma evaluation are related to the type of technique to provide a correct description of the injured areas and which classification techniques should be designed for the problem. These problems are related to all proposed works, and considering the high potential in this research area, new ways to describe the injured areas (that are very similar), and the evaluation of different classifiers can deliver important and substantial improvements to the precision and correct differentiation of both.

As one can observe, Barrett's esophagus automatic identification using machine learning techniques presents a growing interest in the last years. Therefore, there is plenty of room for new works that employ techniques that were not considered in such a context. In this work, we extended and outperformed the approach proposed by Souza et al. [Souza Jr. et al. 2017] by learning proper visual dictionaries using unsupervised OPF, as well as we introduced a variant of supervised OPF (OPF_{knn}) proposed by Papa et al. [Papa, Fernandes e Falcão 2017] in the context of BE identification. The OPF was never applied to such problem in the visual learning step, and this could deliver, besides the novelty in the feature vector calculation, a new way to

evaluate the key points provided by the feature extraction techniques. Last but not least, we introduce the A-KAZE feature extraction technique for the calculation of the key points, for comparison with SURF and SIFT, previously adopted for the BE and adenocarcinoma differentiation context [Souza Jr. et al. 2018]. The results presented in this paper are close to some state-of-the-art recognition rates [Mendel et al. 2017], and it features recent advances to BE automatic identification by means of machine learning and computer vision.Therefore, the main contributions of this paper are five-fold:

- to extend and outperform the recent results obtained by Souza et al. [Souza Jr. et al. 2017] in which the evaluation of BE and adenocarcinoma context were performed using: (i) SURF and SIFT techniques for key points calculation, (ii) k-means and random techniques for the bag-of-visual-words calculation, and (iii) OPF and SVM classifiers for the classification task;
- to introduce OPF$_{knn}$ [Papa, Fernandes e Falcão 2017] for BE and adenocarcinoma automatic diagnosis, considering that Souza et al. [Souza Jr. et al. 2017] employed only the complete graph version of OPF classifier for the classification task;
- to introduce A-KAZE features for the aforementioned context, once such technique has been largely applied in the literature for image description and retrieving;
- to extend the work by Afonso et al. [Afonso et al. 2012] with a more robust evaluation of the unsupervised OPF for learning visual dictionaries;
- to introduce a new representation of feature extraction techniques (such as SURF and SIFT) based on their most representative words in the feature space using the OPF clustering technique.

The remainder of this paper is organized as follows. Sections 6.2 to 9.3 present a theoretical background of unsupervised OPF and the methodology adopted in this work, respectively. Section 9.4 discusses the experiments, and Section 9.5 states conclusions and future works.

6.2 Unsupervised Learning with Optimum-Path Forest

In this section, we briefly present the theoretical background related to unsupervised OPF, which is used to learn proper visual dictionaries.

Let $\mathscr{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ be an unlabeled dataset such that $\mathbf{x}_i \in \Re^n$ stands for a feature vector extracted from some sample (i.e., images in our case) related to the problem to be addressed.

Additionally, let $\mathscr{G} = (\mathscr{D}, \mathscr{A}_k)$ be a graph derived from that dataset, which means $\mathscr{D}$ denotes the set of graph nodes (i.e., vertices) and $\mathscr{A}_k$ stands for a k-nearest neighbors adjacency relation.

In a nutshell, the OPF working mechanism is based on a reward-competition problem, where some samples called “prototypes" employ a competitive process among themselves to conquer the other samples from the dataset $\mathscr{D}$. Such competition ends up partitioning $\mathscr{D}$ into optimum-path trees (OPTs), which are rooted at each prototype node. It is worth mentioning that a sample that belongs to a given OPT is more “strongly connected" to the root and samples of that tree than to any other in the forest (i.e., a collection of all trees in the graph).

At a glance, the whole process can be summarized in the following steps:

1. To establish a proper neighborhood size and build up $\mathscr{A}_k$ (i.e., to find out “suitable" k values);
2. To elect the prototypes and Learning Visual Representations with Optimum-Path Forest and its Applications to Barrett’s esophagus and Adenocarcinoma Diagnosis
3. To start the competition process.

Concerning step 1), a number of different approaches to cope with the task could be considered. Rocha et al. [Rocha, Cappabianco e Falcão 2009] proposed to compute the best value of k (i.e., the neighborhood size), say that k^*, as the one that minimizes the normalized graph cut, which is a measure that considers both the dissimilarity between clusters as well as the similarity within the groups of samples [Shi e Malik 2000].

Soon after computing k^*, the next move concerns finding the prototypes (i.e., step 2), also known as the “roots of the trees". Such essential samples are in charge of ruling the competition process that ends up partitioning the graph into OPTs (i.e., clusters). Those samples will be used as the visual words to compose the final dictionary, as further discussed.

The supervised OPF proposed by Papa et al. [Papa, Falcão e Suzuki 2009] elects the prototypes as the nearest samples from different classes, which can be accomplished by computing a Minimum Spanning Tree (MST) over the training graph. Then, the samples from different classes that are connected in the MST are marked as prototypes. However, unsupervised OPF does not make use of labeled datasets, which motivated Rocha et al. [Rocha, Cappabianco e Falcão 2009] to elect the prototypes as the samples that are located at the center of the clusters. Such samples can be computed by assigning a density score $\rho(\mathbf{x}_i)$ for each dataset sample $\mathbf{x}_i \in \mathscr{D}$. That score is computed using a probability density function (pdf) given by a Gaussian

distribution considered in the neighborhood of each sample as follows:

$$\rho(\mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}k} \sum_{\forall \mathbf{x}_j \in \mathscr{A}_k(\mathbf{x}_i)} \exp\left(\frac{-d(\mathbf{x}_i, \mathbf{x}_j)}{2\sigma^2}\right), \tag{6.1}$$

where $i \neq j$ and $\sigma = d_{max}/3$. In this case, d_{max} stands for the maximum arc-weight in G. Using such formulation, $\rho(\mathbf{x}_i)$ considers all adjacent nodes for the probability computation purposes since a Gaussian function covers 99.7% of the samples within $d(\mathbf{x}_i, \mathbf{x}_j) \in [0, 3\sigma]$.

After computing Equation 6.1 for all nodes, the competition process among samples can take place. Each density value will be used to populate a priority queue, where the idea of the unsupervised OPF algorithm is to end up maximizing the cost of each sample, and thus partitioning the graph.

The definition of “cost" is based on paths on graphs, i.e., a sequence of adjacent samples with no cycles. Let $\pi_{\mathbf{x}_i}$ be a path with terminus at sample $\mathbf{x}_i$ and starting from some root $\mathscr{R}(\mathbf{x}_i)$, where $\mathscr{R}$ stands for the set of prototype samples. Additionally, let $\pi_{\mathbf{x}_i} = \langle \mathbf{x}_i \rangle$ be a trivial path (i.e., a path composed of a single sample) and $\pi_{\mathbf{x}_i} \cdot \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ the concatenation of $\pi_{\mathbf{x}_i}$ and the arc $(\mathbf{x}_i, \mathbf{x}_j)$ such that $i \neq j$.

The OPF algorithm assigns to each path $\pi_{\mathbf{x}_i}$ a value $f(\pi_{\mathbf{x}_i})$ given by a connectivity function $f : \mathscr{X} \to \Re$. In this context, a path $\pi_{\mathbf{x}_i}$ is considered optimum if $f(\pi_{\mathbf{x}_i}) \geq f(\tau_{\mathbf{x}_i})$ for any other path $\tau_{\mathbf{x}_i}$. Such sort of functions are known as “smooth functions", and they figure important constraints that ensure the theoretic correctness of the OPF algorithm [Falcão, Stolfi e Lotufo 2004].

Among different path-cost functions that have been proposed in the literature, unsupervised OPF employs the following formulation for $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathscr{D}$ such that $i \neq j$:

$$f(\langle \mathbf{x}_i \rangle) = \begin{cases} \rho(\mathbf{x}_i) & \text{if } \mathbf{x}_i \in \mathscr{R} \\ \rho(\mathbf{x}_i) - \delta & \text{otherwise,} \end{cases} \tag{6.2}$$

and

$$f(\pi_{\mathbf{x}_i} \cdot \langle \mathbf{x}_i, \mathbf{x}_j \rangle) = \min\{f(\pi_{\mathbf{x}_i}), \rho(\mathbf{x}_j)\}, \tag{6.3}$$

where $\delta = \min_{\forall (\mathbf{x}_i, \mathbf{x}_j) \in \mathscr{A}_k | \rho(t) \neq \rho(s)} |\rho(t) - \rho(s)|$. In a nutshell, δ stands for the smallest quantity required to avoid plateaus in the regions nearby the prototypes (i.e., areas with the highest density).

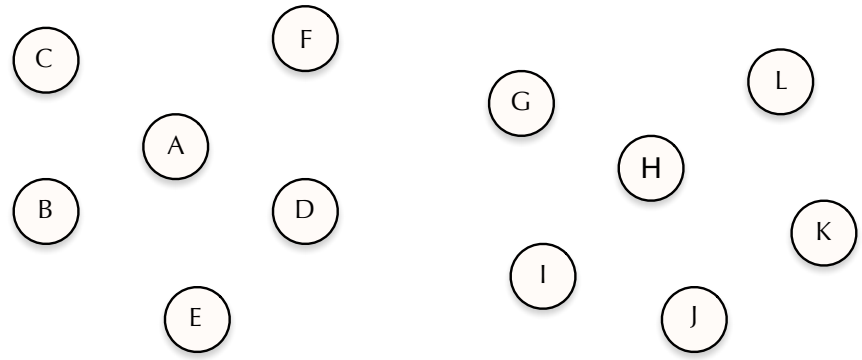
Among all possible paths $\pi_{\mathbf{x}_i}$ from the maxima of the pdf, the method assigns to sample $\mathbf{x}_i$ a final path whose minimum density value along it is maximum. Such final path value is

represented by a cost map $\mathscr{C}$, as follows:

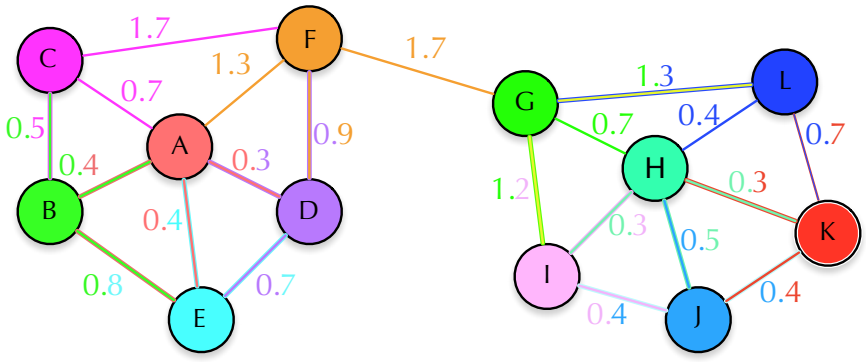
$$\mathscr{C}(\mathbf{x}_i) = \max_{\forall \pi_{\mathbf{x}_j} \in (\mathscr{D}, \mathscr{A}_k), i \neq j} \{f(\pi_{\mathbf{x}_j} \cdot \langle \mathbf{x}_j, \mathbf{x}_i \rangle)\}. \tag{6.4}$$

The OPF algorithm maximizes the connectivity map $\mathscr{C}(\mathbf{x}_i)$, $\forall \mathbf{x}_i \in \mathscr{D}$, by computing an optimum-path forest over the dataset. Such forest is encoded as a predecessor map $\mathscr{P}$ with no cycles that assigns to each sample $\mathbf{x}_i \notin \mathscr{R}$ its predecessor $\mathscr{P}(\mathbf{x}_i)$ in the optimum path from $\mathscr{R}$, or a marker *nil* when $\mathbf{x}_i \in \mathscr{R}$.

Figures 6.1 to 6.3 depict a toy example concerning the unsupervised OPF working mechanism. Figures 6.1a and 6.1b illustrate an unlabeled dataset and its 3-nearest neighbors graph, respectively (we assume $k = 3$ to explain step 1). For the sake of visualization purposes, we assigned the same color to each graph node and the arcs corresponding to its 3-nearest neighbors.



(a)



(b)

Figure 6.1: Toy example: (a) unlabeled dataset and its (b) 3-nearest neighbors graph.

Notice the arcs are also weighted by the distance (e.g., Euclidean distance) among their corresponding nodes. One can observe that some arcs and their weights are double-colored, which means their corresponding nodes share the very same 3-neighborhood.

Figure 6.2 illustrates the density computation step to further elect the prototypes (i.e., step 2). Therefore, given the arc-weights depicted in Figure 6.1b, we can use Equation 6.1 to compute $\rho(\mathbf{x}_i)$, $\forall \mathbf{x}_i \in \mathscr{D}$. Notice the density values are computed over the adjacency relation encoded by $\mathscr{A}_k$. One can realize that the samples located at the center of the clusters tend to be the ones with the highest value of ρ since they are connected by smaller arc-weights.

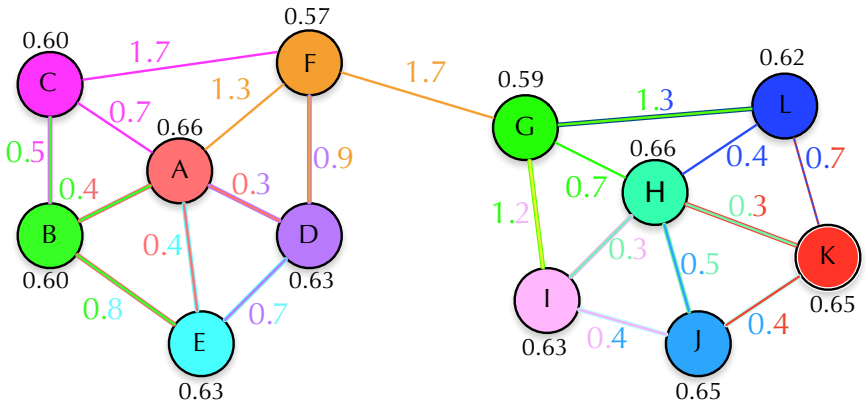


Figure 6.2: Computing the densities of each graph node according to its 3-neighborhood. The values under/over the nodes stand for their density values computed using Equation 6.1.

The density values are then stored in a priority queue (i.e., a max-heap) that pops out the sample $\mathbf{x}_i$ with the highest $\rho(\mathbf{x}_i)$. Concerning the toy example depicted in Figure 6.2, the first sample to come out of the queue is either 'H' or 'A' since both have the highest densities. Suppose 'H' has been added first to the queue. Since it has no predecessor, it is added to the set $\mathscr{R}$ and assigned $f(H) = \rho(\mathrm{H}) = 0.66$ according to Equation 6.2.

Further, the competition process (i.e., step 3) takes place. In short, sample 'H' evaluates its neighbors 'I', 'J', and 'K' to offer better costs to them (i.e., costs that are greater than the ones they have already). Therefore, one has $f(\mathrm{H}\cdot\langle \mathrm{H,I}\rangle) = \backslash min\{0.66, 0.63\} = 0.63$, $f(\mathrm{H}\cdot\langle \mathrm{H,J}\rangle) = \backslash min\{0.66, 0.65\} = 0.65$, and $f(\mathrm{H}\cdot\langle \mathrm{H,K}\rangle) = \backslash min\{0.66, 0.65\} = 0.65$. Since the costs offered by 'H' are greater or equal than the costs of its neighbors, they are conquered by sample 'H'. Such process is encoded by the aforementioned predecessor map $\mathscr{P}$, i.e., after this first move of sample 'H', one has that $\mathscr{P}(\mathrm{I}) = \mathrm{H}$, $\mathscr{P}(\mathrm{J}) = \mathrm{H}$, and $\mathscr{P}(\mathrm{K}) = \mathrm{H}$.

The next sample to start the competition process is sample 'A', and the very same process

mentioned earlier is repeated until all samples have played in the competition process. The resulting optimum-path forest is depicted in Figure 6.3. Notice one can obtain a different number of clusters based on the value of k_{max}. In this toy example, we obtained two clusters, which are labeled with the same color of its prototype/root of the tree (i.e., the dashed nodes 'A' and 'H').

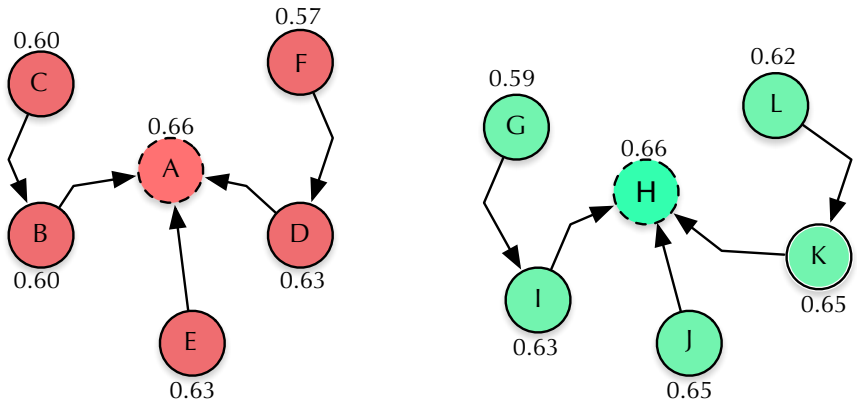


Figure 6.3: Resulting optimum-path forest with two clusters and prototypes highlighted.

The unsupervised OPF algorithm finds the number of the clusters on-the-fly, which means there is no need to have such information beforehand. The only parameter that needs to be set is the k_{max}, which constraints the search for suitable neighborhood sizes. One can observe that the knowledge required to set k_{max} is considerably lower than the one needed to set the number of clusters used by k-means, for instance. Such skill makes OPF pretty much attractive to the application addressed in this paper, as discussed in the next section.

6.3 Barrett's Esophagus

The BE disease is known as the replacement of squamous cells by columnar cells in the esophagus. This process is a result of a complication of gastroesophageal reflux disease, being able to progress into esophageal cancer [Dent 2011, Sharma et al. 2016].

The incidence of BE and esophageal adenocarcinoma in the western population of the world has risen significantly in the past decade. Their close association with the metabolic syndrome suggest growth in the next years [Lagergren e Lagergren 2010, Dent 2011, Lepage, Rachet e Jooste 2008]. The early diagnosis of Esophageal adenocarcinoma in BE diagnosed patients is critical for remission and justifies the necessity of robust surveillance, detection, and characterization. However, the detection of dysplastic tissues and their characterization of abnormalities

within BE-diagnosed patients can be challenging, especially for manual evaluation made by endoscopists. Even considering the dangerousness of the disease, when detected at the early stages, the disease can be treated with very high rates of remission (93% after 10 years) [Dent 2011, Sharma et al. 2016, Phoa et al. 2016].

The esophagus mucosa is composed of squamous cells (similar to the skin or mouth cells), with a whitish-pink color surface, while the gastric mucosa goes sharply from salmon-pink to red [Dent 2011, Sharma et al. 2016]. The point in which the stomach and the stomach meet is called squamocolumnar junction or “Z-line". BE’s mucosa may extend upward in a continuous pattern, changing the Z-line position [Dent 2011, Sharma et al. 2016, Phoa et al. 2016]. Figure 6.4 shows the two cases in which patients can present long-segment of BE and short-segment of BE in a Z-line variation.

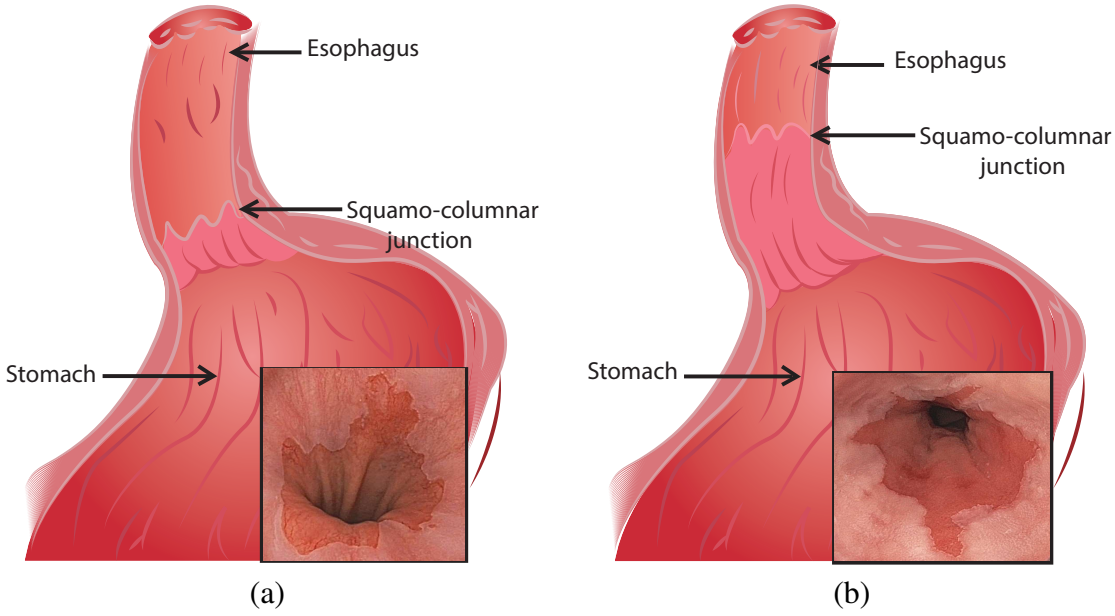


Figure 6.4: BE’s short-segment (a) and BE’s long-segment (b), with their respective endoscopic views (extracted from [Souza Jr. et al. 2018]).

6.4 Methodology and Proposed Approach

In this section, we present the proposed approach and the methodology adopted to cope with the problem of Barrett’s esophagus automatic identification using bag-of-visual-words. First, the proposed method is defined, followed by the datasets used for the experiments, adopted classifiers and experimental delineation.

6.4.1 Proposed Method

As mentioned earlier, one of the leading contributions of this work is to evaluate the robustness of the OPF clustering for learning visual dictionaries. To fulfill that purpose, we considered three distinct feature descriptors based on key point extraction from images: (i) SIFT, (ii) SURF, and (iii) A-KAZE. Although any other approaches could be used, we opted to employ these mainly because they are well known and widely considered in the literature of bag-of-visual-words for both image classification and retrieval, but any other techniques could be applied considering the generalization of the learning visual dictionaries.

For the inicial step of the proposed model, given a set of training images, it is needed to build a bag of key points extracted from them. In hands of a feature extraction technique for the description of an image (being SURF, SIFT or A-KAZE, in this specific case), the model aims to provide the most discriminative key points in the feature dimension based on the entire feature domain. Therefore, taking into account the key points of the entire dataset, a clustering algorithm can be used to group the key points into clusters that share similar properties for choosing the “best key point" from each cluster and use it as the representative of that group. Such samples will compose the final bag-of-visual-words. The main contribution of this work is the calculation of such most representative key points from clusters (as we use to call “prototypes") by using the OPF clustering technique. After obtaining the bag-of-visual-words, the last step of the model is known as “quantization" and computes the new representation for both training and testing images. For each image, it is computed the frequency of each visual word from the bag in the given image by finding the most similar visual word to each key point based on a distance metric. The outcome of that process is a histogram (feature vector) where each bin has the number of key points that are similar to its corresponding visual word. Notice that the representation of both training and testing images are computed based on the same bag. Finally, in hands of the feature vectors, each image of the evaluated dataset shows the exact same number of features for its description, but with the calculation based on the entire feature space domain. The training and testing may be conducted as the final step of the model.

In this work, we propose to cluster the dataset of key points using the OPF technique presented in the previous section and then use the prototypes to compose the bag-of-visual-words. As aforementioned, the prototypes are located in the regions of highest densities, which means they are pretty much suitable to describe the clusters. Another decisive point about OPF concerning other optimization-based clustering techniques relates the fact of not being attracted to local optima, such as k-means or k-medoids, for instance, which are widely used for learning dictionaries due to their simplicity and low computational cost.

As mentioned earlier, OPF finds the clusters on-the-fly, i.e., the clustering process is dynamic, and the forest configuration can change until the last sample finishes the conquering process. Instead of varying the size of the dictionary, one can change the value of k_{max} and then may find the different number of clusters. The cluster calculation comprises one of the most important steps of the proposed method. The prototype computation is performed in an unsupervised process, turning the calculation of the feature vectors based only on the key points themselves, and providing a high generalization for this task. The problem of a different number of key points for each image can be solved using this bag-of-visual-words approach, proposing a consistent way of regular description for images evaluated by feature extraction techniques. Figure 6.5 depicts the pipeline adopted in this work.

Since the images are colored, a gray-scale normalization is applied to the images so that the key points can be extracted. Later, such PoIs are then mapped onto a feature space for clustering purposes. An example of the outcome of the clustering process is depicted at the bottom of Figure 6.5. Each color stands for a different group and the dashed nodes represent the prototypes selected by OPF to be part of the visual dictionary. As aforementioned, a histogram is built upon the training PoIs and the visual words for the further design of the final set of handcrafted features. For the selected visual words, an evaluation of their appearance is performed in the PoIs of each dataset image aiming to calculate the final cumulative histogram that represents each feature vector, with dimension depending on the number of visual words generated in the clustering calculation of the bag.

6.4.2 Datasets

An in-depth analysis concerning the robustness of the proposed approach is provided through two datasets. The first dataset comprises a set of images from a benchmark dataset provided at the “MICCAI 2015 EndoVis Challenge"[1] was considered, hereinafter called “MICCAI 2015" dataset, which aimed at differentiating Barrett’s esophagus from cancerous images. Such dataset is composed of 100 endoscopic pictures of the lower esophagus captured from 39 individuals, 22 of them being diagnosed with early-stage Barrett’s, and 17 showing signs of esophageal adenocarcinoma. Each patient has several endoscopic images available, ranging from one to a maximum of eight. The database comprises a total of 50 images displaying cancerous tissue areas as well as 50 images showing dysplasia without signs of cancer. Suspicious lesions observed in the cancerous images had been delineated individually by five endoscopy experts.

[1] https://endovissub-barrett.grand-challenge.org/home/

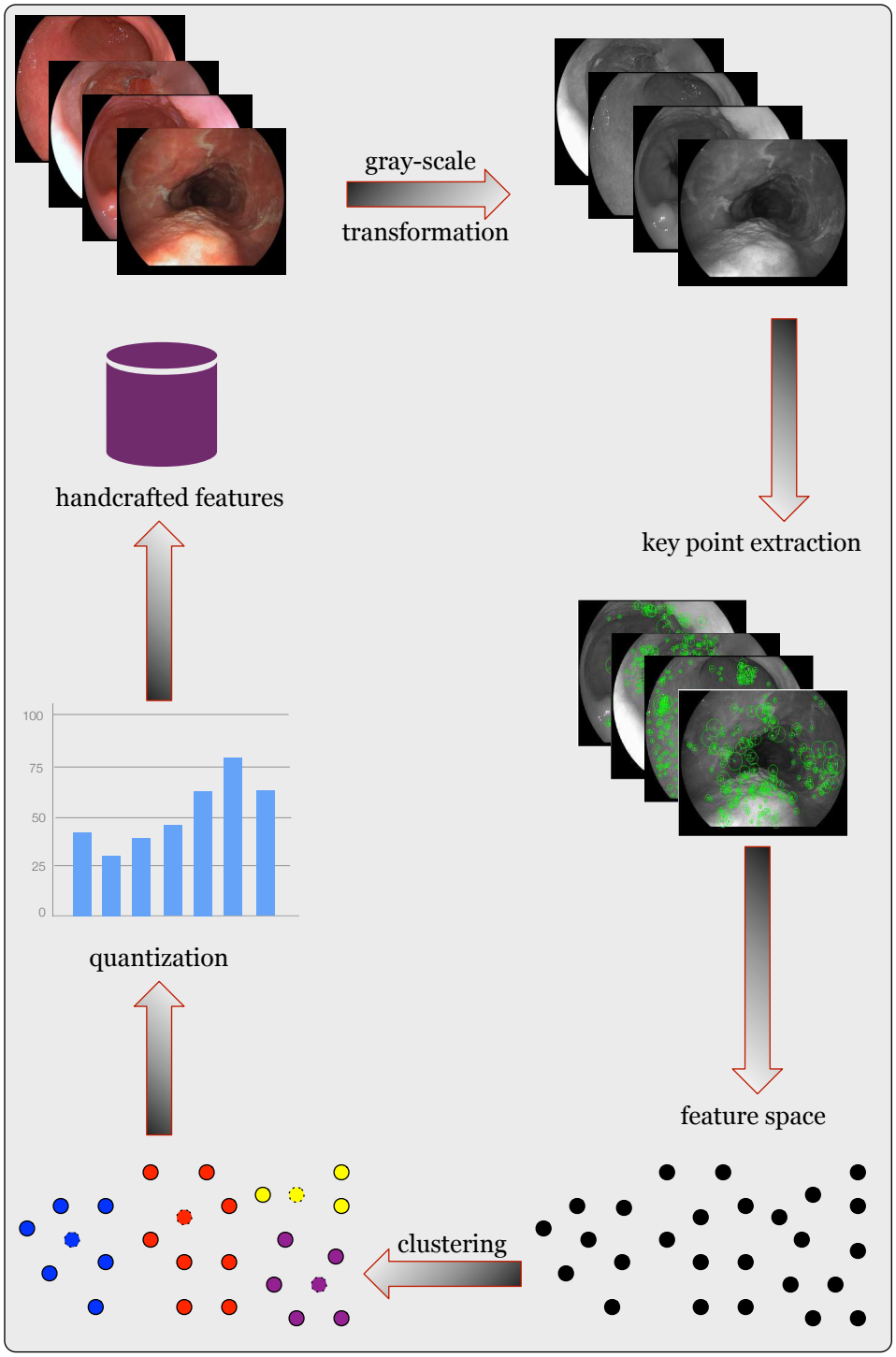


Figure 6.5: Pipeline adopted in this work for Barrett's esophagus identification.

Additionally, a dataset provided by the Augsburg Klinikum, Medizinische Klinik III was also used for the experiments. Such dataset is composed of 76 endoscopic images (esophagus) captured from different patients with adenocarcinoma (34 samples) and BE (42 samples). The images were annotated (manual segmentation of the adenocarcinoma's and Barrett's area, respectively) by an expert from the Augsburg Klinikum, and the diagnosis was provided using

biopsy. Since we are dealing with a classification problem, the annotations provided by the experts were not considered in our work.

Figure 6.6 depicts some examples of the MICCAI 2015 dataset positive for cancer (i.e., negative for BE) and their respective delineations performed by five experts. However, we are not working with the delineation information since we compute the PoIs for the whole image. One could use the information about the delineated regions to extract PoIs from that areas only, which could guarantee that pure adenocarcinoma PoIs are computed, but the problem still concerns the fact that delineations are not available to all real-world images.

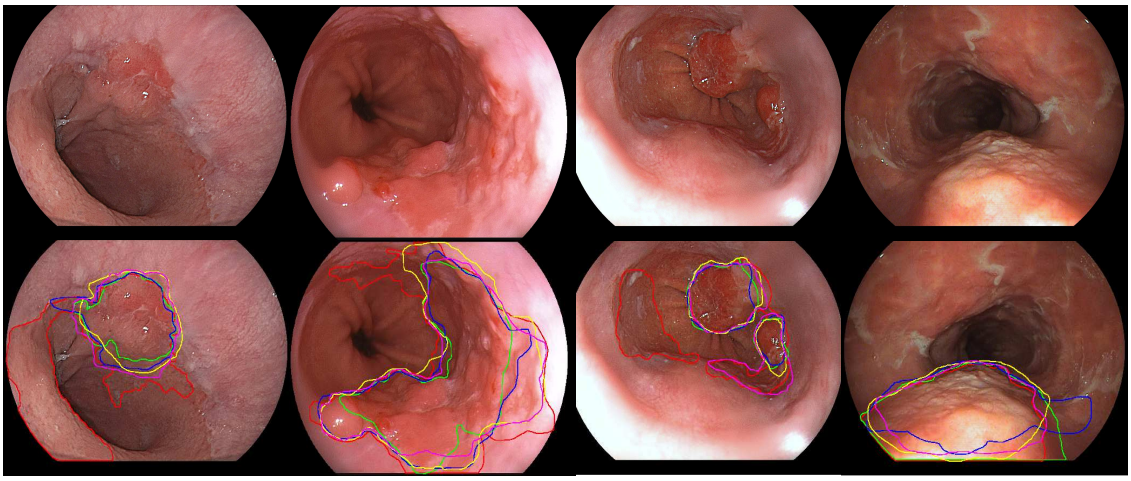


Figure 6.6: Some examples of images positive for cancer and their respective delineations (MICCAI 2015 dataset).

Figure 6.7 displays some images positive for cancer from Augsburg dataset. In this case, we have only one delineation per image. Once again, such information is not used in this work since we are interested mostly in the differentiation of Barrett's esophagus and adenocarcinoma rather than its segmentation.

6.4.3 Adopted Classifiers

We considered different supervised pattern recognition techniques to assess the robustness of unsupervised OPF for learning visual dictionaries:

- OPF$_{cpl}$: supervised OPF with complete graph proposed by Papa et al. [Papa, Falcão e Suzuki 2009, Papa et al. 2012];
- OPF$_{knn}$: supervised OPF with k-nn graph proposed by Papa et al. [Papa, Fernandes e Falcão 2017];

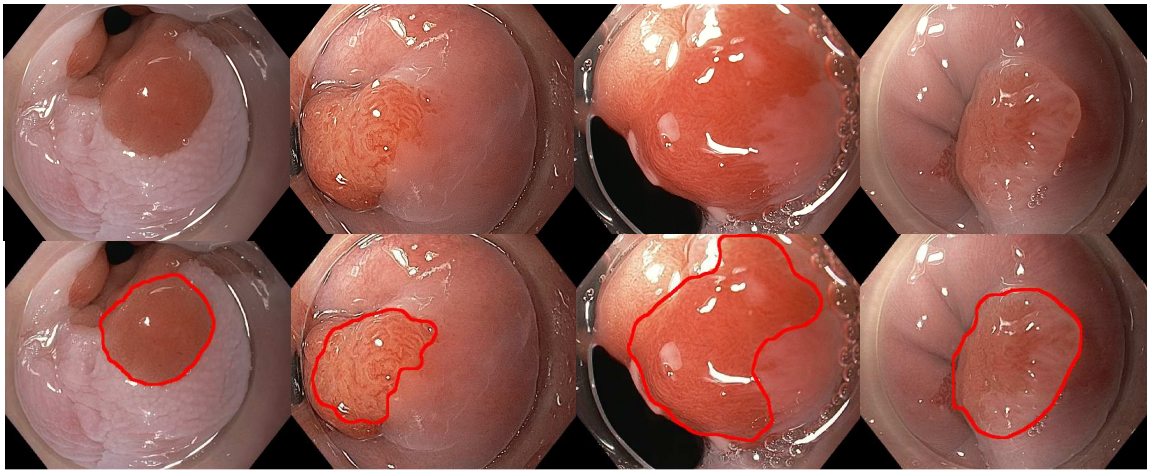


Figure 6.7: Some examples of images positive for cancer and their respective delineations (Augsburg dataset).

- SVM-RBF: Support Vector Machines with Radial Basis Function kernel and parameters optimized by cross-validation [Chang e Lin 2011];
- SVM-Linear: Support Vector Machines with Linear kernel and parameters optimized by cross-validation [Chang e Lin 2011];
- Bayes: standard Bayesian classifier.

Regarding the OPF-based classifiers, we used the LibOPF [Papa e Falcão], which is an open-source library that implements both the supervised as well as the unsupervised versions of the OPF used in this work. With respect to the Bayesian classifier, we employed our own implementation.

6.4.4 Experimental Delineation

To compose the set of experiments, we considered three different sizes for the dictionaries: 100, 500, and $1,000$. The main idea is to evaluate the robustness of the techniques used in this work under different scenarios. As we shall discuss later, the usage of dictionaries with 500 visual words seemed to achieve better results, as stated in a previous work [Souza Jr. et al. 2017], which motivated us to set $k_{max} = 500$ for this one. However, this not implies in constraining OPF to find exactly 500 clusters, just to limit the size of the neighborhood of each sample to be 500. Regarding OPF_{knn}, its parameter k is fine-tuned within the range $[1,500]$, and the value that maximized the accuracy over the training set was used.

Regarding the experimental validation, it was considered a cross-validation approach with 20 runs and using 70% of the dataset for training purposes, as well as the remaining 30%

for classification. Moreover, the experimental results were assessed using a statistical analysis using the Wilcoxon signed-rank test with confidence as of 5% [Wilcoxon 1945]. All experiments were conducted on an 8GB-memory computer equipped with an Intel Core i5 - 2.30 GHz processor. Additionally, we employed the OpenCV [OpenCV 2015] implementation for feature extraction using SIFT, SURF, and A-KAZE.

6.5 Experimental Results

In this section, we present the experiments used to evaluate the proposed approach. Five supervised classifiers were considered to discriminate between samples positive and negative to adenocarcinoma: OPF_{cpl}, OPF_{knn}, SVM-RBF, SVM-Linear, and Bayesian classifier (hereinafter called Bayes). For all the classifiers adopted for such evaluation, there was no need for setting any parameter, as long as they were used in the default set. The same experimental protocol was applied to all techniques using cross-validation, i.e., three distinct feature representations were considered (SURF, SIFT, and A-KAZE, with the metric threshold of all sets is default), and with different bag sizes (i.e., 100, 500 and $1,000$ visual words). The results are presented and discussed considering each dataset individually.

A statistical evaluation using the signed-rank Wilcoxon test [Wilcoxon 1945] was used for comparison purposes as follows:

1. For each dictionary generation approach (i.e., clustering by k-means, random or unsupervised OPF), it was verified the classification results and the best ones were highlighted in bold. Statistically similar results were highlighted in bold.

2. For each feature extractor (i.e., A-KAZE, SIFT, and SURF), the best statistical results were underlined.

3. Additionally, the best results among all configurations were marked with a '$\star$' symbol.

This very same procedure was adopted to both datasets.

In this work, we used the following accuracy rate:

$$A = \frac{TP+TN}{TP+TN+FP+FN} \cdot 100, \tag{6.5}$$

where TP and TN stand for the true positives and true negatives, respectively, and FN and FP denote the false negatives and false positives, respectively. In a nutshell, the above equation

computes the ratio between the number of correct classifications (i.e., $TP+TN$) and the size of the dataset (i.e., all correct and wrong classifications).

6.5.1 MICCAI 2015 Dataset

Tables 6.1, 6.2, and 6.3 present the results related to A-KAZE, SURF, and SIFT descriptors, respectively, concerning MICCAI 2015 dataset. Regarding the A-KAZE results presented in Table 6.1, one can draw the following conclusions: (i) OPF_{cpl} obtained the best results for all dictionary generation techniques, and (ii) OPF clustering achieved the best results (77.6% of recognition rate with $1,000$ visual words) for BE recognition among all configurations, although being statistically similar to k-means with OPF_{cpl} with 500 and $1,000$ visual words as well.

Table 6.1: Mean accuracy results using A-KAZE features with 100**,** 500**, and** $1,000$ **visual words.**

Dictionary		100	500	1000
k-means	OPF_{cpl}	73.6%	***76.3%**	***77.2%**
	OPF_{knn}	59.2%	61.7%	68.1%
	SVM-RBF	60.7%	65.9%	66.1%
	SVM-Linear	58.5%	63.0%	67.4%
	Bayes	56.8%	60.0%	60.9%
Random	OPF_{cpl}	59.5%	63.9%	**70.3%**
	OPF_{knn}	58.3%	61.7%	62.3%
	SVM-RBF	62.1%	65.6%	63.7%
	SVM-Linear	55.3%	59.0%	59.1%
	Bayes	55.5%	62.2%	61.1%
OPF clustering	OPF_{cpl}	72.2%	73.1%	***77.6%**
	OPF_{knn}	62.3%	60.1%	66.2%
	SVM-RBF	61.9%	65.1%	70.9%
	SVM-Linear	55.8%	60.5%	66.8%
	Bayes	55.8%	58.0%	61.3%

The average number of PoIs used for training and test sets concerning A-KAZE feature extractor were $16,024$ and $6,868$, respectively, taking an average computational load of 4.05 minutes. A training set composed of around $16,000$ visual words is enough to support the sizes of the dictionaries we used to build the feature vector of each image, i.e., 100, 500, $1,000$. Larger dictionaries may not be interesting since there will be numerous small-sized clusters, which means less spatial information about the visual words is captured.

Table 6.2 presents the results concerning the SURF feature extractor. Once again, OPF_{cpl} achieved the best classification results regarding all dictionary generation approaches, and OPF clustering allowed the best results among all, i.e., it could learn better dictionaries for image

representation. In this context, a dictionary of size 500 computed by k-means also achieved the best recognition rates according to the statistical test. The average number of PoIs used for training and test sets concerning SURF feature extractor were $14,411$ and $6,189$, respectively, taking an average computational load of 13.77 minutes.

Table 6.2: Mean accuracy results using SURF Features and 100, 500, and $1,000$ visual words.

Dictionary		100	500	1000
k-means	OPF_{cpl}	70.0%	**74.8%**	**73.6%**
	OPF_{knn}	64.1%	66.0%	65.1%
	SVM-RBF	63.6%	64.8%	62.6%
	SVM-Linear	62.0%	58.6%	62.8%
	Bayes	56.4%	56.9%	57.4%
Random	OPF_{cpl}	**69.7%**	**70.2%**	**66.1%**
	OPF_{knn}	58.0%	58.4%	61.8%
	SVM-RBF	61.0%	63.4%	62.1%
	SVM-Linear	51.7%	57.6%	56.5%
	Bayes	50.5%	53.5%	56.9%
OPF clustering	OPF_{cpl}	69.4%	***78.4%**	***77.1%**
	OPF_{knn}	63.6%	69.6%	71.6%
	SVM-RBF	67.5%	71.8%	70.9%
	SVM-Linear	65.1%	66.9%	66.7%
	Bayes	53.3%	56.8%	57.1%

One can observe that SVM did not obtain proper recognition rates in both situations, i.e., A-KAZE and SURF feature extractors. One possible reason concerns the number of training samples, which is usually lower than the number of features. Therefore, SVM will map samples to a lower-dimensionality feature space instead of a higher one, thus neglecting the assumption of linearity in higher-dimensionality spaces.

Table 6.3 presents the results considering the SIFT feature extractor. Once again, OPF_{cpl} achieved the best results so far, with OPF_{knn} and SVM-RBF being statistically similar for k-means with $1,000$ words and a random generation of dictionaries with $1,000$ words. However, the best global results were achieved using OPF clustering with OPF_{cpl} with 500 and $1,000$ visual words, outperforming by far the other results with SIFT feature extractor. The average number of PoIs used for training and test sets concerning SIFT feature extractor were $28,137$ and $12,059$, respectively, taking an average computational load of 5.95 minutes.

Last but not least, the best results among all three feature extractors (i.e., the ones marked with '$\star$') were obtained using OPF clustering for dictionary generation and OPF_{cpl} for classification with 500 and $1,000$ visual words considering SURF and SIFT, and the same pair (i.e.,

Table 6.3: Mean accuracy results using SIFT Features and 100, 500, and $1,000$ visual words.

Dictionary		100	500	1000
k-means	OPF$_{cpl}$	68.3%	**72.3%**	**71.4%**
	OPF$_{knn}$	67.0%	**71.8%**	**72.1%**
	SVM-RBF	67.3%	**71.4%**	**71.9%**
	SVM-Linear	55.2%	56.8%	67.3%
	Bayes	53.5%	60.0%	60.7%
Random	OPF$_{cpl}$	**66.4%**	**70.7%**	**71.2%**
	OPF$_{knn}$	58.1%	63.9%	**66.1%**
	SVM-RBF	62.1%	65.6%	63.7%
	SVM-Linear	53.2%	54.5%	52.7%
	Bayes	50.2%	53.0%	54.4%
OPF clustering	OPF$_{cpl}$	71.2%	***77.7%**	***78.9%**
	OPF$_{knn}$	63.9%	71.3%	**75.7%**
	SVM-RBF	68.0%	70.2%	69.7%
	SVM-Linear	61.3%	64.7%	64.4%
	Bayes	50.2%	53.0%	54.4%

OPF clustering and OPF$_{cpl}$) regarding A-KAZE with $1,000$ visual words, and finally k-means and OPF$_{cpl}$ with 500 words. Notice the best absolute result was obtained using OPF clustering for visual words generation and OPF$_{cpl}$ for classification purposes with SIFT-based features (i.e., 78.9%).

6.5.2 Augsburg Dataset

Tables 6.4, 6.5, and 6.6 present the results related to A-KAZE, SURF, and SIFT descriptors, respectively, concerning Augsburg dataset. Starting with the A-KAZE feature extractor, one can observe the best results were mostly obtained by both OPF$_{cpl}$ and OPF$_{knn}$. The best global results were achieved by OPF clustering, Random and k-means, but the most accurate one (i.e., absolute results) was OPF clustering for visual dictionary generation and OPF$_{cpl}$ for classification purposes with accuracy of 72.6%. Such result is slightly less accurate than the same feature extractor considering MICCAI 2015 dataset since Augsburg dataset is more challenging due to different levels of adenocarcinoma. The average number of PoIs used for training and test sets concerning A-KAZE feature extractor were $40,064$ and $17,170$, respectively, taking an average computational load of 4.18 minutes.

Table 6.5 presents the results concerning the SURF feature extractor. Once again, OPF-based classifiers obtained the best results in most of the scenarios, being OPF clustering and k-means the best approaches for visual dictionary generation. The best absolute classification

Table 6.4: Mean accuracy results using A-KAZE Features and 100, 500, and 1,000 visual words.

Dictionary		100	500	1000
k-means	OPF_{cpl}	60.7%	⋆**69.4%**	65.6%
	OPF_{knn}	61.9%	**66.1%**	⋆**70.1%**
	SVM-RBF	60.4%	63.5%	63.1%
	SVM-Linear	55.1%	60.4%	62.1%
	Bayes	56.9%	60.1%	61.3%
Random	OPF_{cpl}	59.4%	**68.4%**	⋆**69.9%**
	OPF_{knn}	59.9%	61.4%	62.4%
	SVM-RBF	57.9%	62.2%	63.2%
	SVM-Linear	55.3%	58.8%	58.9%
	Bayes	56.5%	57.1%	61.0%
OPF clustering	OPF_{cpl}	68.4%	⋆**68.7%**	⋆**72.6%**
	OPF_{knn}	67.4%	⋆**69.3%**	⋆**70.3%**
	SVM-RBF	59.4%	63.0%	⋆**69.8%**
	SVM-Linear	57.7%	57.3%	62.7%
	Bayes	62.4%	60.7%	63.1%

results were obtained by OPF_{cpl} and Bayes with accuracies nearly to 68%. The average number of PoIs used for training and test sets concerning SURF feature extractor were 14,251 and 6,108, respectively, taking an average computational load of 9.23 minutes.

Table 6.5: Mean accuracy results using SURF Features and 100, 500, and 1,000 words.

Dictionary		100	500	1000
k-means	OPF_{cpl}	**66.3%**	⋆**67.9%**	61.5%
	OPF_{knn}	62.8%	63.2%	**65.4%**
	SVM-RBF	57.1%	61.1%	62.9%
	SVM-Linear	56.7%	57.1%	59.4%
	Bayes	60.8%	59.9%	61.1%
Random	OPF_{cpl}	60.0%	**62.2%**	**63.5%**
	OPF_{knn}	54.2%	58.1%	60.8%
	SVM-RBF	61.3%	**61.9%**	**62.0%**
	SVM-Linear	57.1%	55.4%	56.4%
	Bayes	51.9%	59.0%	59.1%
OPF clustering	OPF_{cpl}	59.2%	62.1%	**66.1%**
	OPF_{knn}	61.1%	63.9%	64.5%
	SVM-RBF	58.5%	62.0%	**65.4%**
	SVM-Linear	53.5%	60.8%	64.6%
	Bayes	59.8%	**67.0%**	⋆**67.9%**

The Augsburg dataset figured out as being more challenging than MICCAI 2015 dataset

due to the considerably low results achieved (Table 6.2). SVM-RBF presented better results with higher-dimensionality bags (i.e., 65.4% with 1,000 words with OPF clustering), and the same behavior can be observed regarding SVM-Linear.

Table 6.6 presents the results considering the SIFT feature extractor. In this case, OPF-based classifiers and SVM-RBF figured as the most accurate techniques and OPF clustering as the best one for visual dictionary generation (absolute results). A comparison against A-KAZE and SURF showed these to be quite more accurate than SIFT, an opposite situation that occurred over MICCAI 2015 dataset, where SIFT achieved the best recognition rates. Additionally, the average number of PoIs used for training and test sets concerning SIFT feature extractor were 89,514 and 38,363, respectively, taking an average computational load of 8.71 minutes.

Table 6.6: Mean accuracy results using SIFT Features and 100, 500, and 1,000 visual words.

Dictionary		100	500	1000
k-means	OPF_{cpl}	**60.3%**	**60.5%**	59.3%
	OPF_{knn}	58.9%	**60.6%**	**62.0%**
	SVM-RBF	60.8%	61.8%	**59.8%**
	SVM-Linear	55.5%	57.1%	**59.9%**
	Bayes	53.1%	54.8%	58.7%
Random	OPF_{cpl}	59.2%	60.5%	**61.6%**
	OPF_{knn}	57.0%	58.4%	60.5%
	SVM-RBF	57.8%	**62.6%**	**62.1%**
	SVM-Linear	54.4%	55.6%	**61.5%**
	Bayes	51.9%	57.0%	59.0%
OPF clustering	OPF_{cpl}	60.4%	**62.8%**	**62.1%**
	OPF_{knn}	58.1%	61.6%	**63.9%**
	SVM-RBF	57.0%	60.5%	**62.1%**
	SVM-Linear	58.8%	58.9%	58.7%
	Bayes	61.1%	**62.2%**	61.8%

6.5.3 Discussion

In this section, we aim at providing a more in-depth discussion about the experiments, as well as insightful conclusions regarding the usage of bag-of-visual words in the context of computer-aided differentiation between Barrett's esophagus and adenocarcinoma. Table 6.7 presents a summary with the best results obtained in the previous two sections concerning the number of visual words and feature extractor. Concerning both datasets, OPF_{cpl} figured as the more accurate classification technique, meanwhile OPF clustering appears as the best dictionary generation approach.

Table 6.7: Summarization of the results.

Dataset	Accuracy	Feature Extractor	#visual words
MICCAI 2015	78.9%	SIFT	1,000
Augsburg	72.6%	A-KAZE	1,000

The results support the primary contributions stated previously, which are related to the robustness of OPF-based classifiers for both supervised and unsupervised learning in the context of automatic adenocarcinoma identification. Additionally, the number of visual words strongly affects the results, but we believe a trade-off between the size of the dictionary and the information it carries on shall be established beforehand.

Table 6.8 summarizes the mean sensitivity and specificity results of both datasets with the best configuration of the number of visual words, dictionary generation approach, feature extractor, and classification technique mentioned above. Sensitivity stands for the classification rate considering adenocarcinoma identification, i.e., those positive to Barrett's esophagus and to adenocarcinoma, and specificity denotes the accuracy regarding those negative to adenocarcinoma, i.e., positive only to BE. Considering such sensitivity and specificity results, some conclusions can be drawn: (i) for the MICCAI 2015 dataset, the sensitivity results presented higher values than the specificity ones, suggesting a very good generalization in the positive adenocarcinoma identification. Even with lower results, the specificity still showed a convincing value, and the misclassification can be justified by two factors: the fuzzy region (region in which the experts disagree in the annotation) and lack of enough key points in the non-cancerous regions during the feature vector calculation. For the Augsburg results of sensitivity and specificity, a better trade-off between the correct classification of positive and non-positive adenocarcinoma samples could be found, but still with lower results when compared to the MICCAI 2015 dataset ones. The Augsburg dataset presents images with different behavior and acquisition technology when compared to the MICCAI 2015 ones, thus justifying the worse results.

Table 6.8: Mean sensitivity (i.e., positive to BE) and specificity (i.e., negative to BE) results.

Dataset	Sensitivity	Specificity
MICCAI 2015	81.7%	76.4%
Augsburg	70.9%	74.9%

To provide more insightful comments and to better understand the working mechanism of visual words in the context of computer-assisted BE identification, we performed some addi-

tional experiments with cancerous images that were classified either as cancer or as Barrett's esophagus since we have their delineated regions. In a nutshell, the main idea is to compute the percentage of PoIs located inside those regions with respect to the remaining ones (i..e, those located outside cancerous areas). This information allows us to compare whether the number of PoIs placed inside the delineated regions are enough or not to provide accurate classifications.

Table 6.9 presents the mean percentage of PoIs located inside the cancerous area for the whole dataset, as well as the average percentage of PoI inside the cancerous area concerning the misclassified images (i.e., cancerous images that were classified as BE). Since we conducted a cross-validation approach with 20 runs, the average percentages concerning the misclassified images (i.e., Cancer→BE) were computed to each run, for the further computation of the average value of all. Additionally, since MICCAI 2015 dataset comprises delineations from five experts, we took the intersection of them all as the final delineated area to compute the percentage of PoIs into account.

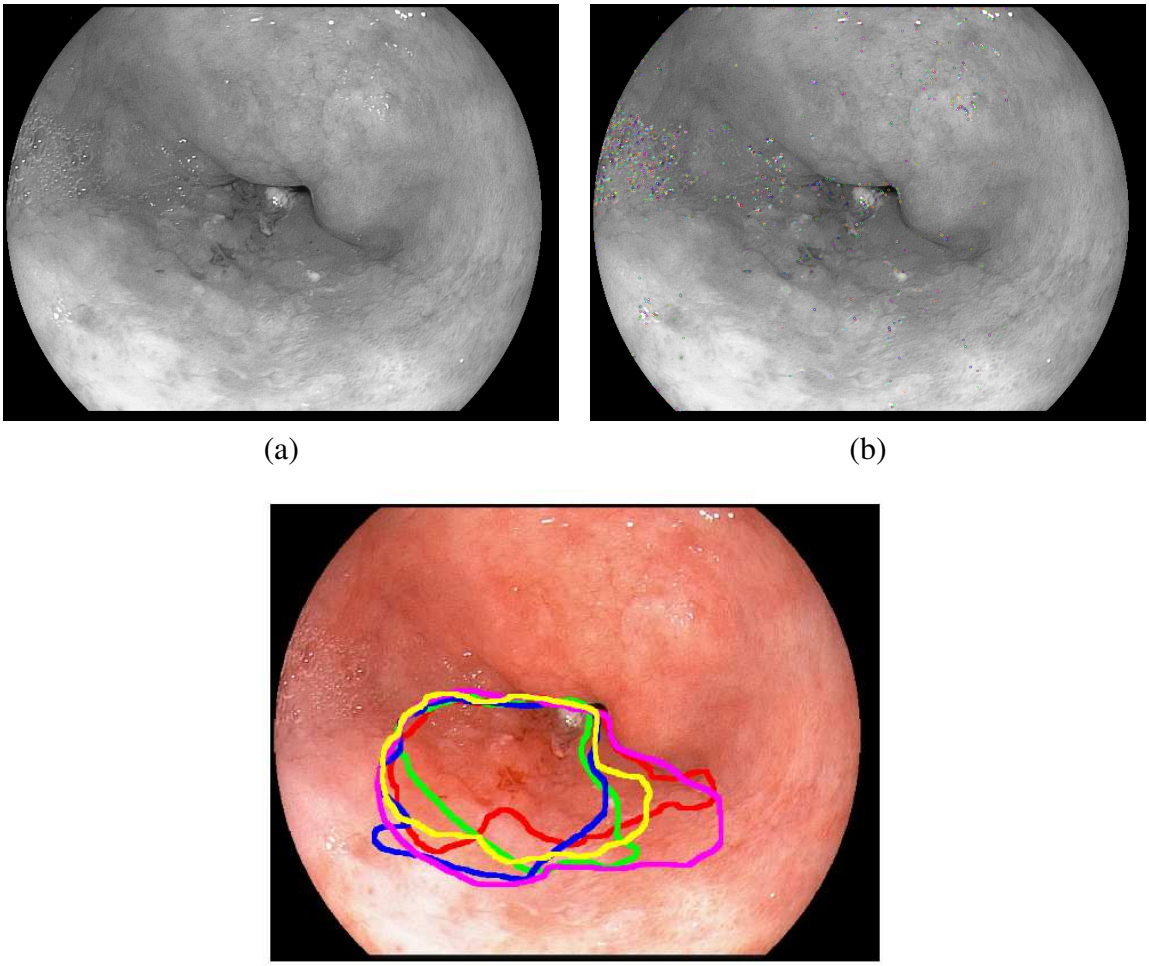
One can observe that the percentage of PoIs inside the cancerous images were more significant than the values obtained from the misclassified images. Such assumption is pretty interesting since we can conclude that the number of PoIs inside the delineated regions are essential to achieve accurate results and to avoid misclassifications. The only exception stands for the Augsburg dataset with A-KAZE features, where the number of PoIs were slightly higher for the misclassified images. Note that the percentage of PoIs inside the cancer region is in general higher for the Augsburg databaset than for the MICCAI 2015 dataset. This can be explained because the Augsburg images use the near-focal imaging technique, in which the suspicious region is displayed larger.

Table 6.9: Percentage of PoIs inside the delineated (cancerous) ares.

Dataset	Feature Extractor	Cancer PoI %	Cancer→BE PoI %
MICCAI 2015	A-KAZE	30.34%	21.69%
MICCAI 2015	SURF	25.58%	23.05%
MICCAI 2015	SIFT	30.73%	23.04%
Augsburg	A-KAZE	53.77%	55.70%
Augsburg	SURF	42.97%	39.54%
Augsburg	SIFT	48.34%	44.06%

For visualization purposes, Figures 6.8 to 6.9 depict some cancer patients that were misclassified as BE from both datasets. The PoIs showed in Figure 6.8 were calculated using SIFT and belong to the MICCAI 2015 dataset, and their percentage of incidence is 21.72%, which is

slightly lower considering the average percentage presented in Table 6.9 (23,04%).



(a) (b)

(c)

Figure 6.8: Misclassified image (patient 31) from MICCAI 2015 dataset: (a) gray-scale, (b) PoIs (SIFT), and (c) RGB version with delineations.

One can observe a considerable amount of PoIs located at the left-middle portion of Figure 6.8b, mainly due to some air bubbles and foam. Problems with light (upper part of the image) also contribute to placing PoIs outside the delineated area.

The PoIs showed in Figure 6.9 were calculated using A-KAZE on an image from the Augsburg dataset. Their percentage of incidence is 7.5%, which is quite low considering the average percentage presented in Table 6.9 (53.77%). In this case, the main reason for placing PoIs outside the delineated area concerns illumination problems (brighter areas).

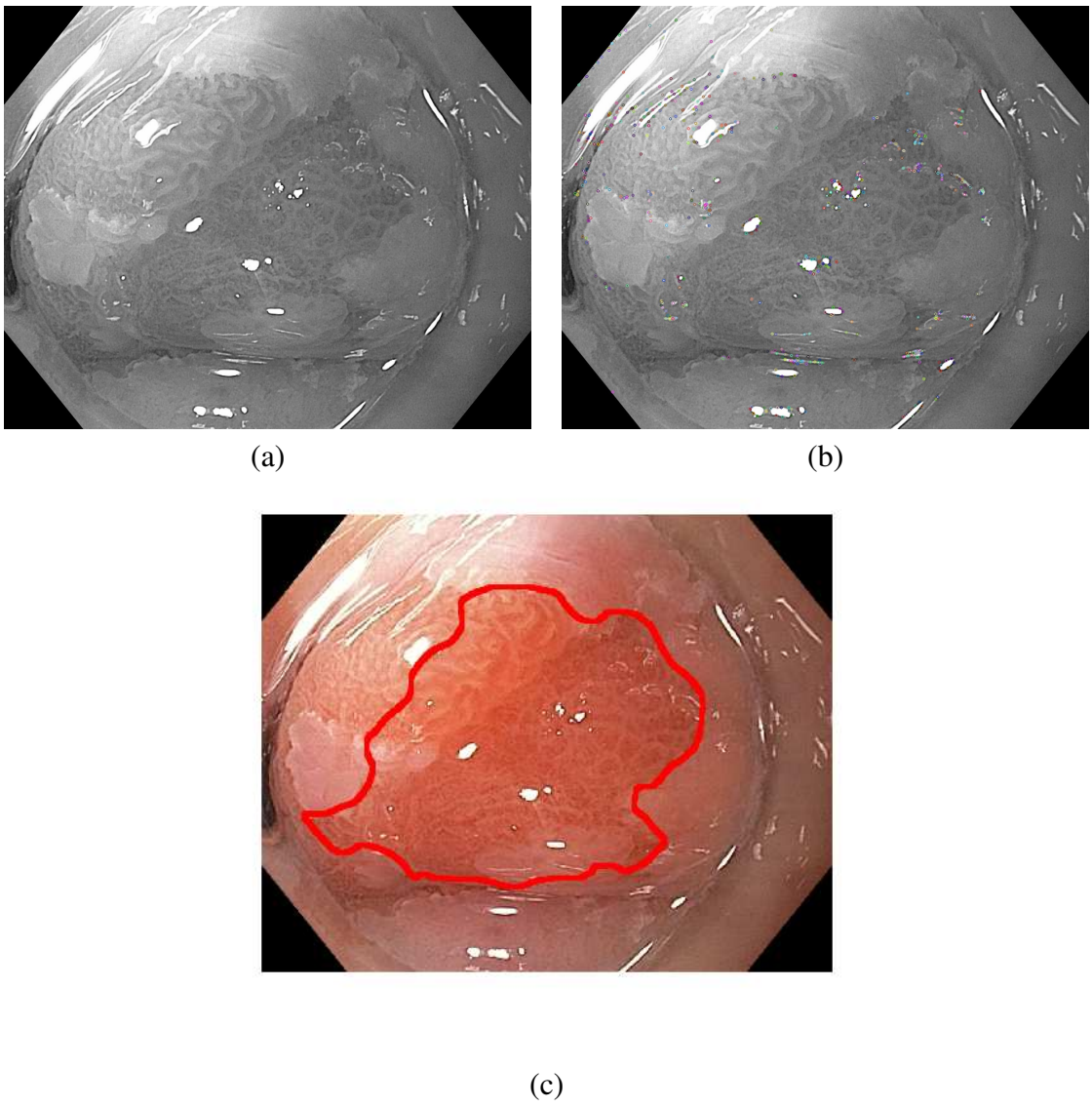


Figure 6.9: Misclassified image (patient 39) from Augsburg dataset: (a) gray-scale, (b) PoIs (A-KAZE), and (c) RGB version with delineation.

6.6 Conclusions and Future Works

In this paper, we dealt with the problem of computer-assisted Barrett's esophagus identification by means of bag-of-visual-words calculated using the OPF clustering technique. Such technique showed promising results, outperforming the previous handcrafted feature results in the same context. This suggests the generalization relevance of such technique, which can improve previous results in the same field not only for the BE context but for other in which the image representation configures the context to be evaluated. BE stands for an illness that is likely to be confused with adenocarcinoma, and its early detection and prevention is of great concern.

We observed that only a very few works attempted at coping with the problem of automatic BE identification using computer vision and machine learning techniques to date. In this work, we fostered the research towards such area by introducing a supervised variant of the Optimum-Path Forest classifier for automatic BE recognition, as well as we showed how to build proper visual dictionaries using unsupervised OPF learning, outperforming the results obtained in some recent works in which the same database and protocol were applied [Souza Jr. et al. 2017, Souza Jr. et al. 2017]. Considering some previous works [Souza Jr. et al. 2017, Souza Jr. et al. 2018, Souza Jr. et al. 2017], the use of handcrafted features were based on the SURF and SIFT PoIs, but without the use of the OPF clustering as a way of dimensional reduction of the problem. Moreover, considering the improvements of the results, the use of the OPF clustering provides a new and promising way of BE and adenocarcinoma problem evaluation based on extracted key points. The presented results showed the relevance of such technique addressed to the BE and adenocarcinoma evaluation and description, contributing to the context literature and influencing the evaluation and description of other tissue diseases. Comparing the proposed method with others already published, we can ensure that with the use of the OPF for the BoVW step, improvements could be achieved considering the higher results obtained. Also, such technique provides advantages in the dimension reduction of the feature vector calculation, once even with a different number of key points per image, a standard method of feature calculation is established. Again, the OPF clustering may provide flexibility and time saving for such task.

The experimental results were considered over two datasets: (i) MICCAI 2015, and (ii) Augsburg. For both scenarios, we evaluated five classification techniques and three unsupervised learning approaches to build the visual dictionaries. Also, we considered dictionaries with three distinct sizes and even three different feature extractors.

The experiments pointed out that bag-of-visual-words techniques are suitable to handle BE automatic identification, and there must be a trade-off between the number of visual words and the amount of information they can encode (i.e., size of the clusters). Additionally, both supervised and unsupervised OPF-based classifiers achieved the most accurate results, thus supporting the main contributions of this paper.

In the following, a bullet list of trends based on the achieved results is presented:

- the OPF classifier presented the highest results of accuracy in all experiments, and may be highly recommended considering the high generalization that provided for such a context, even for different description scenarios;

- the representation of BE and adenocarcinoma by means of image description techniques may provide encouraging results, and with less computation processing cost as needed in more sophisticated techniques;
- the A-KAZE features showed the very best results for BE and adenocarcinoma description in the Augsburg dataset evaluation, suggesting to be a very important technique for the description of such diseases;
- the use of handcrafted features still has potential to evaluated for BE and adenocarcinoma problem, considering the several number of techniques, such as fisher vectors and sparse coding;
- the use of OPF clustering improved the current results and can be applied to a large number of cases of image description of the BE and adenocarcinoma regions;
- the way of improving the selection of key points in each region (cancerous and non-cancerous) still shows potential considering the influence of the number of key points in each region for the correct classification result.

Regarding future works, we aim at considering deep learning and post-processing techniques after the construction of the bags, such as feature selection (i.e., visual word selection). Additionally, this post-processing can be performed using the large number of machine learning techniques, such as SVM, OPF or even Convolution Neural Networks, providing intermediate learning for the dictionaries calculation. More techniques for image description are also considered to be evaluated using the bag-of-visual-words provided by the OPF clustering.

6.7 Chapter's Considerations

In a direct continuation of the work proposed in Chapters 3, 4, and 5, a new object detector technique was introduced in this Chapter, along with a brand new evaluation of feature location after calculating the key-points.

Here, AKAZE features, an object detection and description method that operates in a non-linear scale space are employed. Not like the previous methods such as SIFT or SURF, that find features in the Gaussian scale space, not respecting the natural boundaries of objects and may smooth details and noise from the original image, AKAZE, using a nonlinear diffusion, detects and describes features in nonlinear scale-spaces, keeping up important details and removing noise through the scale-spaces. Finally, AKAZE employs a mathematical process cal-

led Fast Explicit Diffusion to increase the nonlinear space scale computation. As a result, a high-dimensional description of size 61 is obtained for each key-point the technique detects.

The evaluation protocols were again based on full-images, and the definition of features based on object detection techniques (SURF, SIFT, and AKAZE) and BoVW representation. This time, the introduction of unsupervised-OPF for calculating the most representative features, or prototypes, for further feature vector computation was a strong contribution of this study. The unsupervised-OPF build the optimum trees concerning a maximum number of neighbors, defining the samples that best represent each one, further called the prototypes to be compared of all the features in the computation of the feature vector of each input sample. As in the previous Chapters, this one conducted experiments over a 20-fold cross-validation design, with 80% of samples randomly selected for training and the remaining 20% for testing. Also, BoVW with bigger sizes was proposed to evaluate the impact of the number of features in the correct detection of cancerous tissues.

Remarkable results could be achieved after the experimental step. First, the best results could be achieved with the biggest visual word size, suggesting that the representation of more features gives the classifier more information to describe the tissues we aim to differentiate. SIFT and AKAZE techniques, in combination with unsupervised-OPF BoVW, presented the best classification results for MICCAI and Augsburg datasets, respectively. Such a result highlights the powerful space-scale representation SIFT and AKAZE perform, extremely representative in properly describing early cancerous and BE tissues. Finally, the spatial position of the defined features was clearly evaluated, and the highest accuracies were always related to a higher amount of features automatically detected inside the experts' annotation area. Due to that, this manuscript suggests a relation between experts' and computational description for the correct and positive definition of adenocarcinoma in BE samples.

Next, some extensions of this work would be evaluated, based on brand new ways of optimizing the feature selection and positioning during the description of cancerous and noncancerous tissues. Also, more classifiers should be evaluated based on the optimization we want to conduct. Finally, after extensively evaluating handcrafted feature representations, deep learning techniques shall be employed in the identification of early cancer in BE context.

Chapter 7

BARRETT'S ESOPHAGUS ANALYSIS USING INFINITY RESTRICTED BOLTZMANN MACHINES

This chapter presents the paper entitled Barrett's Esophagus Analysis Using Infinity Restricted Boltzmann Machines, published at Journal of Visual Communication and Image Representation [Passos et al. 2019] as an extension from the idea presented in [Passos e Papa 2017] applied to medical issues.

7.1 Introduction

The incidence of adenocarcinoma in patients with (BE) faced a major increase in western populations in the last 10 years, explained by risk factors such as obesity and smoking [Lagergren e Lagergren 2010, Dent 2011, Lepage, Rachet e Jooste 2008], and an expectation to rise in the next years. The bad prognosis of patients suffering from esophageal adenocarcinoma is related to its late diagnosis. Despite the dangerousness of the disease, when detected at the early stages the dysplastic tissue can be treated achieving very high rates of the disease remission (93% after 10 years, still presenting 5% of morbidity and 0% of mortality) [Dent 2011, Sharma et al. 2016, Phoa et al. 2016]. Endoscopic resection (mucosal resection and submucosal dissection) and ablation techniques (radiofrequency ablation and cryoablation) appear to be promising methods developed for the management of BE, with the potential to reduce the adenocarcinoma risk in patients with dysplasia. However, limitations in the current methods for monitoring and evaluating the BE level highlighted the necessity to the design of additional tools to improve the detection of dysplasia [Shaheen et al. 2009, Johnston et al. 2005, Overholt, Panjehpour e Halberg 2003].

Many efforts were considered in the last years regarding machine learning and computer-

aided diagnosis. Van der Sommen [van der Sommen et al. 2016], for instance, designed a system capable of automatically extract features for detecting and delineating early neoplastic lesions in Barrett's esophagus. Other works [Souza Jr. et al. 2017, Hassan e Haque 2015] aimed to use features extracted from endoscopic images for the classification of Barrett's esophagus and adenocarcinoma. Furthermore, Mendel et al. [Mendel et al. 2017] proposed a deep learning approach based on Convolutional Neural Networks in the context of BE analysis. Recently, Souza et al. [Souza Jr. et al. 2017] conducted a study in which two approaches were introduced to distinguish between BE and adenocarcinoma: (i) the OPF [Papa, Falcão e Suzuki 2009, Papa et al. 2012] classifier; and (ii) the use of BoVW [Csurka et al. 2004, Peng et al. 2016] using PoIs extracted from endoscopic images using SURF [Bay et al. 2008] and SIFT [Lowe 2004] techniques [Souza Jr. et al. 2018] for the feature vector extraction.

Restricted Boltzmann Machines (RBMs) are nondeterministic neural networks composed of two layers of neurons, i.e., visible and hidden, whose main idea is to produce a probabilistic representation of a given input data in the hidden layer, such that the network is capable of reconstructing the data in the visible layer [Larochelle et al. 2012, Schmidhuber 2015]. The process is conducted using the minimization of the system's energy, analogous to the Maxwell-Boltzmann distribution law of thermodynamics. RBMs have been highlighted in the scientific community over the last years, as well as some variants concerning deep learning models, e.g., Deep Belief Networks (DBNs) [Hinton, Osindero e Teh 2006] and Deep Boltzmann Machines (DBMs) [Salakhutdinov e Hinton 2012], due to their outstanding results in a number of domains, such as human motion [Taylor, Hinton e Roweis 2006], classification [Larochelle et al. 2012], spam [Silva et al. 2016], anomaly detection [Fiore et al. 2013], and collaborative filtering [Salakhutdinov, Mnih e Hinton 2007], just to cite a few.

However, one of the main concerns related to RBMs is associated with the number of hidden units, which is application-dependent and has a great impact on the final results. Montufar and Ay [Montufar e Ay 2011] showed that an RBM with $2^{m-1} - 1$ hidden units is a universal approximator, where m stands for the number of visible (input) units. Moreover, such a large scale representation may not be efficient in practice, which motivated researchers to study models that can automatically increase their capacity during learning.

Cotê and Larochelle [Côté e Larochelle 2016] proposed an extension of the RBM that does not require specifying the number of hidden units, and it can increase its capacity (i.e., number of hidden units) during training, hereinafter called infinite RBM (iRBM). The learning achieved by adding new units in the hidden layer, where each one is trained gradually from left to right. Effectively, the number of hidden units increases automatically to a capacity that is

similar to the universal approximator (i.e., when the number of hidden units tends to infinite), though being much smaller. Such model is possible due to the following assumptions: (i) that a finite number of hidden units has non-zero weights and biases, and (ii) the parametrization of the per-unit energy penalty (β) ensures the infinite sums during probability computation will converge. Since the role of this energy penalty is only to ensure the iRBM is properly defined, the penalty imposed in the energy function can be compensated by the learned parameters (i.e., weight decay). Therefore, we can remove one hyper-parameter from the project with the cost of introduction of a less sensible one, i.e., the number of hidden units is removed and an extra parameter is introduced in the model.

Despite the advantage that iRBM brought by removing the need to properly select the number of hidden neurons, it also came up with a shortcoming related to the slow convergence. Peng et al. [Peng, Gao e Li 2018] attribute the problem to the initial correlation that is given by the ordering effect present in the iRBM, and proposed a solution by adding a probability of flipping the position of some neurons in the hidden layer while training, avoiding the dependency among each other. Additionally, they also proposed a mechanism to use the iRBM not only for binary image reconstruction but also for discriminative tasks, employing a “one-hot"vector representation of the sample’s label together with the feature vector while training the model for further classification of the test set.

Regardless dropping out the hidden units that are usually required beforehand, iRBM still demands the selection of the remaining hyper-parameters, such as the learning rate, the weight decay, and the β hyper-parameter, which despite less sensitive than the number of hidden units, still requires a proper fine-tuning. To deal with this problem, Passos et al. [Passos e Papa 2017] proposed to employ meta-heuristic optimization techniques to fine-tune the aforementioned hyper-parameters regarding binary image reconstruction since it has provided suitable results [Papa et al. 2015, Papa et al. 2015, Papa, Scheirer e Cox 2016, Rosa et al. 2016, Rosa et al. 2015, Rodrigues, Yang e Papa 2016]. However, as far as we know, such techniques were never used to fine-tune iRBMs regarding classification tasks. In this paper, we propose to find suitable hyper-parameters concerning the discriminative iRBM model using eight meta-heuristic techniques: Particle Swarm Optimization (PSO) [Rodrigues et al. 2015], Bat Algorithm (BA) [Yang e Gandomi 2012], Cuckoo Search (CS) [Yang e Deb 2010], Brain Storm Optimization [Shi 2011], Firefly Algorithm (FA) [Yang 2010], and the Harmony Search (HS) [Geem 2009]. Although one can use any other optimization technique, we opted to use these mainly because they are well recognized in the literature, and they do not require computing derivatives as usually demanded by standard optimization techniques.

In this paper, we also introduce the infinity Restricted Boltzmann Machines in the context of automatic classification of Barrett's esophagus using information extracted with SURF and SIFT techniques in the "MICCAI 2015 Endovis Challenge"dataset. The experiments performed a comparison of the aforementioned meta-heuristic optimization techniques regarding iRBM meta-parameter fine-tuning to the task of Barrett's esophagus classification. Additionally, we also considered both linear and Radial Basis Function SVMfor comparison purposes.

Therefore, the main contributions of this paper are fourfold: (i) to introduce iRBM in the context of Barrett's esophagus recognition, (ii) to promote the scientific literature concerning iRBMs, (iii) to foster the scientific literature concerning Barrett's esophagus, and (iv) to deal with the problem of iRBM hyper-parameter optimization concerning discriminative tasks. The remainder of this paper is organized as follows. Sections 7.2 and 7.3 present theoretical background and the proposed fine-tuning process, respectively. Section 7.4 discusses the methodology and Section 7.5 presents the experimental results. Finally, Section 7.6 states conclusions and future works.

7.2 Theoretical Background

In this section, we briefly explain the theoretical background related to the discriminative Infinity Restricted Boltzmann Machines and the dinamic training strategy.

7.2.1 Discriminative Infinity Restricted Boltzmann Machines

Larochelle and Bengio introduced a discriminative version of the RBM [Larochelle e Bengio 2008] to the task of classification, and Peng et al. [Peng, Gao e Li 2018] adapted the idea to the iRBM domain. In order to couple the labels in the formulation, the energy function is redefined as follows:

$$\mathbf{E}(\mathbf{v},\mathbf{h},y,z) = -\sum_{i=1}^{m} a_i v_i - \mathbf{e}_y d - \sum_{j=1}^{z} b_j h_j - \sum_{j=1}^{z} h_j \left(\sum_{i=1}^{m} \left(v_i w_{ij} \right) + \mathbf{e}_y u_{yj} + \beta_j \right), \tag{7.1}$$

where d is the bias of the label vector, $\mathbf{e}_y = \left(1_{y=1}\right)_{i=1}^{C}$ stands for the so-called "one-hot"representation of label $y \in \{1,2,\ldots,C\}$, and u_{yj} is the element from matrix $\mathbf{U}$ connecting the jth hidden unit to $\mathbf{e}_y$.

The distribution over y given the energy function (Eq. 7.1) is given by:

$$P(h_j = 1|\mathbf{v}, z) = \begin{cases} \phi\left(\sum_{i=1}^{m} w_{ij} v_i + u_{yj}\mathbf{e}_y + b_j\right) & \text{if } j \leq z \\ 0 & \text{otherwise,} \end{cases} \tag{7.2}$$

and

$$P(y|\mathbf{h}, z) = exp\left(\sum_{j=1}^{z} h_j\left(\mathbf{e}_y u_{yj}\right) + d_y\right) \Big/ \sum_{y'} exp\left(\sum_{j=1}^{z} h_j\left(\mathbf{e}_{y'} u_{y'j}\right) + d_{y'}\right), \tag{7.3}$$

having $P(v_i = 1|\mathbf{h}, z)$ defined exactly the same as in Equation 7.2. Figure 7.1 depicts the Discriminative iRBM.

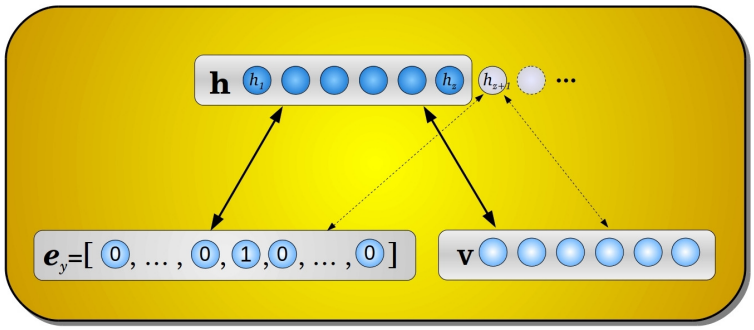


Figure 7.1: Discriminative iRBM. Both visible (v) and label ($\mathbf{e}_y$) layers are employed for training the model. A new hidden unit $\mathbf{h}_{z+1}$ is introduced in the model for learning purposes.

7.2.2 Dynamic Training Strategy

Despite the advantages achieved using iRBM, such as the absence of a hyper-parameter to be fine-tuned, i.e., the number of units in the hidden layer, it presents a shortcoming related to a slow convergence while training the network. The explanation concerns the time required by the filters to diverge from each other given an initial correlation imposed by the ordering effect intrinsic to iRBMs, where each newly added hidden unit suffers from the influence of the previously added ones, learning the features jointly and not by itself.

To cope with the issue, Peng et al. [Peng, Gao e Li 2018] proposed to use the approximated gradient descent algorithm together with the dynamic training strategy, which assumes that changing the order of the hidden units at each gradient descent step and jointly training iRBMs with all possible orders it is possible to alleviate the bias inherited from the ordering effect. The model employs a variable Q_t, which controls the proportion of units regrouped at step t. Additionaly, $\tilde{\mathbf{o}}$ stands for the vector of indexes to be permuted given a probability distribution.

The process is illustrated in Figure 7.2.

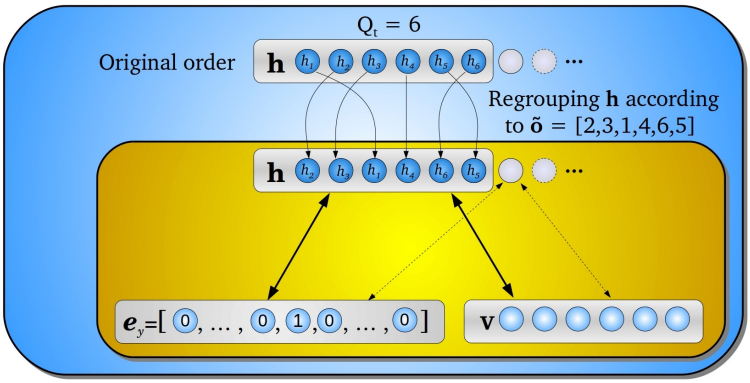


Figure 7.2: Dynamic training strategy proposed by Peng et al. [Peng, Gao e Li 2018], where Q_t Hidden units are permuted at time step t accordingly to the indexes defined in õ.

7.3 Infinity RBM Fine-Tuning as an Optimization Problem

The proposed approach requires the optimization of three hyper-parameters of the iRBM: (i) the learning rate η, (ii) the weight decay λ regularization parameter, and (iii) the β parameter. Notice the ADAGRAD stochastic gradient technique [Duchi, Hazan e Singer 2011] is employed as the learning rate. In this case, we have a per-dimension learning rate method, i.e., $\boldsymbol{\eta} \in \Re^n$, with $\varepsilon = 10^{-6}$ [Côté e Larochelle 2016]. In a nutshell, meta-parameters η and β can be interpreted as an n-sized vector, where n stands for the current number of hidden units. This latter parameter stands for a small number to avoid numerical instabilities. Cotê and Larochelle [Côté e Larochelle 2016] claim that one can throw away the parameter n, thus replacing the RBM model by the iRBM one. However, $\boldsymbol{\beta}$ is still a hyper-parameter to be optimized. Notice the regularization parameter $\boldsymbol{\beta}$ is way less sensitive than the number of hidden units $\mathbf{n}$.

Figure 7.3 depicts the proposed approach to optimize the iRBM model, where the idea is to initialize all decision variables at random, and then the optimization algorithm takes place. In this work, we used the following ranges concerning the parameters: $\boldsymbol{\eta} \in [0,0001, 0,5]$, $\boldsymbol{\beta} \in [0,01, 1,5]$ and $\lambda \in [0,00001, 0,01]$.

In order to fulfill the requirements of any optimization technique, one shall design a fitness function to guide the search into for best solutions. In this paper, we used the average accuracy of the training set considering the task of classification as the fitness function. Furthermore, we present the mean results obtained over 20 runs to provide a statistical comparison.

In short, the optimization technique selects the set of hyper-parameters that maximizes the classification accuracy over the training set considering a set of features extracted from

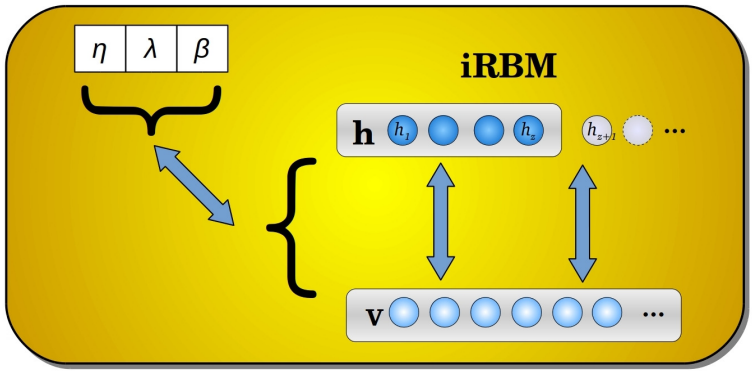


Figure 7.3: Proposed approach to model the iRBM fine-tuning problem as an optimization task.

endoscopic images using BoVW over SIFT and SURF features as an input to the model. After learning the hyper-parameters, one can proceed to the classification step concerning the testing samples. Regarding this work, the following approaches are conducted: (i) the set of meta-parameters that best fits the model is selected using the validation set over a reduced number of 150 epochs for convergence purposes, and (ii) afterwards, the selected set of best meta-parameters are used to train the network using $1,500$ epochs and to perform the classification over the test set.

7.4 Methodology

In this section, we present the methodology employed to evaluate the optimization techniques, the dataset, and the experimental setup.

7.4.1 Optimization Techniques

This work employs six metaheuristic techniques to the task of iRBM fine-tuning, i.e., PSO, BA, CS, BSO, HS, and FA, presented in Section 7.1.

Table 7.1 presents the parameters used for each aforementioned optimization technique, where five agents (initial solutions) were used for all optimization techniques during 10 iterations for convergence purposes. Notice these parameters were set empirically. In regard to PSO, w stands for the inertia weight, and c_1 and c_2 control the step size towards the best local and global solutions, respectively. With respect to BA, f_{min} and f_{max} bound the minimum and maximum frequency values, and A and r denote the loudness and pulse rate values, respectively. Regarding BSO, p_{gen} defines a probability whether a new solution will be generated by one or two other individuals, k stands for the number of clusters composed of similar ideas, and

$p_{oneCluster}$ and $p_{twoCluster}$ stand for the probability of creating a new solution based on only one or two clusters, respectively. Parameters φ and τ are used to avoid the technique getting trapped from local optima. FA uses μ and γ, which stand for a random perturbation and the light absorption coefficient, respectively. Variable ς denotes the attractiveness of each firefly. Furthermore, Harmony Memory Considering Rate (HMRC) and Pitch Adjusting Rate (PAR) are used by HS for responsible creating new solutions based on previous experience of the music player and applying some disruption to the created solution in order to avoid local optima, respectively. Finally, CS uses Γ to compute the Lévy distribution, ζ for the switching probability (i.e., the probability of replacing the worst nests by new ones), and s for the step size.

Table 7.1: Parameter configuration for each optimization technique.

Technique	Parameters
BA	$r = 0.5$, $A = 1.5$ $f_{min} = 0$, $f_{max} = 100$ $\varphi = 0.9$, $\tau = 0.9$
BSO	$p_{gen} = 0.4$, $k = 2$ $p_{oneCluster} = 0.8$, $p_{twoCluster} = 0.5$
CS	$\Gamma = 1.5$, $\zeta = 0.25$, $s = 0.8$
FA	$\gamma = 1$, $\varsigma = 1$, $\mu = 0.2$
HS	$HMCR = 0.7$, $\rho = 10$, $PAR = 0.7$
PSO	$c_1 = 1.7$, $c_2 = 1.7$, $w = 0.7$

7.4.2 Datasets

The information (i.e., features) were extracted from a dataset of images from Barrett's esophagus and adenocarcinoma called "MICCAI 2015", which was provided at the "MICCAI 2015 EndoVis Challenge"[1]. Such dataset is composed of 100 endoscopic images of the lower esophagus from 39 individuals, 22 presenting esophageal adenocarcinoma and 17 diagnosed with early-stage Barrett's esophagus. For each patient, several endoscopic images were made available, ranging from one to eight. A total of 50 images showing cancerous tissue areas and 50 images showing dysplasia without cancer compose the dataset. The injured tissue observed in the cancerous images have been delineated by five different endoscopy experts. Figure 7.4 shows some dataset samples and their respective delineation performed by the experts.

[1] https://endovissub-barrett.grand-challenge.org/

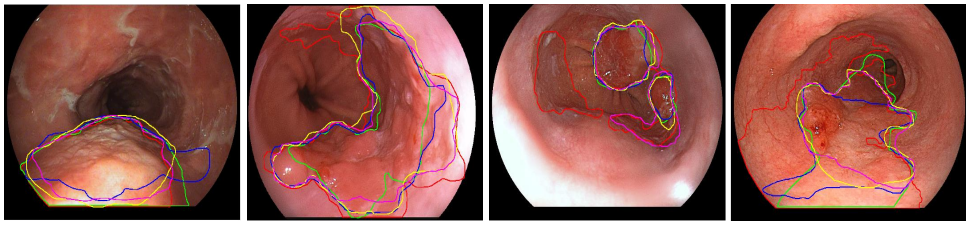


Figure 7.4: Some samples from the Barrett's Endovis 2015 Challenge [Souza Jr. et al. 2017].

7.4.3 Experimental Setup

This work employs a cross-validation procedure with 20 runs to provide a statistical analysis using the Wilcoxon signed-rank test with a significance of 0.05 [Wilcoxon 1945]. Regarding the task of meta-parameter optimization, we conducted the experiments over 150 epochs to find the meta-parameters that lead to the best classification accuracies regarding the validation set. Finally, the network was trained once again over $1,500$ epochs using the best parameters found by each meta-heuristic technique to classify the testing set. The learning procedure was conducted using Persistent Contrastive Divergence (PCD) [Tieleman 2008] using three Gibbs sampling steps with mini-batches of size 5.

Finally, the codes used to reproduce the experiments are available on GitHub[2,3]. The experiments were conducted using an Ubuntu 16.04 Linux machine with 8Gb of RAM running an Intel Core™i5 – 2410*M* with a frequency of 2.30 GHz and a GPU GeForce® GT540*M* with 2GB. The source-codes run on top of Matlab and C for the iRBM and optimization approaches, respectively.

7.5 Experiments

This section describes the experiments as follows: Section 7.5.1 presents the image feature extraction using points-of-interest together with a Bag-of-Visual-Words schema, and Section 7.5.2 discusses the optimization steps as well as the time consumption regarding the role optimization process. Sections 7.5.3 and 7.5.4 present the procedures adopted while training and testing the model, respectively.

[2]iRBM: https://github.com/Boltzxuann/RP-iRBM
[3]LibOPT [Papa et al. 2017]: https://github.com/jppbsi/LibOPT

7.5.1 Feature Extraction

The points-of-interest were calculated using the Speed-Up Robust Features and the Scale-Invariant Feature Transform techniques, and then the feature vectors were calculated using Bag-of-Visual-Words. The SURF technique ensures scale and spatial invariance, seeking for maxima of the determinant of the Hessian matrix, demarcating specific key points which are explored in their local neighborhood resulting in a feature vector of size 64. The SIFT algorithm operates on image regions calculating features that are invariant to scaling and rotation. It seeks for the scale-space extrema detection evaluating the image scales (difference-of-Gaussian function) providing feature vectors of size 128. Finally, the BoVW technique uses points-of-interest from a set of reference images to generate a visual dictionary that is employed in the training and testing phases. For this work, we considered dictionaries with two different sizes: 500 and 1,000 [Souza Jr. et al. 2017]. In order to compose such dictionaries, two well-known techniques were considered: (i) k-means and (ii) random selection. Figure 7.5 illustrates the feature vector calculation for the experiments.

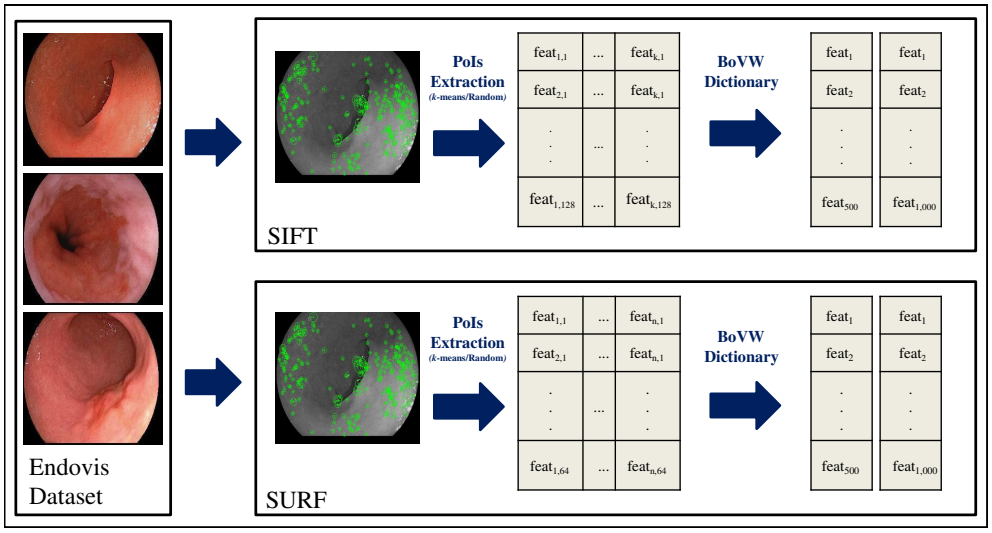


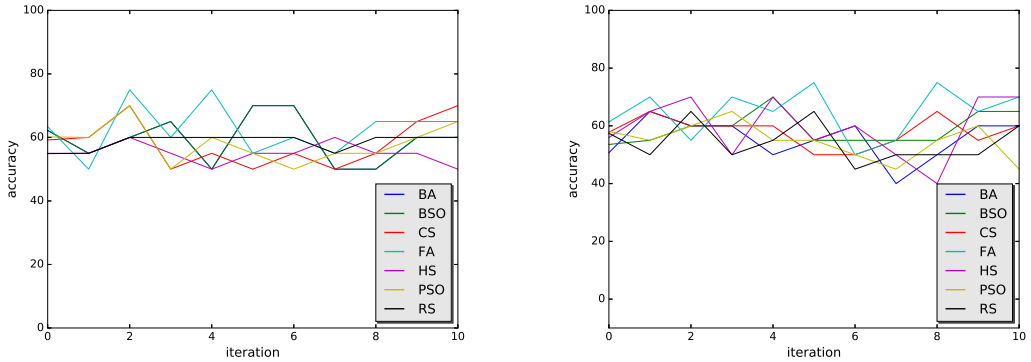
Figure 7.5: Descriptor calculation for the experiments using SURF, SIFT and BoWV techniques (adapted from [Souza Jr. et al. 2017]).

7.5.2 Optimization

Six meta-heuristic optimization techniques, i.e., BA, BSO, CS, FA, HS, and the PSO, were employed in this work to fine-tune the iRBM meta-parameters: learning rate η, weight decay λ, and the beta β. All techniques were initialized with five agents and executed during 10 iterations over 150 epochs. Additionally, we started the model using random variables and executed the iRBM for 15 runs over 150 epochs, hereinafter called Random Search (RS) for comparison

purposes.

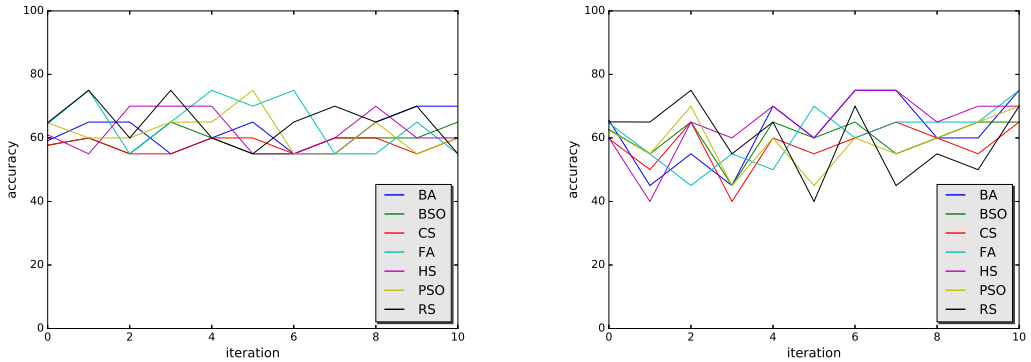
Figures 7.6 and 7.7 present the results obtained while fine-tuning the model over the validation sets regarding 500 and 1,000 visual words, respectively. The most accurate iterations were selected for each technique for visualization purposes. Notice that iteration zero stands for the average of the five initial agents for the optimization techniques, as well as the first five runs for RS.



(a) (b)

Figure 7.6: Classification accuracies over the validation set during the meta-parameter optimization process concerning 500 visual words for SIFT (a) and SURF (b).

Despite the oscillatory behavior, one can notice that FA obtained the highest results for both SIFT and SURF techniques over 500 visual words, reaching 75% of accuracy during a few iterations (Figure 7.6). It also reached the best results at the end of the optimization steps, which reflects the best values obtained over the testing set.



(a) (b)

Figure 7.7: Classification accuracies over the validation set during the meta-parameter optimization process concerning 1,000 visual words for SIFT (a) and SURF (b).

Regarding 1,000 visual words, Figure 7.7 depicts a behavior similar to Figure 7.6, with considerable oscillation and FA obtaining the best results, which once again reflects on the

results of the test set, presented in Table 7.5. The Harmony Search and the Bat Algorithm behave similarly to FA, also achieving 75% over a few iterations. Additionally, the random search also obtained results statistically similar to the ones found by FA, which can be explained by the short number of agents and iterations employed for the optimization convergence, i.e., 5 and 10, respectively. Furthermore, the random search presents an even more oscillatory behavior, as shown in Figure 7.7(b).

Tables 7.2 and 7.3 present the average execution time regarding 20 executions concerning 500 and 1,000 visual words, respectively. Clearly, HS and RS obtained the lowest execution time for both configurations. Since HS (and also RS) updates a single agent for each iteration and the remaining techniques update all the agents for each iteration, it is expected that HS to be faster. In a nutshell, for the configuration employed in this work with 5 agents over 10 iterations, HS and RS evaluate the fitness function 15 times, while the others evaluate 5 times (initialize each agent) and then update each one for 10 times, ending up in 55 executions. Notice CS also presents a small execution time, due to the solutions discarded without evaluation (eggs abandoned by the host bird).

		BA	BSO	CS	FA	HS	PSO	RS
k-means	SIFT	49.55	50.08	22.81	45.12	14.28	49.92	14.29
	SURF	49.96	51.01	23.42	46.38	14.17	50.44	14.50
Random	SIFT	44.93	51.97	24.05	45.29	14.45	51.19	14.99
	SURF	48.74	52.33	25.08	45.31	14.37	50.48	15.27

Table 7.2: Mean computational load (in minutes) of each technique applied to the BE and adenocarcinoma problem using 500 visual words for the feature vector calculation.

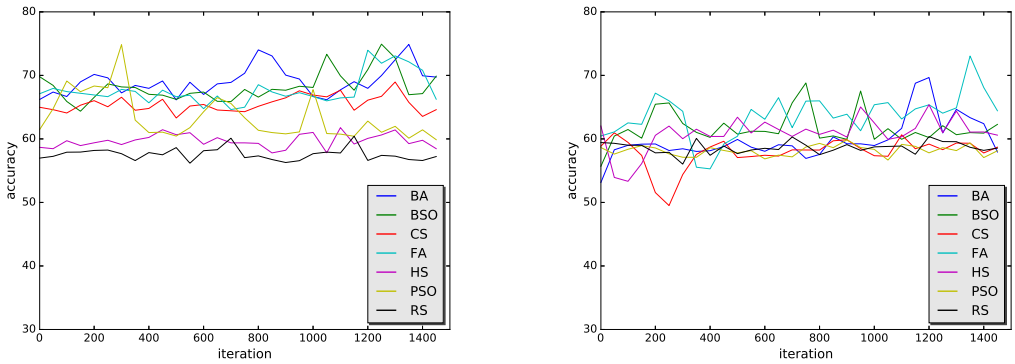
		BA	BSO	CS	FA	HS	PSO	RS
k-means	SIFT	50.45	53.27	23.84	47.17	14.10	48.76	14.34
	SURF	49.75	52.15	23.20	45.45	14.29	50.12	14.45
Random	SIFT	49.94	50.98	23.27	46.92	14.25	48.63	14.45
	SURF	49.99	51.64	24.12	45.15	14.29	50.73	14.38

Table 7.3: Mean computational load (in minutes) of each technique applied to the BE and adenocarcinoma problem using 1000 visual words for the feature vector calculation.

7.5.3 Training

The experiments presented in this section employ the best meta-parameters obtained in Section 7.5.2 for each meta-heuristic optimization technique, i.e., the combination of learning rate, weight decay, and β that provided the best accuracies over the validation set during 150 epochs. The iRBM was trained once again, however using 1,500 epochs.

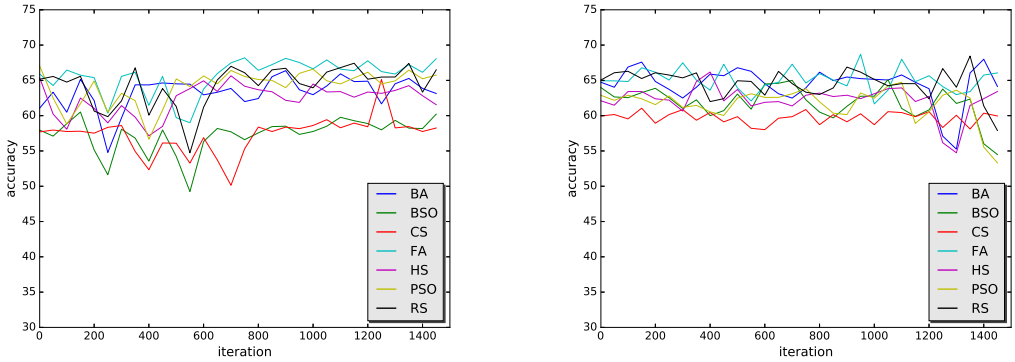
Figure 7.8 depicts the learning steps concerning SIFT and SURF features over the 500-sized dictionary. Despite the oscillatory behavior, which probably can be attributed to the dynamic training strategy [Peng, Gao e Li 2018] described in Section 7.2.2, one can notice that FA, BSO, and BA interchanged the highest results regarding Figure 7.8(a), while the random search obtained the less accurate results. Concerning SURF, FA also presented the best results, as depicted in Figure 7.8(b). Furthermore, FA appears more inclined to a slight growth behavior than the others techniques, regardless the oscillations.



(a) (b)

Figure 7.8: Classification accuracies during the training convergence process concerning a dictionary composed of 500 **visual words for SIFT (a) and SURF (b).**

A similar behavior is observed in Figure 7.9, which concerns SIFT and SURF with a dictionary of size $1,000$. However, FA interchanges the top results with BA, HS, PSO and even the random search. In spite of such exchange, FA seems to stand for the best optimization technique overall.



(a) (b)

Figure 7.9: Classification accuracies during the training convergence process concerning a dictionary composed of $1,000$ **visual words for SIFT (a) and SURF (b).**

7.5.4 Classification Step

The experiments conducted in this section are divided according to the feature extraction technique and visual dictionary sizes. Tables 7.4 and 7.5 present the mean accuracy and the standard deviation concerning the classification over the test set during $1,500$ training epochs using 500 and $1,000$ visual words, respectively. The training employs the combination of meta-parameters that provided the best accuracies over the validation set during the fine-tuning step, presented in Section 7.5.2. Additionally, results are compared against two versions of Support Vector Machines with RBF and linear kernels, as well as the well-known Bayes classifier. Moreover, the experiments were executed for 20 runs for statistical analysis using Wilcoxon signed-rank test [Wilcoxon 1945] with 0.05 of significance, being the most accurate results in bold.

	SIFT		SURF	
	k-means	Random	*k*-means	Random
iRBM-BA	**63.50±10.6184**	55.65±6.1846	61.60±7.0183	57.40±9.1311
iRBM-BSO	**64.85±9.3770**	51.85±2.9924	**62.70±8.5032**	54.65±3.2247
iRBM-CS	61.40±8.6628	57.45±7.7751	60.97±4.3011	55.90±3.7381
iRBM-FA	**66.15±11.5023**	49.95±2.0504	**66.35±4.7022**	58.80±4.1389
iRBM-HS	59.95±6.2954	55.75±8.1319	**65.35±6.4570**	56.00±5.2749
iRBM-PSO	**65.55±6.5370**	52.50±1.5003	**64.80±6.0156**	51.20±2.7254
iRBM-RS	58.45±5.0028	56.90±3.7025	60.20±6.45883	59.15±7.5520
SVM-RBF	**65.50±9.8467**	**65.60±10.4255**	**64.80±8.2496**	**63.40±11.2731**
SVM-Linear	56.80±4.5527	54.50±6.8608	58.60±6.5392	57.60±4.4458
Bayes	59.98±3.449	53.00±3.5192	56.86±3.1080	53.52±4.6362

Table 7.4: Best accuracy values for the "MICCAI 2015 Endovis Challenge"Dataset using 500 visual words.

Regarding Table 7.4, one can notice that iRBM fine-tuning with FA outperformed all the other techniques, obtaining the most accurate classification results. Additionally, iRBM optimized with all meta-heuristic techniques, except CS and the random search, achieved similar results concerning the Wilcoxon test, as well as SVM-RBF. Such results may lead to two assumptions: (i) meta-heuristic optimization techniques are suitable for fine-tuning iRBM meta-parameters concerning classification tasks, as well as for reconstruction tasks [Passos e Papa 2017], since the results obtained using such techniques outperformed the ones obtained with a random initialization of the weights, and (ii) iRBM is appropriate for classification tasks since it outperformed the results obtained by some well-established classifiers, such as SVM and Bayes.

Concerning the results presented in Table 7.5, the most accurate results were obtained by

	SIFT		SURF	
	k-means	Random	*k*-means	Random
iRBM-BA	60.81±7.4686	60.01±7.2947	**65.38±7.6846**	60.04±7.4913
iRBM-BSO	58.40±8.8101	60.61±6.7935	60.41±8.6898	58.79±7.5418
iRBM-CS	59.57±10.4727	59.43±9.2710	62.09±7.9337	58.71±8.9159
iRBM-FA	**66.01±9.6896**	62.21±6.3956	**67.00±8.1187**	60.4±9.2183
iRBM-HS	61.32±6.7106	61.08±11.1132	**65.38±6.4580**	59.46±7.1040
iRBM-PSO	62.84±6.8391	60.60±8.0681	59.80±7.8694	58.10±10.0872
iRBM-RS	**65.10±6.0221**	**63.79±6.1503**	**66.30 ±9.3325**	61.71±6.2307
SVM-RBF	**64.10±8.0761**	**63.70±7.0523**	**62.60±7.2304**	**62.10±6.3215**
SVM-Linear	**67.30±9.8649**	52.70±3.5027	**62.80±4.9553**	56.50±4.0050
Bayes	60.70±3.7278	54.37±3.0303	57.43±4.2186	56.86±6.2625

Table 7.5: Best accuracy values for the "MICCAI 2015 Endovis Challenge"Dataset using $1,000$ **visual words.**

SVM-Linear and once again the iRBM fine-tuned using the FA technique, which suggests the idea that FA is the most effective meta-heuristic optimization technique with respect to classification tasks using iRBM. Furthermore, similar results were obtained by SVM-RBF, as well as the iRBM using the HS, BA, and the RS as parameters fine-tuning. The explanation for finding competitive accuracies concerning a random initialization of the meta-parameters may be due to the short number of agents and iterations employed for meta-heuristic optimization techniques. Moreover, one can notice that the configuration using *k*-means obtained the best results regarding both SIFT and SURF, as well as in with both 500 and $1,000$ words, which suggests that dictionaries composed by employing *k*-means generate better features.

Considering the very best results obtained for all the techniques, i.e., iRBM, SVM-RBF, SVM-Linear, and Bayes, Tables 7.6 and 7.7 present the sensitivity (SE) and the specificity (SP) results concerning the configuration over 500 and $1,000$ words, respectively. Notice the best values are in bold.

	Accuracy	Sensitivity	Specificity
iRBM-FA	**66.35%**	**0.644**	0.687
SVM-RBF	65.60%	0.612	**0.706**
SVM-Linear	58.60%	0.582	0.593
Bayes	59.98%	0.593	0.605

Table 7.6: Mean SE and SP values for the selected best results obtained using dictionaries of 500 **words.**

One can observe that iRBM obtained the best Sensitivity for both configurations, which

	Accuracy	Sensitivity	Specificity
iRBM-FA	67.00%	**0.655**	0.692
SVM-RBF	64.10%	0.632	0.686
SVM-Linear	**67.30%**	0.583	**0.767**
Bayes	60.70%	0.581	0.638

Table 7.7: Mean SE and SP values for the selected best results obtained using dictionaries of $1,000$ words.

indicates a higher rate of true positives correctly identified. Such outcome is particularly interesting for medical issues, since a correct identification of some illness, specially in early stages, may prevent the progress of the disease. Additionally, it can be also observed that using either configurations, i.e., dictionaries of 500 and $1,000$ words, does not impact in the final classification accuracy since both scenarios achieved similar results.

7.6 Conclusions and Future works

This work dealt with the problem of automatic Barrett's esophagus identification using infinity Restricted Boltzmann Machines for classification purposes. The approach employs SURF and SIFT techniques to extract the points-of-interest, which are used to build structural patterns in endoscopy images in association with a BoVW. Such descriptors were calculated over a set of images previously annotated by five experts, making the identification of malignant lesions available for classification. Additionally, experiments were conducted over two different dictionary configurations, i.e., 500 and $1,000$ words.

From the experiments, we can conclude that: (i) infinity Restricted Boltzmann Machines are convenient for Barrett's esophagus identification task, since it outperformed SVM-Linear and SVM-RBF, as well as the Bayes classifier in one of the configurations, and achieved similar results concerning the other; (ii) meta-heuristic optimization techniques are suitable for iRBM meta-parameter optimization, since they outperformed a random search over an equal number of executions; (iii) the identification of Barrett's esophagus is not a trivial task, once all techniques obtained results under 70% of accuracy.

Based on the last assumption, we can also conclude that Barrett's esophagus identification requires more study and RBM-based approaches may offer an interesting direction in such context, once they are the base blocks for deeper architectures, i.e., Deep Belief Networks and Deep Boltzmann Machines, which are able of extracting deeper characteristics and correlations from data.

Considering future works, we intend to investigate the identification of Barrett's esophagus using RBM-based deeper architectures, as well as other deep learning techniques. Furthermore, we aim to consider the suitability of deeper versions of the iRBM.

7.7 Chapter's Considerations

As a final attempt in the evaluation of handcrafted features for BE and adenocarcinoma identification, in this manuscript we attended to employ iRBM, a nondeterministic neural networks composed of two layers of neurons, i.e., visible and hidden, to produce probabilistic representations in the hidden layer and further reconstruction of data in the visible layer. Plus, iRBM does not require specifying the number of hidden units, making its capacity flexible during training. An extra consideration about iRBM is related to properly selecting its training hyper-parameters, such as learning rate, weight decay, among others, to ensure fast and precise convergence. To cope with such a task, the iRBM hyper-parameter optimization was performed by the application of metaheuristic techniques, such as Cuckoo Search and Particle Swarm Optimization.

Therefore, the experimental delineation was based on selecting features from SURF and SIFT key-points, optimizing the iRBM hyper-parameters by employing metaheuristic techniques during the training phase, and finally, classifying BE and adenocarcinoma samples based on a model with the best hyper-parameters already calculated. We believe that this ablation process of defining the optimal set of hyper-parameters may enhance the generalization performed by iRBM.

Just like in Chapters 4 and 6, the feature vector calculation was based on a BoVW-k-means approach that selects the most representative ones, the prototypes, for further computing the descriptor for each dataset sample. In this manuscript, SURF and SIFT key-points were selected for the feature extraction task. For the hyper-parameter optimization task, we employed the following metaheuristic techniques: Particle Swarm Optimization, Bat Algorithm, Cuckoo Search, Brain Storm Optmization, Firefly Algorithm, and Harmony Search.

In a close observation of the results presented in tables 7.6 and 7.7, the hyper-parameter optimization process could enhance the iRBM performance compared to other classifiers. However, compared to previous results, we could not outperform the ones presented in previous Chapters. In conclusion, we observed that considering how difficult the definition of early cancer is in BE samples, the iRBM classifier could not converge properly, providing a generalization that even presenting an optimized hyper-parameter configuration, could not avoid the visual-biased

design this context presents.

By finishing the evaluation of iRBM with nonempirical parameters, still based on object detector techniques of description, we covered a fair amount of description techniques, classifiers, and experimental delineations for the human representation of early cancer and BE, and in other words, the handcrafted features. Very satisfactory outcomes could be achieved so far, highlighting that the human knowledge, in quantitative and qualitative ways, presents essential influence in the correct representation and further classification of BE and adenocarcinoma using computational resources. From this Chapter on, we explored the use of deep learning techniques for understanding the computational focus in such identification, aiming at comparing its representation with what we achieved until this point in our research.

Chapter 8

ASSISTING BARRETT'S ESOPHAGUS IDENTIFICATION USING ENDOSCOPIC DATA AUGMENTATION BASED ON GENERATIVE ADVERSARIAL NETWORKS

This work introduces Generative Adversarial Networks to generate high-quality endoscopic images, thereby identifying Barrett's esophagus and adenocarcinoma more precisely, in a way to solve the lack of available data drawback. Further, Convolution Neural Networks are used for feature extraction and classification purposes. Such study was published at "Computers in Biology and Medicine"journal [Souza Jr. et al. 2020].

8.1 Introduction

Barrett's esophagus (BE) is a condition that changes the mucosal cells of the lower part of the esophagus considerably. Risk factors, such as obesity, contribute to rising the number of patients, mainly in western populations [Lagergren e Lagergren 2010, Dent 2011, Lepage, Rachet e Jooste 2008]. The remission of the disease is directly related to early diagnosis, thus delivering the treatment with reduced rates of morbidity and mortality, as well as a complete remission after 10 years of treatment [Dent 2011, Sharma et al. 2016, Phoa et al. 2016].

Techniques widely used in clinics, such as optical coherence tomography, confocal laser endomicroscopy, and chromoendoscopy, have been employed to BE and adenocarcinoma screening, turning the manual evaluation more accurate by characterizing esophageal histology through *in vivo* experiments [Sharma et al. 2015]. However, BE can often be misdiagnosed du-

ring endoscopy due to the inability to distinguish the columnar mucosa of the proximal stomach from the metaplastic epithelium in the distal esophagus. Usually, it is recommended to employ the Seattle biopsy protocol to handle lesions (i.e., dysplastic tissue) presented in white light endoscopy examination. Even though the protocol recommends the extraction of four biopsies sampled at every 1 cm, such a procedure is still susceptible to failures since those samples may not be large enough for a proper evaluation, ending up in erroneous diagnoses [Abrams et al. 2009, van der Sommen et al. 2014, Sharma et al. 2015].

Despite the developments in interventional therapies for handling adenocarcinoma and Barrett's esophagus, e.g., endoscopic resection and ablation techniques, which present a high potential for reducing the cancer risk in dysplasia-diagnosed patients, their limitations must be handled with methods for monitoring and evaluating the state of the disease, thus improving dysplasia detection [Shaheen et al. 2009, Johnston et al. 2005, Overholt, Panjehpour e Halberg 2003]. Computer-aided analysis of Barrett's esophagus and adenocarcinoma figures as a powerful tool that has been subjected to intensive research in the past years [Souza Jr. et al. 2018]. Works have been conducted to evaluate the use of handcrafted features of endoscopic images based on texture and color, while others evaluate the well-known Convolutional Neural Networks (CNN) to identify the disease automatically. Recent works proposed by Souza Jr. et al. [Souza Jr. et al. 2018, Souza Jr. et al. 2019, Souza Jr. et al. 2017, Souza Jr. et al. 2018, Souza Jr. et al. 2017], Mendel et al. [Mendel et al. 2017], and Passos et al. [Passos et al. 2019] are some examples that make use of artificial intelligence techniques for classification purposes.

The works proposed by Souza Jr. et al. [Souza Jr. et al. 2018, Souza Jr. et al. 2019, Souza Jr. et al. 2017, Souza Jr. et al. 2018, Souza Jr. et al. 2017] and Passos et al. [Passos et al. 2019] studied the use of image representation techniques to describe and classify adenocarcinoma and Barrett's esophagus regions. For such works, the use of feature extraction techniques, such as Speeded-up Robust Features (SURF) [Bay et al. 2008] and Scale-Invariant Feature Transform (SIFT) [Lowe 2004], as well as classifiers, such as Support Vector Machines (SVMs), Optimum Path-Forest (OPF) [Papa, Falcão e Suzuki 2009, Papa et al. 2012], and the infinite Restricted Boltzmann Machine (iRBM) [Peng, Gao e Li 2018], were considered to provide the class prediction of the injured regions. Mendel et al. [Mendel et al. 2017, Ebigbo et al. 2020, Ebigbo et al. 2019] introduced the use of deep learning techniques to classify expert-annotated images of the esophagus presenting adenocarcinoma and Barrett's esophagus. A Convolutional Neural Network, together with transfer learning, was applied in a leave-one-patient-out protocol. Recently, some new deep learning approaches were also considered for the automatic classification and identification of neoplasia regions in endoscopic images [de Groof et al. 2020, van der Putten et al. 2020].

The most common drawbacks related to computer-assisted BE and adenocarcinoma identification are related to the absence of data mostly, once most of the datasets figure only a limited number of patients. Another obstacle concerns the lack of public datasets, which limits the development of more effective methods to detect early-stage illness in medical images. Recently, the data acquisition bottleneck has been coped with data augmentation (DA) methods, which figure as an interesting option to increase the amount of data. The application of data augmentation can be justified for several reasons. The first one concerns oversampling the minority class when trying to learn from imbalanced datasets. Such a scenario is challenging since the classifier tends to be biased toward the majority class. The other reason is to avoid the use of the original dataset for privacy reasons. Datasets, especially from the medical field, may provide personal legal information that can pose risks when misused [Tanaka e Aranha 2019].

Among the methods for data augmentation, Generative Adversarial Networks (GANs) [Goodfellow et al. 2014] have presented significant improvements in image generation, mainly for medical imaging [Han et al. 2019]. The primary idea related to GANs is to train a generator and a discriminator simultaneously. While the former employ inputs from easy-to-sample random sources (e.g., noise) for artificial image generating, the latter determines whether the samples belong to a real or an artificial image based on a learning process. The main goal of GANs is to generate artificial images confident enough to confuse the discriminator, which sends back an output gradient to the generator during the training process to improve the quality of the synthetic images. Concerning the GANs original formulation proposed by Goodfellow et al. [Goodfellow et al. 2014], the final discriminator's output is the probability of a set of images to belong or not to the original dataset, and hypothetically, the convergence is given when the generator and the discriminator reach the Nash-equilibrium. The use of GANs-based synthetic images for medical imaging presents a definite trend to be followed, thus handling several problems, such as lack of data or fragmented samples that compose the dataset. Some examples include using GANs to identify brain tumors in magnetic resonance images [Han et al. 2020], skin lesion detection [Frid-Adar et al. 2018], and segmentation of lung nodules in computed tomography images [Jin et al. 2018].

Currently, a wide range of works has proposed to using GANs for many purposes, such as a new approach called SaliencyGAN that is composed of concatenated GANs with partially shared parameters [Wang et al. 2020]; the generation of high-quality and precise medical data using an invariant deformation model based on the deformation-invariant CycleGAN (DicycleGAN) architecture and the spatial transformation network [Wang et al. 2019]; and the reconstruction of compressed sensing magnetic resonance images and fast magnetic resonance images by the application of Deep De-Aliasing GANs [Yang et al. 2018, Yu et al. 2017]. These recent investi-

gations show up how interesting and trending the use of GANs may be for an extensive amount of different research fields. Some interesting works addressed the use of GANs to provide additional data and represent new and not trivial expected scenarios due to their large applicability. The work proposed by Han et al. [Han et al. 2019] concerns CNNs trained on synthetic samples of magnetic resonance images of the brain. The main goal was to propose a two-step-based GAN method for generating, separately, samples with and without tumors, for further classification purposes. Finally, Sandfort et al. [Sandfort et al. 2019] proposed a study in which a CycleGAN was employed in the generation of non-contrast-computed tomography samples for further data augmentation and segmentation using a pre-trained U-Net.

This work aims at investigating the feasibility of using Generative Adversarial Networks to generate synthetic images from the esophagus and further assist the identification of Barrett's esophagus and adenocarcinoma. Therefore, the main contributions of this work are the following:

- to introduce Generative Adversarial Networks to augment data from endoscopic images aiming at distinguishing between Barrett's esophagus from adenocarcinoma;
- to assess how the amount of synthetic data added to the original datasets in the data augmentation process can affect BE and adenocarcinoma identification;
- to evaluate whether it is more effective to generate synthetic patches or the whole image. It is also critical to define that the generation of full images presents a high computational cost, while the patch generation requires less computational power due to the lower output resolution. Even making sense generating both outputs, it is reasonable to assess which one fits better the final evaluation considering such experimental design's pros and cons.

The remainder of this work is presented as follows. Section 8.2 presents a theoretical background concerning the Generative Adversarial Networks and the variant used in this work, while Section 8.3 describes the methodology and the proposed method. Finally, Sections 8.4 and 8.5 state the experiments and conclusions, respectively.

8.2 Generative Adversarial Networks

Goodfellow et al. [Goodfellow et al. 2014] introduced the concept of GANs, which is a framework composed of two networks, i.e., a generator G and a discriminator D, that contest with each other to generate data with the same (or similar) statistics (i.e., data distribution) as

the training set. The generator attempts to learn, from random inputs, how to generate data to fool the discriminator in a game-like approach. The learning procedure aims at minimizing the following loss function:

$$\min_G \max_D \mathscr{L}(D,G) = E_x[\log D(x)] + E_z[\log(1 - D(G(z)))], \tag{8.1}$$

where E_x stands for the expected value over all real data instances, and E_z corresponds to the expected value over all random inputs to the generator. In other words, the latter term defines the expected value over all fake instances generated by G. Last but not least, $D(x)$ and $D(G(z))$ are the probabilities given by the discriminator that instance x is real and that the fake instance z generated by G is also real, respectively. While the goal of the generator is to minimize its loss, the discriminator tries to maximize it. In other words, the Nash-equilibrium is reached during training, in which neither the discriminator nor the generator can improve their utilities unilaterally. Figure 8.1 depicts the overall mechanism concerning the Generative Adversarial Networks.

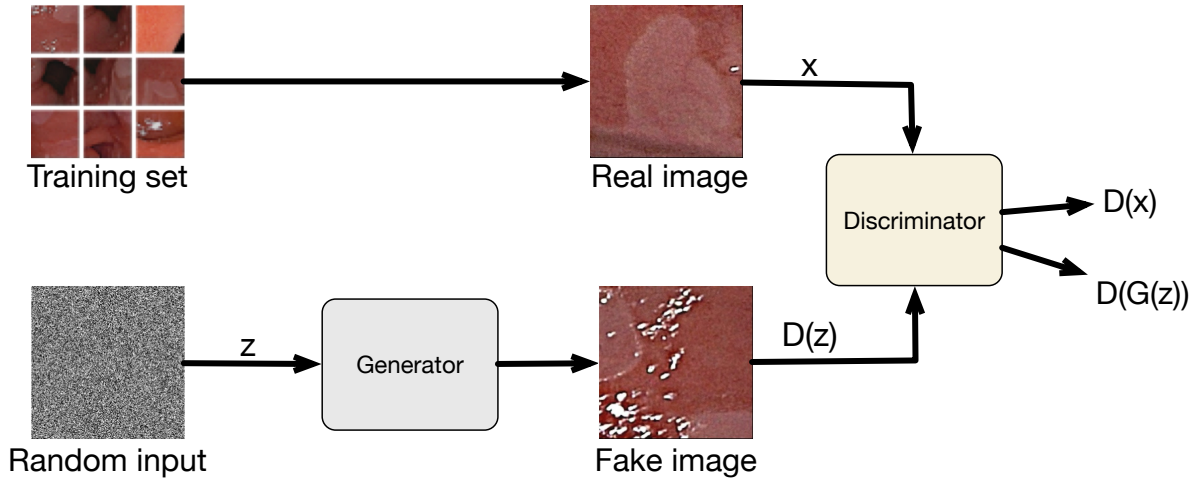


Figure 8.1: Overall mechanism of standard Generative Adversarial Networks. Based on the training set and a random input (noise, for instance), the generator network keeps producing synthetic samples to be evaluated by the discriminator network. In the end, the main goal is to provide samples as similar as possible to the training set, so that the discriminator will not be able to classify them correctly as "fake".

Concerning the medical applications, GANs can be of great interest due to their capabilities in generating synthetic data. The work proposed by Shin et al. [Shin et al. 2018] showed improvements in brain tumor identification when augmented data was introduced. The study conducted by Karras et al. [Karras et al. 2017] presented promising results in the implementation of a progressive growing of GANs to generate realistic mammography images with resolution up to $1,280 \times 1,024$ pixels. Realistic-looking retinal images were also synthesized using GANs by Zhao et al. [Zhao et al. 2018] based on a small dataset and a style transfer. Finally,

the work conducted by Middel et al. [Middel, Palm e Erdt 2019] proposed a method in which GANs were trained to generate realistic medical images obtained from the National Institutes of Health repository.

Among the number of GANs variants, the Deep Convolutional GANs (DCGAN) was proposed by Radford et al. [Radford, Metz e Chintala 2015] and motivated by scaling up GANs using of CNN models. The core of the DCGAN approach comes by adopting and modifying three changes to CNN architectures. First, the DCGAN generator and discriminator use all convolutional nets, replacing the deterministic spatial pooling functions with stridden convolutions, which allow the network to learn its own down or upsampling steps [Springenberg et al. 2015, Radford, Metz e Chintala 2015]. Second, by eliminating fully-connected layers from the top and applying global average pooling instead, it was found that the model stability could be increased, but at the cost of convergence speed. Therefore, a middle ground of directly connecting the highest convolutional features to the input and output of both generator and discriminator was applied and presented satisfactory results. Besides, the first DCGAN layer, which takes a uniform distribution as input, is reshaped into a 4-dimensional tensor to be used as a convolution-stack start point. The last convolutional layer of the discriminator is flattened and feeds a single sigmoid function to provide the output.

The third change concerns avoiding the application of batch normalization in the generator's output and discriminator's input layers. The batch normalization helps to deal with training problems related to poor initialization and gradient flow. However, directly applying batch normalization to all layers results in oscillatory behavior and model instability [Ioffe e Szegedy 2015, Radford, Metz e Chintala 2015]. Finally, the Rectified Linear Unit (ReLU) activation was applied in the generator layers, except its output, and within the discriminator, the leaky rectified activation showed to perform very well, especially for higher resolution modeling. Figure 8.2 illustrates the generator's pipeline based on noisy input.

Despite the advantages mentioned above, we considered using DCGAN based on successful cases observed in recent medical imaging augmentation works. Diaz et al. [Diaz-Pinto et al. 2019], for instance, employed DCGAN for retinal image synthesis, providing an in-depth analysis of the qualitative results and quantitative evaluation of the structural properties of synthetic and real images. Further, Doman et al. [Doman, Konishi e Mekada 2020] proposed using DCGAN to synthesize lesion images from metastatic liver cancer. The authors stated that the method provides a significant improvement in terms of detection rating, raising from 65% to 95%. Meanwhile, Alyafi et al [Alyafi, Diaz e Marti 2020] proposed a similar work employing DCGAN to generate realistic breast mass X-ray mammographies. Finally, Anicet and Luna [Za-

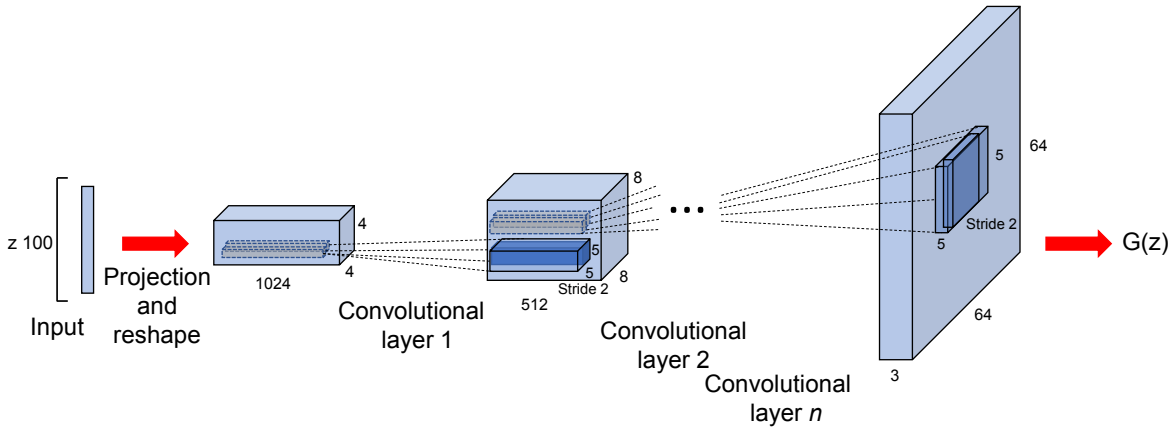


Figure 8.2: DCGAN's generator: a 100-dimensional uniform distribution z is projected to a small spatial extent convolutional representation with many feature maps. Four fractionally-stridden convolutions convert the high-level representation into a 64×64 image output. As one can observe, no fully connected or pooling layers are applied to the architecture. The discriminator is defined in an analogous fashion [Radford, Metz e Chintala 2015].

nini e Colombini 2020] introduced DCGAN to generate Parkinson's disease electromyography signals. The work introduced different frequencies and amplitudes to simulate the patient's tremor patterns. DCGAN presents several convolutional layers, no fully connected layers, and batch normalization applied to every layer of the generator (except the last one), thus turning the learning process more stable and scalable to a wide range of applications. Such an approach is able to deal better with the drawback related to time-consuming, computational power and output quality, and easy achieving convergence during the training and synthetic sample generation process [Cao et al. 2018, Radford, Metz e Chintala 2015]. The listed advances led us to employ DCGAN to BE and adenocarcinoma evaluation using GAN-based data augmentation.

While DCGAN presents the advantage of being suitable to a wide range of applications, other GAN architecutes offer different features in their concepts, such as: (i) Conditional GAN (CGAN) that adds a constraint model variable to guide the data generation process, making the convergence to a specific target faster; (ii) Wasser GAN (WGAN) that adds the weight pruning during the data synthesis, making the training process more stable, (iii) Wasser GAN with Gradient Penalty (WGAN-GP) that replaces the WGAN pruning by gradient penalties, eliminating the necessity of generator and discriminator balancing during the training min-max game, (iv) CycleGAN that introduces to the training a cycle loss and a self-constraint, providing the ability of generating random samples based on two different distributions, (v) Laplacian GAN (LAPGAN) that employs Laplacian Pyramids with upsampling, Gaussian Pyramids with downsampling and a training process based on a cascade convolutional fashion, resulting in a step-by-step independent training that learns residuals, and finally, (vi) Self-attention GAN

(SAGAN) that introduces the self-attention and spectral normalization to the training process, turning the networks more stable and faster. Even having such particularities, these architectures also present disadvantages such as slow convergence (WGAN-GP), inappropriate weight pruning (WGAN), more requirements for the data, such as tags or delineations (CGAN), not flexible resolution of output samples (CycleGAN), and necessity of being trained under supervision (LAPGAN). Finally, DCGAN still showed more application benefits due to its simplicity, stability, and large scalability.

8.3 Methodology and Proposed Method

This section presents the proposed approach and the methodology adopted to cope with Barrett's esophagus data augmentation and identification using DCGAN and CNN.

8.3.1 Proposed Method

As mentioned earlier, one of the leading contributions of this work is to cope with the small number of samples and to evaluate the robustness of data-augmented databases concerning BE and adenocarcinoma using different CNN architectures. To fulfill that purpose, we considered the DCGAN architecture for the data augmentation due to its simple implementation and high generalization, as well as two CNN architectures, i.e., LeNet-5 and AlexNet. Such architectures were considered due to their extensive usage in the literature, but any other model could also be used. Another certain point concerning the CNN architectures used in this work relates to the fact that LeNet-5 and AlexNet do not present a massive number of convolutional layers, thus turning the learning process faster.

For a comprehensive understanding of the deep networks' generalization capabilities, different batch-size evaluation experiments are proposed to investigate a reasonable model for the automatic diagnosis of BE and adenocarcinoma. Moreover, it is imperative to assess how different CNN architectures can deal with the problem addressed here and whether they can outperform handcrafted feature extraction approaches or not (Section 8.5). Figure 9.1 depicts the pipeline proposed in this work.

8.3.2 Datasets

Two High-definition white-light endoscopic datasets were used for an in-depth analysis concerning the robustness of the proposed approach. The first dataset is composed of images

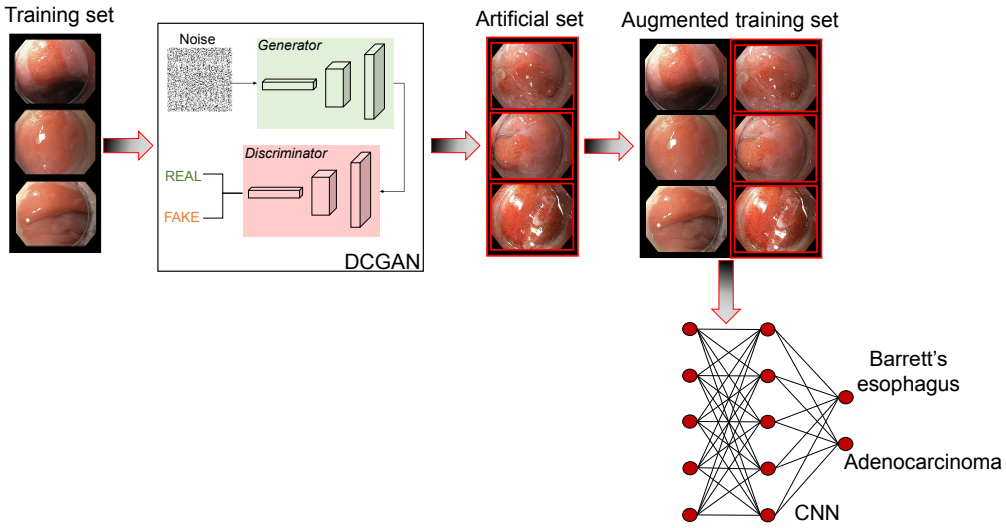


Figure 8.3: Pipeline adopted in the work. The first step is related to the synthetic sample generation using DCGAN, followed by the definition of the augmented data sets for further classification using different CNN architectures.

from a benchmark dataset provided at the "MICCAI 2016 EndoVis Challenge"[1], from now on called "MICCAI 2015" which was published aiming at encouraging researchers to conduct studies for differentiating Barrett's esophagus from cancerous images. Such dataset comprises 100 lower esophagus endoscopic images captured from 39 individuals, 22 being diagnosed with Barrett's esophagus, and 17 showing early-stage signs of esophageal adenocarcinoma. Each patient figures different endoscopic images, ranging from one to a maximum of eight. The dataset presents, in total, 50 images displaying cancerous tissue areas as well as 50 images showing only Barrett's esophagus. Five different experts have individually delineated suspicious regions observed in the cancerous images. Figure 9.2 depicts some instances positive to adenocarcinoma from MICCAI 2015 dataset and their five respective annotated delineations.

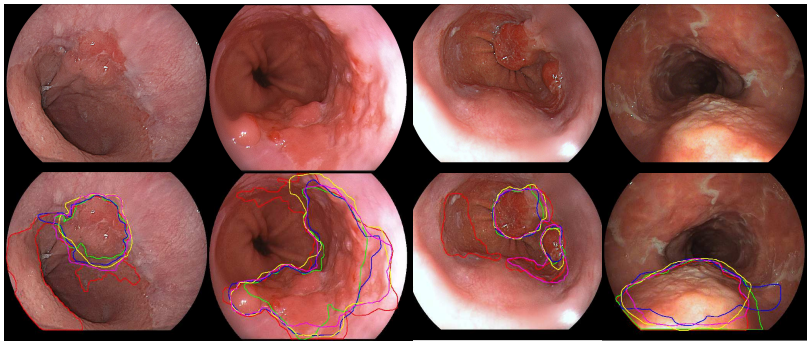


Figure 8.4: Some images from MICCAI 2015 dataset positive to adenocarcinoma and their respective delineations.

The second dataset was provided by the University Hospital Augsburg, Medizinische Klinik

[1] https://endovissub-barrett.grand-challenge.org/home/

III, Germany. The dataset is composed of 76 endoscopic images captured from different patients with BE (42 samples) and early adenocarcinoma (34 samples). One expert manually annotated the cancerous images with the final diagnosis provided by biopsy. Figure 9.3 displays some images from the Augsburg dataset labeled as positive to adenocarcinoma.

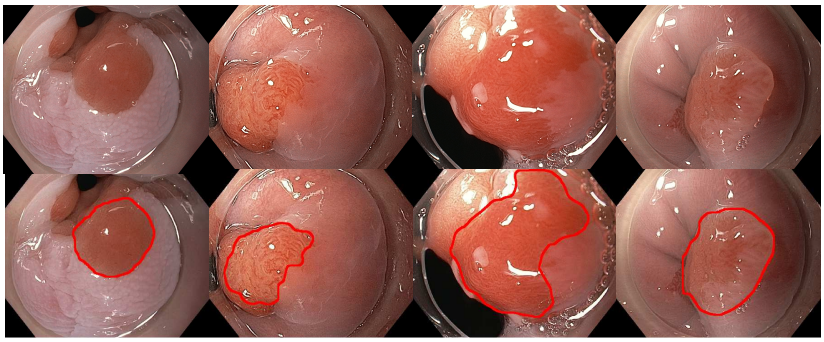


Figure 8.5: Some images from Augsburg dataset positive to adenocarcinoma and their respective delineation.

8.3.3 Experimental Delineation

In this section, we present the methodology used to conduct the data augmentation and classification steps.

8.3.3.1 Data Augmentation

To cope with the data augmentation step, two different batch sizes were evaluated, i.e., 16 and 32 samples. The reason for testing batches of size 16 concerns checking how the batch size influences the output results since the number of input samples for the image-based approach was not that significant for both datasets. The remaining DCGAN parameters were kept as the standard ones, i.e., a learning rate of 0.0001, $\beta_1 = 0.5$, $\beta_2 = 0.999$, and feature map sizes for the inner-networks[2].

The experiments were conducted over $12,000$ epochs, generating 525 and 64 synthetic samples at every $2,000$ iterations, for patch-based and full-images-based approaches, respectively. The output sample amount was related to computational limitations. We employed two different strategies to sample the output (i.e., synthetic) images: (i) the best batch (total amount of output samples at a certain iteration) during learning (i.e., the one at the smallest error), and (ii) the five best batches. These approaches are hereinafter called “very-best" and “5-best" respectively. The experimental results were assessed through a statistical analysis using the Wilcoxon

[2]The learning rate, β_1, and β_2 are parameters of the Adam optimizer [Kingma e Ba 2015] used in this work.

signed-rank test with confidence of 5% [Wilcoxon 1945].

For the sake of comparison purposes, data augmentation using rotation, zooming, and horizontal mirroring were applied to original datasets using random parameters (e.g., rotation angle and degree of zooming)[3]. Synthetic data were generated at the same proportion as GANs, so the comparison against them is the fairest possible. Last but not least, the experiments were conducted on a 96GB-memory computer equipped with an NVIDIA TitanX Graphics Card of 12 GB.

8.3.3.2 Data Classification

We employed two different strategies for the pipeline depicted in Figure 9.1: a patch- and image-based approaches. Regarding the pre-processing step, the images were split into patches whose sizes were chosen based on the work by Mendel et al. [Mendel et al. 2017]. The idea is to cover the entire image with a sliding window of 200×200 pixels and overlapping of 50 pixels in both horizontal and vertical directions. Concerning the image-based approach, the pre-processing step considered resizing the images to feed the data augmentation and the classification networks. Images from MICCAI 2015 dataset ended up with a resolution of 512×370 pixels, while the images from the Augsburg dataset figured a resolution of 512×410 pixels.

Additionally, it is worth noting that the experts' annotations were admitted for the patch labeling process. Regarding the MICCAI 2015 database, the intersection among the five delineations was assigned to the valid adenocarcinoma region. If a patch crosses the boundary between positive and negative regions, the number of cancerous and non-cancerous pixels was taken into account, and that patch is classified with the same label of the majority pixels within it. With respect to the Augsburg dataset, the single delineated area available was considered, and the same procedure for labeling patches mentioned earlier was conducted.

Regarding the patch-based approach, the entire set of patches was first considered for the data augmentation process. Further, a new augmented dataset was built to perform the classification experiments. Accordingly, 80% of all patches were randomly selected for training purposes, and the remaining 20% were used for the testing phase. Such a partitioning process was conducted for 20 runs for more robust evaluation. The image-based experiment was conducted similarly to the first one (i.e., patch-based), where the data augmentation procedure took place first. Further, a 20-fold cross-validation protocol with 80% for training and 20% for tes-

[3]The images to be mirrored were chosen at random. Apart from the number of images generated artificially, one-third accounts for rotation, one-third for mirroring, and the remaining stands for zoomed images.

ting purposes was performed to evaluate the experiments' effectiveness. Besides, the very same protocol was applied to both MICCAI 2015 and Augsburg datasets.

For classification purposes, we employed two CNN architectures pre-trained over the Imagenet dataset: LeNet-5 and AlexNet. The CNN batch sizes matched with the DCGAN ones, and $12,000$ epochs were used for learning purposes. The feature map from the last layer of each CNN (i.e., fully-connected) was reduced using the Principal Component Analysis for further classification using naive Bayes, Optimum-Path Forest [Papa e Falcão] and Support Vector Machines [Chang e Lin 2011]. Besides, we also considered the softmax layer on top of each CNN for comparison purposes. Last but not least, we compared the performance of all models with and without the augmented data, and no hyperparameter fine-tuning was employed concerning the deep networks since their default setting was employed. We used batches of size 16 for training purposes since they showed the best option so far.

8.4 Experimental Results

This section presents the experiments used to evaluate the proposed method. Considering the synthetic image selection, three different techniques were applied, i.e., the Fréschet Inception Distance (FID) [Heusel et al. 2017], the Structural Similarity Index (SSIM) [Zhou et al. 2004] and the Mean Squared Error (MSE) [Davies, Twining e Taylor 2008]. FID measure uses an Inception network pre-trained with the Imagenet dataset to calculate how similar are the original and synthetic distributions using the Fréschet distance. The lower the result, the closer the distributions, corresponding to good visual synthetic images. The SSIM metric is based on three local comparisons: luminance, contrast, and structure. The result is a measure from within the interval $[-1,1]$, being 1 a very similar synthetic image. Finally, the MSE measures the difference between the pixels of two images using the squared error. The most similar synthetic images should provide a result close to 0.

Let $\mathscr{B} = \{\mathscr{B}_1, \mathscr{B}_2, \ldots, \mathscr{B}_m\}$ be the set of batches generated during the augmentation step, and $\mathscr{I} = \{I_1, I_2, \ldots, I_n\}$ be the set of training images. For each synthetic image $\mathscr{S}_j \in \mathscr{B}_i$, $j = 1, 2, \ldots, |\mathscr{B}_i|$, we computed its closest training sample, i.e.,:

$$F_{\mathscr{B}_i}(\mathscr{S}_j, \mathscr{I}) = \arg\min_k G(\mathscr{S}_j, \mathscr{I}_k), \tag{8.2}$$

where $G(\cdot)$ stands for any similarity measure, i.e., SSIM and MSE in the manuscript. The final similarity value concerning batch $\mathscr{B}_i$ is computed as follows:

$$S(\mathscr{B}_i) = \frac{1}{|\mathscr{B}_i|} \sum_{j=1}^{|\mathscr{B}_i|} F_{\mathscr{B}_i}(\mathscr{S}_j, \mathscr{I}). \tag{8.3}$$

Regarding the classification step, we used the standard recognition rate, i.e., the ratio between the number of correct classifications and the number of dataset samples.

Additionally, a statistical evaluation using the signed-rank Wilcoxon test [Wilcoxon 1945] was considered for comparison purposes as follows:

1. For the data augmentation step, two different amounts of samples were added to the original datasets to assess the impact of new synthetic data on the classification result. We have two distinct approaches: (i) *very-best*, which means the best batch of samples, i.e., the samples with the highest similarity considering Equation 8.3 were added to the original dataset[4]; and (ii) *5-best*, which means that the final five best batches were added to the original dataset.

2. The best results (including the ones statistically similar) concerning the classification task were highlighted in bold.

3. The best overall recognition rate is highlighted with a $\star$ symbol.

8.4.1 MICCAI 2015 Dataset

In this section, we discuss the results concerning MICCAI 2015 dataset considering the augmentation and classification experiments.

8.4.1.1 Data Augmentation

The data augmentation step is evaluated qualitatively and quantitatively. The qualitative approach is based on visual insights, with the algorithm capable of generating samples realistic enough to fool a non-expert at first glance for both image and patch-based approaches. Default hyperparameters led to suitable results, as one can observe in Figures 8.6 and 8.7, which illustrate a comparison between original and synthetic samples to corroborate the quality of the synthetic images generated by DCGAN considering Barrett's esophagus and adenocarcinoma context.

[4]Notice that FID measure encodes the notion of a batch of samples, thus it has not been used the formulation provided by Equation 8.3.

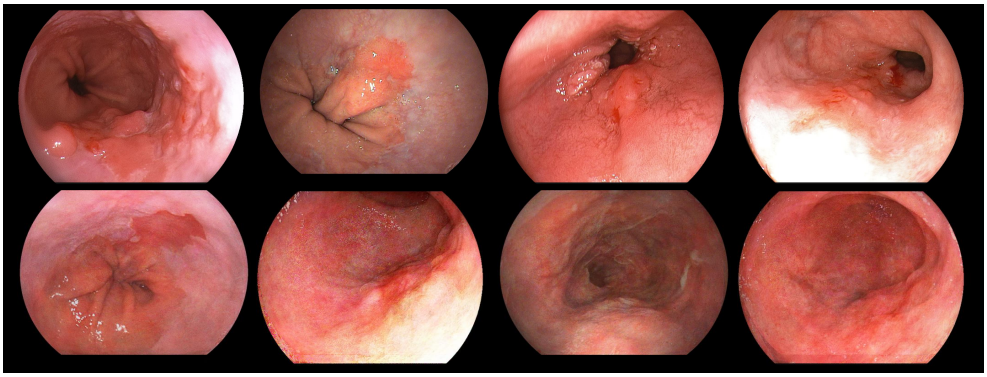


Figure 8.6: MICCAI 2015 dataset experiment using full images: original (top) and synthetic (bottom) images.

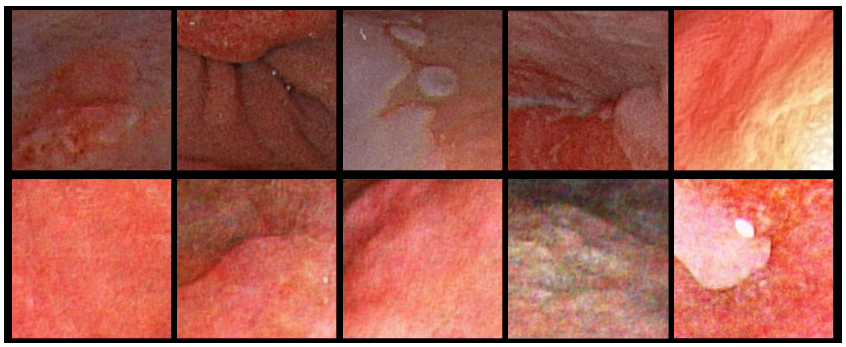


Figure 8.7: MICCAI 2015 dataset experiment using patches: original (top) and synthetic (bottom) images.

Table 8.1 presents the quantitative evaluation using FID and the similarity values considering SSIM and MSE measures using Equation 8.3, in which S_{SSIM} and S_{MSE} stand for the similarity values using Equation 8.3 considering SSIM and MSE measures. Looking carefully, one can observe that, for all measures, the best results were obtained using the patch-based approach. The same statement holds for both classes of interest, i.e., Barrett's esophagus and adenocarcinoma. Working with patches allows us considerably larger datasets, which strongly influences the GANs training step.

8.4.1.2 Classification

Table 8.2 presents the results related to MICCAI 2015 dataset. Regarding the patch-based approach, one can draw the following conclusions: (i) the augmented dataset presented the highest accuracy results for both batch sizes compared to the original dataset classification; (ii) LeNet-5 classifier showed the best classification performance with a batch size of 16 (90.04% of recognition rate with the 5-best approach for data augmentation) for BE recognition among all configurations, although being statistically similar to LeNet-5 considering the very-best ap-

Table 8.1: Quantitative experiments concerning MICCAI 2015 dataset for the very-best and 5-best approaches.

		Barrett's esophagus		**Adenocarcinoma**	
		very-best	5-best	very-best	5-best
Patches	FID	134.49	177.75	174.3	201.5
	S_{SSIM}	0.85	0.80	0.77	0.73
	S_{MSE}	199.43	243.02	244.1	265.4
Images	FID	215.14	257.43	266.6	294.5
	S_{SSIM}	0.77	0.71	0.74	0.69
	S_{MSE}	346.42	399.42	357.7	417.3

Table 8.2: Accuracy results considering MICCAI 2015 dataset.

		very-best		**5-best**		**Original Dataset (no data augmentation)**		**+ rotation translation mirroring**	
	Batch Size	LeNet-5	AlexNet	LeNet-5	AlexNet	LeNet-5	AlexNet	LeNet-5	AlexNet
Patches	16	**0.88 ± 0.04**	0.83 ± 0.03	⋆ **0.90 ± 0.03**	**0.87 ± 0.02**	0.81 ± 0.03	0.82 ± 0.03	0.83 ± 0.04	0.83 ± 0.05
	32	**0.86 ± 0.04**	0.84 ± 0.04	0.85 ± 0.03	0.86 ± 0.06				
Images	16	0.82 ± 0.07	0.82 ± 0.05	0.85 ± 0.06	0.82 ± 0.05	0.76 ± 0.03	0.82 ± 0.02	0.79 ± 0.06	0.83 ± 0.05
	32	0.81 ± 0.09	0.80 ± 0.08	0.83 ± 0.07	0.80 ± 0.07				

proach for both batch sizes (16 and 32), and; (iii) LeNet-5 provided the highest accuracy in comparison AlexNet classification over the same configuration.

Regarding the image-based approach results presented in Table 8.2, a similar outcome could be observed: (i) the augmented dataset presented the highest accuracy results for both batch sizes when compared to the original ones; (ii) LeNet-5 classifier showed the best classification result (0.8503% of recognition rate for the 5-best approach with batch size of 16) for BE recognition among all configurations, although being statistically similar to the same configuration and classifier, but with a different batch size (32); and (iii) LeNet-5 outperformed AlexNet accuracy results in almost all experiments. Clearly, one can observe that data augmentation with a patch-based approach outperformed others to a large extent considering the two neural architectures.

8.4.2 Augsburg Dataset

In this section, we discuss the results concerning the Augsburg dataset considering the augmentation and classification experiments.

8.4.2.1 Data Augmentation

In this section, the evaluation of the data augmentation step is performed. Once again, we considered both qualitative and quantitative insights, being the latter based on the same measures used for MICCAI 2015 dataset. Figures 8.8 and 8.9 provide synthetic samples for both patch-based and full image approaches concerning the Augsburg dataset, depicting the visual quality of generated images compared to the original samples. It was possible to generate, for both images and patches, realistic samples to non-expert concerns.

Figure 8.8: Augsburg dataset experiment using full images: original (top) and synthetic (bottom) images.

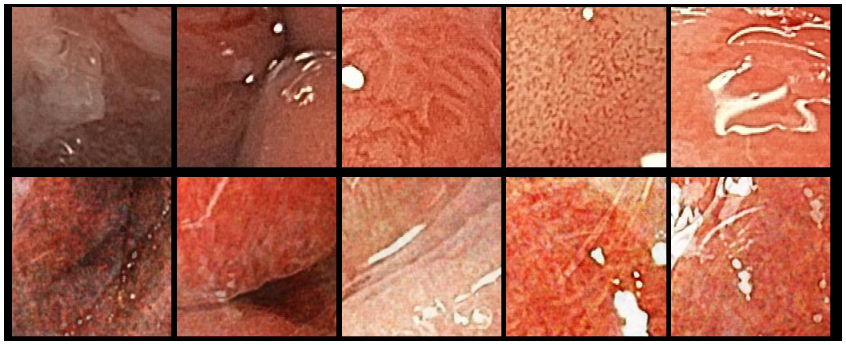


Figure 8.9: Augsburg dataset experiment using patches: original (top) and synthetic (bottom) images.

Table 8.3 presents the quantitative evaluation. Once again, the SSIM measure presented the highest similarity result for both very-best and 5-best approaches for data augmentation, similar to the one obtained for MICCAI 2015. Taking into account the results presented in Table 8.1, one can observe that the Augsburg dataset poses a more challenging task due to the average performance below the one obtained over MICCAI 2015 dataset. The Augsburg dataset figures images that are usually found in clinics and hospitals, where the specular lights contribute considerably to make the task worse.

Table 8.3: Quantitative experiments concerning Augsburg dataset for the very-best and 5-best approaches.

		Barrett's esophagus		**Adenocarcinoma**	
		very-best	5-best	very-best	5-best
Patches	FID	210.15	284.86	300.5	384.9
	S_{SSIM}	0.80	0.73	0.72	0.68
	S_{MSE}	419.25	453.62	444.2	467.8
Images	FID	300.04	329.47	341.1	376.3
	S_{SSIM}	0.70	0.65	0.67	0.63
	S_{MSE}	549.33	581.24	559.8	609.5

Table 8.4: Accuracy results considering Augsburg dataset.

		very-best		**5-best**		**Original Dataset (no data augmentation)**		**+ rotation translation mirroring**	
	Batch Size	LeNet-5	AlexNet	LeNet-5	AlexNet	LeNet-5	AlexNet	LeNet-5	AlexNet
Patches	16	0.85 ± 0.01	0.83 ± 0.02	$\star$**0.88 ± 0.05**	**0.86 ± 0.04**	0.73 ± 0.04	0.73 ± 0.05	0.75 ± 0.06	0.83 ± 0.04
	32	0.82 ± 0.03	0.81 ± 0.03	**0.87 ± 0.04**	0.84 ± 0.05				
Images	16	0.81 ± 0.06	0.80 ± 0.05	0.83 ± 0.03	0.80 ± 0.05	0.70 ± 0.05	0.72 ± 0.01	0.72 ± 0.07	0.76 ± 0.05
	32	0.77 ± 0.08	0.74 ± 0.07	0.80 ± 0.08	0.77 ± 0.09				

8.4.2.2 Classification

Table 8.4 presents the results related to the Augsburg dataset classification task. Analogously to MICCAI 2015, the Augsburg results regarding the patch-based approach were as follows: (i) the augmented dataset presented the highest accuracy results for both batch sizes when compared to the original dataset; (ii) LeNet-5 classifier also showed the best classification result (88.24% of recognition rate in the 5-best approach for data augmentation); and (iii) LeNet-5 outperformed the AlexNet classification performance on almost every experiment.

The same behavior could be observed in the image-based approach results: (i) the augmented dataset presented the highest accuracy results for both batch sizes compared to the original ones; (ii) LeNet-5 classifier presented the best classification result among all the experimental design for such datasets (83.32% of recognition rate for the 5-best approach), without any statistical similarity to the other experimental sets.

8.5 Discussions and Conclusions

In this paper, we dealt with computer-assisted Barrett's esophagus and adenocarcinoma identification by means of data augmentation and through deeply learnable features computed

using LeNet-5 and AlexNet networks. Such methods showed promising results, outperforming the previous handcrafted feature results, and suggesting inspiring progress with the application of deep learning techniques for such context. It also shows the relevance of generalization in deep learning techniques, which can improve previous results in the same field not only for the BE context but for others in which the image representation configures the context to be evaluated.

We observed that only a very few works attempted at coping with the problem of automatic BE and adenocarcinoma identification using computer vision and machine learning techniques to date, and even more, we could not observe any previous study proposing the data augmentation and validation of such a task for BE and adenocarcinoma problem. In this work, we fostered the research toward such tasks by introducing a deep learning data augmentation of BE and adenocarcinoma samples using a variant of Generative Adversarial Networks, and we showed both qualitative and quantitative evaluations of such tasks, outperforming the results obtained in some recent classification works with similar database and protocol [Souza Jr. et al. 2018, Souza Jr. et al. 2019, Souza Jr. et al. 2017, Souza Jr. et al. 2018, Souza Jr. et al. 2017]. Table 8.5 presents a more detailed comparison of the current results against a fair selection of recent state-of-the-art works in which similar protocol or dataset was employed, including the statistical evaluation among them all (bold values mean statistical similarity). Table 8.6 also presents a comparison among the proposed works and some recent state-of-the-art studies focused on the application of different datasets, protocol, and evaluation tasks.

Table 8.5: Comparison against state-of-the-art works with the application of similar evaluation protocols.

Dataset	Method	Approach	Accuracy
MICCAI 2015	Souza Jr. et al. [Souza Jr. et al. 2017]	Images	0.74 ± 0.02
	Souza Jr. et al. [Souza Jr. et al. 2019]		0.79 ± 0.04
	Souza Jr. et al. [Souza Jr. et al. 2018]		0.79 ± 0.06
	Passos et al. [Passos et al. 2019]		0.67 ± 0.10
	Proposed Method		0.85 ± 0.06
	Souza Jr. et al. [Souza Jr. et al. 2018]	Patches	0.88 ± 0.11
	Proposed Method		$\mathbf{0.90 \pm 0.03}$
Augsburg	Souza Jr. et al. [Souza Jr. et al. 2019]	Images	0.73 ± 0.05
	Souza Jr. et al. [Souza Jr. et al. 2018]		0.68 ± 0.07
	Proposed Method		0.83 ± 0.03
	Souza Jr. et al. [Souza Jr. et al. 2018]	Patches	0.86 ± 0.12
	Proposed Method		$\mathbf{0.88 \pm 0.05}$

DCGAN architecture has been recently applied for several data augmentation tasks ([Kim et al. 2019, Du et al. 2019, Doman, Konishi e Mekada 2020, Alyafi, Diaz e Marti 2020]). Considering how difficult is the task of differentiating BE and early-cancerous samples, DCGAN was selected as the approach to aid the automatic classification of BE and adenocarcinoma,

Table 8.6: Comparison against recent state-of-the-art works with the application of different protocols or datasets.

Data type	Dataset	Method	Approach	Protocol	Evaluation	Result
HD White-light endoscopy	Private	van der Sommen et al. [van der Sommen et al. 2016]	Selected Region of Interest	Leave-one-out cross validation	Sensitivity specificity	0.86 0.87
Volumetric laser endomicroscopy	Private	Swager et al. [Swager et al. 2017]	Selected Region of Interest	Leave-one-out cross validation	Area under curve	0.95
White-light endoscopy	Private	van Riel et al. [van Riel et al. 2018]	Selected Region of Interest	Leave-one-out cross validation	Area under curve	0.92
White-light endoscopy	Private	de Groof et al. [de Groof et al. 2020]	Resized images	4-fold cross validation	Accuracy, sensitivity specificity segmentation score	0.88 0.93 0.83 0.92
White-light endoscopy and Narrow-band imaging	Private	Struyvenberg et al. [Struyvenberg et al. 2020]	Zoomed images	4-fold cross validation	Accuracy, sensitivity specificity	0.84 0.88 0.78

providing promising results. Data augmentation posed as an up-and-coming tool for the improvement of BE and adenocarcinoma identification. Concerning the synthetic image evaluation, SSIM seemed to perform satisfactorily in the generated medical image field, following some state-of-the-art works [Middel, Palm e Erdt 2019, Brock, Donahue e Simonyan 2018]. The proposed approach can handle numerous bottlenecks related to the lack of data, or even rights related to medical datasets, increasing data flexibility for a wide range of medical field evaluations based on images. Images with good quality could be generated for both MICCAI and Augsburg datasets. However, the results suggested that the Augsburg dataset posed a higher challenge during synthetic image generation, which can be explained by the fewer samples available for such a dataset.

After data augmentation, the classification is performed to validate the artificial samples and how the amount of data influences the accuracy result. Concerning the MICCAI 2015 dataset, all the augmented dataset results were considerably better using the LeNet-5 architecture for both patch- and image-based approaches. When the 5-best batches were added, results close to 90% of accuracy could be achieved. The same behavior was achieved for the image-based approach, in which LeNet-5, with the 5-best methodology for data augmentation, provided the very best classification results. As a conclusion for MICCAI 2015 augmented data, the volume of data showed to be an essential parameter in the accuracy result. The more samples the dataset presented, the best the result obtained for both patch- and image-based approaches. The smallest batch sizes also presented better results for all patch- and image-based experiments, suggesting improvements when smaller batch sizes are applied.

For the Augsburg dataset, the results presented similar behavior to the previous ones. Lower classification results were obtained for all experimental designs. However, such an outcome is an expected result, once the Augsburg samples seem harder to be generalized by the model.

[Souza Jr. et al. 2018, Souza Jr. et al. 2019, Souza Jr. et al. 2017, Souza Jr. et al. 2018, Souza Jr. et al. 2017]. Despite the generalization difficulty, the results could outperform the ones obtained by previous works [Souza Jr. et al. 2018, Souza Jr. et al. 2019, Souza Jr. et al. 2017, Souza Jr. et al. 2018], highlighting not only the importance of the data augmentation but also the CNN generalization ability for BE and adenocarcinoma context. For such an experimental dataset, LeNet-5 also presented the best result and with no statistical similarity to any original dataset.

Considering the higher accuracy results obtained with the patch-based approach for all the datasets and experimental delineations, besides the higher amount of samples available for the GANs synthetic generation, such results may be justified by the following arguments:

- first, we have many more patches than full images for training.
- second, patch-based classification is “fuzzier" than the whole image approach, which can be classified into positive or negative to cancer using its full content. Concerning the patches, they are labeled as cancerous only when at least 70% of the pixels are positive to that. Such an approach is less conservative than using the entire image.
- patches consider local information only. On the other hand, using the full-image content may turn the learning process prone to errors during the approximation of the distribution of the pixels that belong to healthy and cancerous regions.

Furthermore, the comparison between GANs and transformation-based data augmentation approaches shows that the former was able to generate synthetic images better, despite that both methodologies had outperformed the experiments over the original datasets. Finally, as a bullet list of trends based on the achieved results, we shall list:

- Data augmentation presented very relevant and high-quality synthetic samples, evaluated qualitatively and quantitatively, based on visual insights and GANs evaluation techniques, such as FID, SSIM, and MSE, with SSIM presenting the best data augmentation results;
- The augmentation of BE and adenocarcinoma datasets through artificial sample addition provided encouraging results, suggesting how relevant the amount of data is in the evaluation context;
- The endoscopic image description using CNN outperformed the previous results using the handcrafted features. Although the use and generalization of CNN is still based on black-boxes, the generalization ability of such features represent an essential step to improve the classification accuracy considering BE and adenocarcinoma;

- LeNet-5 provided the best classification results for the majority of experimental settings;
- The addition of the 5-best approach for data augmentation presented the best results for all delineation sets. Despite some statistical similarity, the more the data present in the dataset, the higher the generalization obtained.

A further analysis was performed to confirm the impact of adding new synthetic samples in the classification rates using GANs. For the sake of clarity, we considered only the best results concerning the pair CNN architecture and protocol (Tables 8.2 and 8.4). From the synthetic dataset, we randomly added 25%, 50%, and 75% of the samples to the original training set. Table 8.7 presents additional experiments concerning different percentages for data augmentation. One can observe that the more synthetic samples are added for learning purposes, the higher the model's accuracy. Such finding reinforces the impact of high-quality data on the model's generalization when dealing with endoscopic imagery.

Table 8.7: Evaluating different percentages of synthetic samples for the augmented set.

Dataset	Approach	Best CNN	+25%	+50%	+75%	Full synthetic dataset
MICCAI 15	Patches	LeNet-5	0.84 ± 0.04	0.85 ± 0.05	0.87 ± 0.04	$\mathbf{0.90 \pm 0.03}$
	Images	LeNet-5	0.78 ± 0.09	0.80 ± 0.09	0.82 ± 0.07	$\mathbf{0.85 \pm 0.05}$
Augsburg	Patches	LeNet-5	0.83 ± 0.03	0.84 ± 0.05	0.86 ± 0.06	$\mathbf{0.88 \pm 0.05}$
	Images	LeNet-5	0.75 ± 0.06	0.77 ± 0.07	0.80 ± 0.05	$\mathbf{0.83 \pm 0.03}$

Despite the promising results, some drawbacks of such a study may be highlighted: the high computational cost related to the synthetic image generation can affect the DCGAN's output resolution, demanding even more powerful processing and graphical units for the generation of high-resolution images. To deal with such a problem, more powerful systems may be employed for the experimental task, or upsampling techniques can also help after the generation of small samples when the high-resolution is necessary for the context, and there is not enough computational power.

Regarding future work, we aim to consider different GANs variants for the data augmentation task and to compare the results with different GANs architectures, aiming to define which architecture fits better to aid the differentiation between BE and adenocarcinoma. Additionally, more CNN classifiers shall be applied for comparison purposes, considering deeper and more sophisticated networks. Finally, the hyperparameter optimization concerning GANs may be considered as well, thus avoiding the empirical factor of selecting such parameters.

8.6 Chapter's Considerations

After the evaluation of handcrafted features, based on several image description methods, the human knowledge could be computationally described and employed in the identification of cancerous tissues in BE images. From this Chapter on, the application of deep learning techniques to cope with the description and evaluation of early cancer in BE samples was investigated, aiming to compare its results with previously obtained ones. According to state-of-the-art works, such as the ones conducted by Mendel et al. [Mendel et al. 2017] and Ebigbo et al. [Ebigbo et al. 2020], the classification of cancerous tissues from esophageal regions can be successfully performed using deep learning techniques, so our job would be to enhance such promising results and propose enhancements to increase the classification accuracy of such a context. Moreover, it would be extremely interesting to propose designs in which the interpretation of deep learning results could be assessed.

In this first attempt, we employed the classification of early cancer using basic CNN models, AlexNet and LeNet-5. The extra contribution would be increasing the amount of samples for training and testing by creating artificial samples based on GANs. GANs are, in principle, deep architectures that, based on learning the process of generating artificial data from input samples, create new and very convincing images. After such a task, we created new experimental designs in which artificial and original samples composed the datasets that fed CNN models. As a result, we would be able to evaluate the generalization capability of AlexNet and LeNet-5 and the impact of the amount of data in the correct representation of neoplasia in BE-tissue. Several state-of-the-art works conducted the data augmentation task by introducing zooming, rotated, or translated samples to the training process, however, we assume GAN outputs can be more trustworthy in comparison to such trivial techniques of augmentation. To select the best samples of the extended datasets, we employed techniques such as MSE, FID, and SSIM, quantitatively comparing artificial and original samples.

From the obtained, our results highlight that the CNN models we selected present high potential for describing and detecting cancer in the esophagus, outperforming several of the results from previous Chapters. However, it is imperative to present that such a generalization ability is highly sensitive to the amount of available data. In general, as esophagus' examinations, personal data relates to legal information, making its acquisition a hard task to accomplish. Due to that, datasets are most of the time private and very reduced. Hence, with GAN, a more robust way of promoting a data augmentation was introduced into BE context, showing to be a promising way of generating data to cope with its lack issues, still outperforming other methods such as zooming, rotating, and mirroring.

Finally, the first attempt of representing BE and adenocarcinoma samples using deep learning techniques have shown very interesting results, pushing us to continue its evaluation, described in the next Chapters. For it, more CNN architectures will be proposed and new techniques of describing, filtering, and organizing such architectures. Also, the understanding of the CNN-learning process will be deeply explored, to be further compared to experts' learning process during early-cancer surveillance, focusing on enhancing the correct detection of cancer in BE patients.

Chapter 9

FINE-TUNING GENERATIVE ADVERSARIAL NETWORKS USING METAHEURISTICS: A CASE STUDY ON BARRETT'S ESOPHAGUS IDENTIFICATION

The further evaluation of GAN for BE and adenocarcinoma evaluation based on GAN-finetuned parameters for data augmentation and classification purposed was accepted, presented and published on the conference "Bildverarbeitung für die Medizin 2021" [Souza Jr. et al. 2021]. Considering the parameter impact in the synthetic image generation, several metaheuristics techniques were adopted in order to provide the best data-augmentation parameters for further classification purposes.

9.1 Introduction

Barrett's esophagus (BE) is a dangerous condition in which the mucosal cells of the lower part of the esophagus changes due to chronic gastrointestinal reflux, and may progress into esophageal adenocarcinoma [Lagergren e Lagergren 2010, Dent 2011, Lepage, Rachet e Jooste 2008]. Computer-aided analysis of Barrett's esophagus and adenocarcinoma figures an extremely important tool that has been subjected to intensive research in the past years [Souza Jr. et al. 2018], by the evaluation of: (i) handcrafted features of endoscopic images based on texture and color, and (ii) application of Convolutional Neural Networks (CNN) to identify the disease automatically. Recent works proposed by Souza Jr. et al. [Souza Jr. et al. 2018, Souza Jr. et al. 2019, Souza Jr. et al. 2017, Souza Jr. et al. 2018, Souza Jr. et al. 2017], Mendel et al. [Mendel et al. 2017], and Passos et al. [Passos et al. 2019] are some examples that make use of artificial

intelligence techniques for classification purposes.

A usual drawback related to computer-assisted BE and adenocarcinoma identification is the absence of data, limiting the development and validation of more effective methods to detect early-stage illness in medical samples. Recently, the data acquisition bottleneck has been coped with data augmentation (DA) methods, and among them, Generative Adversarial Networks (GAN) [Goodfellow et al. 2014] have presented significant improvements in image generation, mainly for medical imaging [Han et al. 2019]. The primary idea related to GAN is to train a generator and a discriminator simultaneously, aiming to generate convincing and high-quality synthetic images.

One of the main hindrances regarding GANs, as well as most of the modern deep learning approaches, concerns a proper selection of the architecture hyperparameters since they pose a significant influence in the model's final output. Several works addressed the problem through metaheuristic approaches to fine-tuning deep learning models' hyperparameter. In this context, Passos and Papa [Passos e Papa 2019] introduced such methods to appropriately optimizing Deep Boltzmann Machines [Salakhutdinov e Hinton 2012] hyperparameters.

Metaheuristic Optimization techniques refer to stochastic nature-inspired methods that mimic some natural behavior observed in groups of animals, social conduct, or physical events, among others, to solve complex problems. The paradigm obtained notorious popularity due to positive results in a wide variety of applications. Moreover, it does not require computing derivatives, is independent of the function landscape, and can obtain a sub-optimal solution for complex problems within a reasonably reduced computational burden.

Even though some works considered fine-tuning GAN-based hybrid models through metaheuristic approaches [Huo, Tang e Zhang 2019], as far as we know, no work proposed metaheuristic approaches to optimized GANs hyperparameter itself. Therefore, the main contributions of this work are three-fold:

- to introduce metaheuristic optimization algorithms in the context of GANs hyperparameter optimization;
- to investigate the feasibility of using GAN parameter optimization to generate high-quality synthetic images from the esophagus for further assisting the identification of Barrett's esophagus and adenocarcinoma; and
- to evaluate whether it makes sense to perform such parameter optimization of image generation for further data augmentation and classification purposes.

The remainder of this paper is presented as follows. Section 9.2 introduces the problem of GAN hyperparameter fine-tuning as an optimization problem. Further, Sections 9.3 and 9.4 describe the methodology and the experiments, respectively. Finally, Section 9.5 states the conclusions and future work.

9.2 Generative Adversarial Networks Hyperparameter Fine-Tuning as an Optimization Problem

As with most of the machine learning techniques, the training procedure of GANs demands the user an appropriate selection of the network hyperparameters, which poses a far from straightforward task due to the context-dependence and the sensitivity related to the selected values. To cope with such an issue, this work proposes employing nature-inspired metaheuristic optimization techniques to fine-tune the set of five main hyperparameters $\theta = \{\eta, \beta_1, ngf, ndf, batch\ size\}$ considering a pre-defined range, described as follows: (i) the learning rate $\eta \in [0.0001, 0.001]$, (ii) the Adam optimizer decay control $\beta_1 \in [0.002, 0.5]$, (iii) the $ngf \in [1, 128]$, which is related to the depth of feature maps carried through the generator, (iv) the $ndf \in [1, 32]$, representing the depth of the feature maps propagated through the discriminator, and (iv) the $batch\ size \in [1, 128]$.

The main idea behind metaheuristic optimization techniques consists of stochastically initializing a set of random solutions, and iteratively evolving towards the solution whose decision variables best fit a target objective, i.e., minimizing the quadratic difference between generator and discriminator losses.

The pipeline employed to perform GAN hyperparameter fine-tuning is depicted in Fig. 9.1. In a nutshell, the optimization technique selects the set of hyperparameters that minimize the loss function over the training set, considering a dataset composed of endoscopy images as well as random inputs used in the synthetic images generation process.

9.2.1 Optimization Techniques

This section briefly introduces the metaheuristic techniques employed in this paper. Notice that each technique's parameters were selected empirically based on the suggestion of their authors.

- BSA [Civicioglu 2013, Passos, Rodrigues e Papa 2019]: Backtracking Search Optimization is an evolutionary algorithm that combines stored memories with crossover and

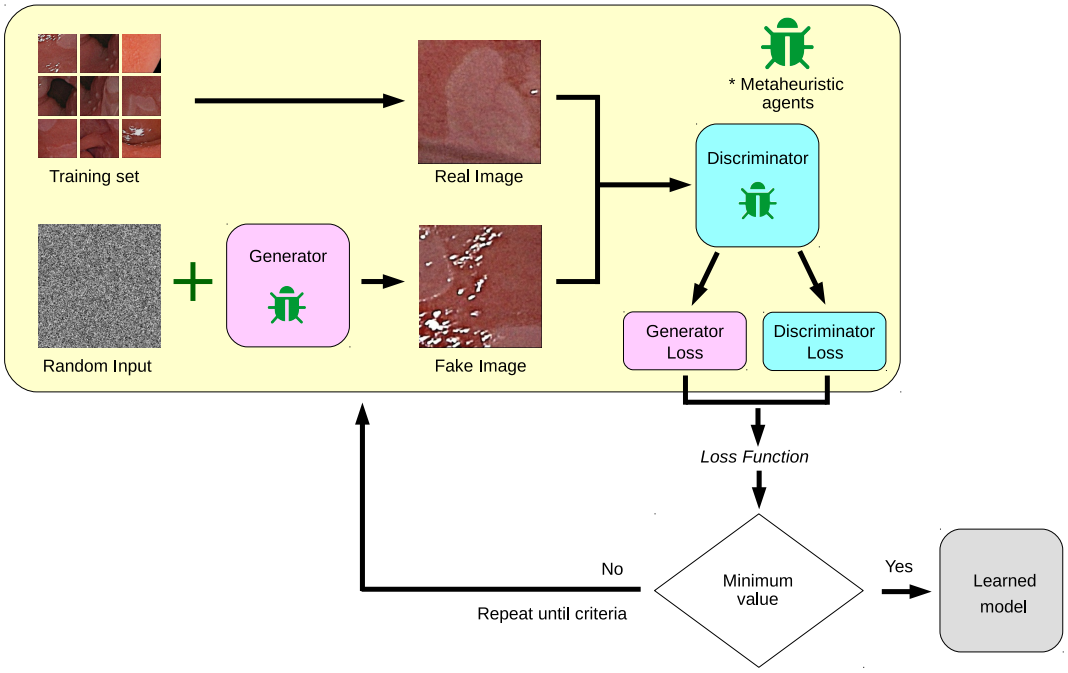


Figure 9.1: Proposed approach to encode the decision variables of each optimization agent.

mutation operations to generate new individuals.

- BSO [Shi 2011]: Brain Storm Optimization is a swarm-based optimization technique inspired by the creative human brainstorming process.
- FA [Yang 2010]: Firefly Algorithm tries to mimic the fireflies' behavior while searching for mating partners and preys.
- FPA [Yang, Karamanoglu e He 2014, Rodrigues et al. 2020]: Flower Pollination Algorithm is a swarm-based optimization method that mimics the pollination process of flowering plants.
- HS [Geem 2009]: Harmony Search models the problem of function minimization inspired on the way musicians create their songs.
- JADE [Zhang e Sanderson 2009]: a differential evolution-based algorithm that implements the "DE/current-to-p-best"mutation strategy.

9.3 Methodology

This section briefly describes the datasets employed in this work, as well as the setup employed during the experiments.

9.3.1 Datasets

Two white-light endoscopic datasets were used for an in-depth analysis concerning the robustness of the proposed approach. The first one, provided at the “MICCAI 2015 EndoVis Challenge"[1], and called “MICCAI" comprises 100 lower esophagus endoscopic images captured from 39 individuals, 22 of them being diagnosed with Barrett’s, and 17 showing early-stage signs of esophageal adenocarcinoma. Five different experts have individually delineated suspicious regions observed in the cancerous images.

The second dataset used for the experiments was provided by the Augsburg Klinikum, Medizinische Klinik III. Such a dataset is composed of 76 endoscopic images captured from different patients with BE (42 samples) and early adenocarcinoma (34 samples). The cancerous images were manually annotated by one expert from the Augsburg Klinikum. The annotations provided by the experts, for both datasets, were considered for the patch label definition.

Fig. 9.2 depicts some examples of the MICCAI dataset’s positive samples (i.e., presenting adenocarcinoma) and their five respective annotated delineations. Fig. 9.3 displays some positive images from the Augsburg dataset. In this case, we have only one delineation per image.

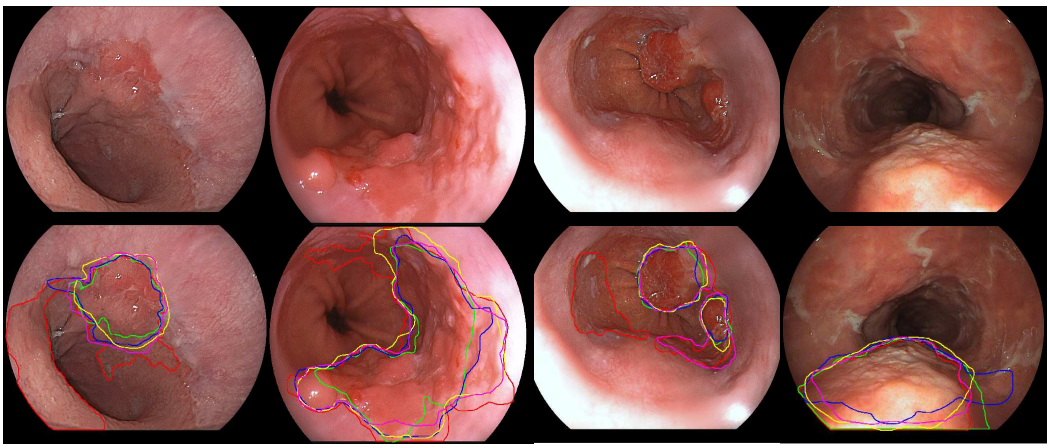


Figure 9.2: MICCAI 2015 positive samples and their respective delineations.

9.3.2 Experimental Setup

Regarding the pre-processing step, the images were split into patches [Mendel et al. 2017]. The idea is to cover the entire image with a sliding window of 200×200 pixels and overlapping of 50 pixels in horizontal and vertical directions. The label of each patch was based on the provided expert annotations of full-images.

[1] https://endovissub-barrett.grand-challenge.org/home/

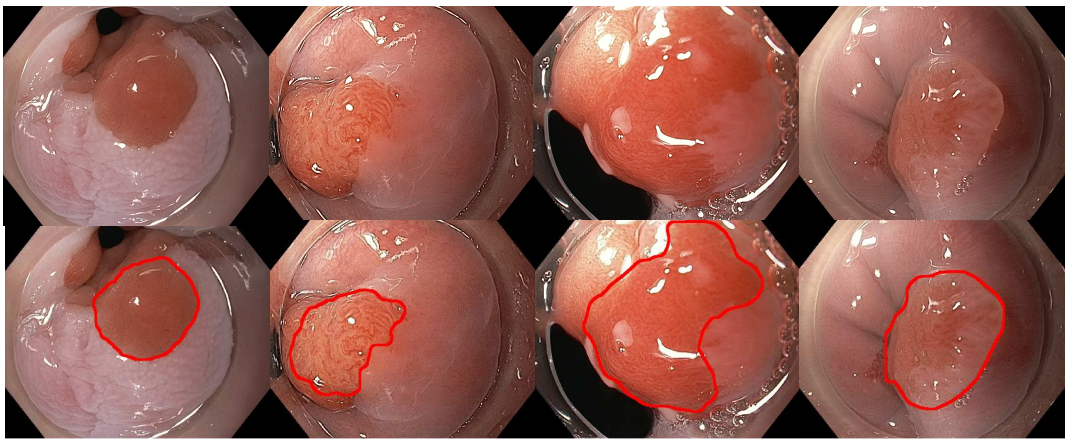


Figure 9.3: Augsburg positive samples and their respective delineations.

For reaching the best parameters for each dataset and class, the meta-heuristic techniques were run for 40 epochs. For the data augmentation evaluation, experiments were conducted over $12,000$ epochs, generating 525 synthetic samples at every $2,000$ iteration, for each sample class (AD and BE) using the best parameters obtained in the meta-heuristic experimental design. The output sample amount was related to computational limitations. We employed two different strategies to sample the output (i.e., synthetic) images: (i) the best batch (total amount of output samples at a certain iteration) during learning (i.e., the one at the smallest error), and (ii) the five best batches. These approaches are called “last" and “5-last" respectively. The experimental results were assessed through a statistical analysis using the Wilcoxon signed-rank test with confidence of 5% [Wilcoxon 1945].

The entire set of original patches was first considered for the data augmentation process. Further, a new augmented dataset was built after the optimized generation of synthetic samples to perform the classification experiments. Accordingly, 80% of the dataset was randomly selected for training purposes, and the remaining 20% was used for the testing phase. Such a partitioning was conducted over 20 runs for more robust evaluation. For the classification step, we employed two CNN architectures pre-trained with the Imagenet dataset: LeNet-5 and AlexNet. The CNNs and DCGAN experiments were run, each, for $12,000$ epochs.

9.4 Experimental Results

This section presents the experimental results for the DCGAN hyperparameters finetuning and the further classification concerning the data augmentation performed with the best meta-heuristic achieved results.

9.4.1 Optimization Results

The meta-heuristic results for the DCGAN hyperparameters finetuning can be observed in Table 9.1. As one can observe, the best results for MICCAI dataset (the values closest to 0) were obtained using BSA and FA, for AD and BE diagnosed patches, respectively, with values of 0.0033 ± 0.0048 and 0.0010 ± 0.0024. Regarding the Augsburg meta-heuristic finetuning, the best results were achieved, respectively, for AD and BE, using FA and BSA, with values of 0.0025 ± 0.0047 and 0.0011 ± 0.0022. Output samples after the synthetic generation of patches using the best parameter results can be observed in Figure 9.4.

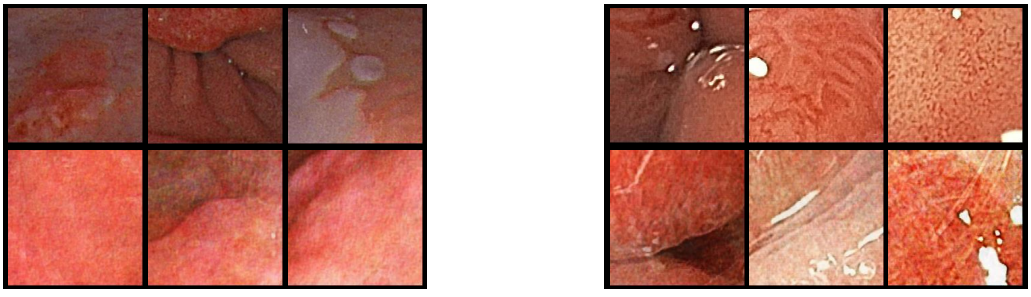
Table 9.1: Loss value and time consumption considering MICCAI and Augsburg datasets.

Dataset	Diagnosis	Metric	BSA	BSO	FA	FPA	JADE	RANDOM
MICCAI	AD	Loss	**0.0033 ± 0.0048**	**0.0056 ± 0.0102**	**0.0046 ± 0.0079**	**0.0037 ± 0.0066**	**0.0057 ± 0.0121**	**0.0310 ± 0.0079**
		Time (h)	2.9238 ± 0.5914	2.7872 ± 1.1002	3.7210 ± 1.1564	3.7254 ± 1.6175	3.4264 ± 1.2135	1.0706 ± 0.4704
	BE	Loss	**0.0045 ± 0.0118**	**0.0029 ± 0.0068**	**0.0010 ± 0.0024**	**0.0011 ± 0.0024**	**0.0018 ± 0.0028**	**0.0034 ± 0.0113**
		Time (h)	13.6447 ± 3.0530	15.1997 ± 5.5386	16.1333 ± 6.9553	14.6826 ± 5.5267	10.8849 ± 2.7443	4.8940 ± 1.3502
Augsburg	AD	Loss	**0.0049 ± 0.0132**	**0.0074 ± 0.0109**	**0.0025 ± 0.0047**	**0.0140 ± 0.0307**	**0.0053 ± 0.0146**	**0.0129 ± 0.0399**
		Time (h)	2.8345 ± 1.4249	2.5835 ± 1.3703	2.1745 ± 1.1661	4.0850 ± 2.8337	2.4363 ± 0.7424	0.6976 ± 0.2499
	BE	Loss	**0.0011 ± 0.0022**	**0.0045 ± 0.0087**	**0.0036 ± 0.0102**	**0.0057 ± 0.0136**	**0.0105 ± 0.0316**	**0.0051 ± 0.0119**
		Time (h)	8.6632 ± 4.3097	8.7128 ± 5.8866	8.7699 ± 1.8434	9.9796 ± 4.6522	9.3854 ± 7.3977	2.6029 ± 0.7025

Additionally, Fig. 9.5 depicts the convergence average and standard deviation concerning each technique. Notice that even though obtained minimal values of loss, it presents high values of standard deviation.

9.4.2 Classification Results

Regarding the classification performed after the data augmentation step, one can observe the results in Table 9.2. We highlight that for MICCAI dataset, the best classification rates



(a) MICCAI (b) Augsburg

Figure 9.4: MICCAI (a) and inhouse (b) dataset experiments using patches: original (top) and synthetic (bottom) images.

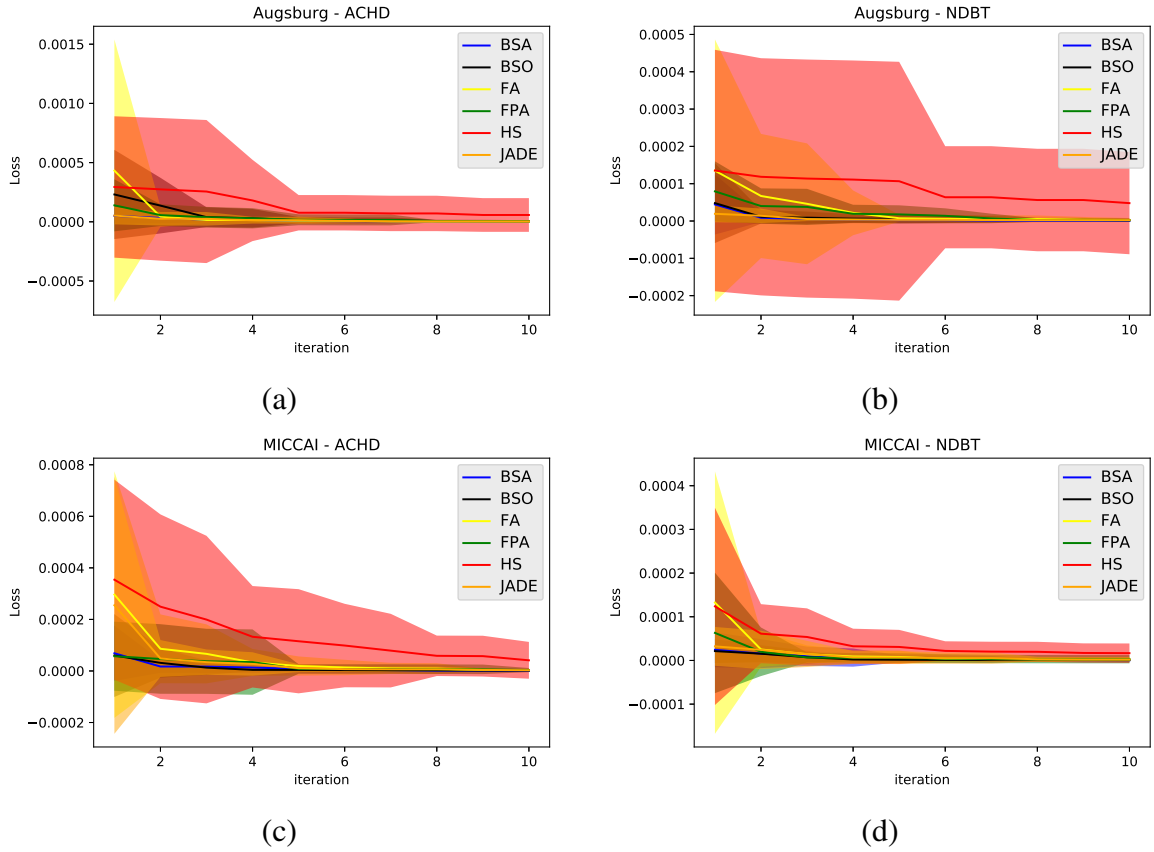


Figure 9.5: Average optimization convergence considering Augsburg dataset over (a) AD and (b) BE, and MICCAI dataset over (c) AD and (d) BE.

Table 9.2: Accuracy results considering MICCAI and Augsburg datasets.

			No augmentation		Standard augmentation		last GAN-augmentation		5-last GAN-augmentation	
Dataset	**AD Parameters**	**BE Parameters**	**LeNet-5**	**AlexNet**	**LeNet-5**	**AlexNet**	**LeNet-5**	**AlexNet**	**LeNet-5**	**AlexNet**
MICCAI	BSA	FA	0.81 ± 0.03	0.82 ± 0.03	0.83 ± 0.04	0.83 ± 0.05	0.89 ± 0.07	0.88 ± 0.07	$\star$ **0.93** $\pm$ **0.04**	**0.91** $\pm$ **0.03**
Augsburg	FA	BSA	0.73 ± 0.04	0.73 ± 0.05	0.75 ± 0.06	0.83 ± 0.04	0.88 ± 0.09	0.87 ± 0.08	0.90 ± 0.11	0.86 ± 0.06

were obtained using BSA and FA data augmentation for AD and BE patches, respectively, with a value of 0.9311 ± 0.0043 using LeNet-5 architecture and “5-last"augmentation protocol. Regarding the Augsburg classification after performing the data augmentation, the best results were achieved, respectively, for AD and BE, using FA and BSA parameters, with an accuracy value of 0.8972 ± 0.1145 also using LeNet-5 and “5-last"augmentation protocol. The wilcoxon statistical test showed up that MICCAI classification results for both LeNet-5 and AlexNet architectures presented statistical similarity, and the same was not obtained with the remaining experimental results. The augmented dataset results using DCGAN finetuned parameter samples outperformed the ones obtained without data augmentation and the ones using augmentation based on rotation, mirroring, and zooming processes.

9.5 Discussion and Conclusions

In this paper, we mainly dealt with computer-assisted Barrett's esophagus and adenocarcinoma identification by means of DCGAN-finetuned data augmentation and through deeply learnable features computed using LeNet-5 and AlexNet networks. The DCGAN hyper-parameter finetuning showed promising classification results after the data augmentation step, outperforming the classification rates of original datasets and datasets augmented with rotated, mirrored, and zoomed samples. These results inspire progress with the application of deep learning techniques for such a context. Related to the search for the best GAN-hyperparameters, we observed that FA and BSA showed the best results for both datasets, suggesting the best performance for BE and adenocarcinoma context concerning the generation of high-quality and trustworthy samples for classification purposes. Along with the introduction of data augmentation based on DCGAN-finetuned hyperparameters, we fostered in this work the validation of DCGAN-image quality by means of classification rates. As one can observe, the final accuracy value presents improvements compared to the previous proposed works, suggesting the high importance of enough data in the training and validation of models for BE and adenocarcinoma context evaluation. In addition, it is highlighted not only the importance of the data augmentation but also the CNN generalization ability to deal with BE and adenocarcinoma distinction problem. Statistically similar, both LeNet-5 and AlexNet presented the higher accuracy results, outperforming the literature results, and also the remaining experimental delineations. Furthermore, as future work, we propose the evaluation of DCGAN-finetuned hyperparameters in the generation of full-image samples, aiming to reinforce the impact of the best hyperparameters selection in the synthetic image generation quality.

9.6 Chapter's Considerations

In a direct continuation of the study proposed in Chapter 8, this work aimed at promoting an optimal GAN configuration for generating artificial samples to compose new datasets, further classified using CNN architectures. As long as metaheuristics were assessed in Chapter 7, here, we employed the same metaheuristics (in addition to JADE technique) to promote the optimal GAN hyper-parameters in the artificial set generation.

For this Chapter, the spatial information was deeply assessed and described using a patch-based protocol for optimal data generation, augmentation, and further classification. By conducting experiments based on patches, as performed in Chapter 5, the sample labeling relies on the majority of pixels from the same class, i.e., more cancerous or not cancerous pixels inside

the patch. Even for deeply-learnable features, it is imperative to understand if the localization of such descriptors correlates with experts' understanding of what defines a cancerous tissue, guiding the interpretation of the problem by ensuring that humans and computers describe the tissues with the same meaning or not.

Once again, this work's results highlight the importance of data, spatial information properly described, and fine-tuning deep architectures in the correct classification of cancerous samples regarding BE context. The best benchmark configuration was conducted over the augmented dataset, using fine-tuned GAN model for both MICCAI and Augsburg datasets. In comparison to baseline data generation (without fine-tuning using metaheuristics), we realized that such an optimization using metaheuristics could significantly increase the fidelity of artificial samples in quantitative and qualitative ways.

Finally, for further experiments, we aimed at focusing on the interpretation of what deep learning models represent in the description step of training. From the results until this point, we already know the importance of deep models and spatial information for the correct classification of cancer. However, can we establish a similar line in the learning process of CNN architectures and experts? The next Chapters aim to answer such a question, while new experimental designs and deep classifiers are also introduced.

Chapter 10

Convolutional Neural Networks for the Evaluation of Cancer in Barrett's esophagus: Explainable AI to lighten up the Black-Box

To propose a quantitative interpretation of the CNN black-box nature, the study published at "Computers in Biology and Medicine" [Souza Jr. et al. 2021] coped with such a task by applying several Explainable AI techniques to highlight discriminative cancerous regions in BE diagnosed images. In addition, five different CNN architectures were employed to generate ML models to be interpreted and compared to cancerous regions' manual segmentation.

10.1 Introduction

Barrett's esophagus (BE) is a condition in which the lower part of the esophagus' mucosal cells change considerably, progressing to esophageal cancer (adenocarcinoma) in severe cases. Some risk factors may increase the number of BE-diagnosed patients, mainly in western populations [Lagergren e Lagergren 2010, Dent 2011, Lepage, Rachet e Jooste 2008]. The remission of esophageal cancer is directly related to early diagnosis, thus delivering the treatment with reduced morbidity and mortality rates, leading to complete remission after 10 years of treatment [Dent 2011, Sharma et al. 2016, Phoa et al. 2016].

Optical coherence tomography, confocal laser endomicroscopy, and chromoendoscopy have been employed to BE and adenocarcinoma screening, enabling a more accurate manual evaluation of the esophageal histology through *in vivo* experiments [Sharma et al. 2015]. However,

BE may be misdiagnosed during endoscopy due to the inability to distinguish the columnar mucosa of the proximal stomach from the metaplastic epithelium in the distal esophagus. The Seattle biopsy protocol is highly recommended for BE lesion evaluation (i.e., dysplastic tissue), suggesting the extraction of four biopsies sampled at every 1 cm. Unfortunately, such a procedure is still susceptible to failures since the evaluated samples may not be large enough for proper screening [Sharma et al. 2015, Abrams et al. 2009, van der Sommen et al. 2014].

The limitations related to interventional therapies (e.g., endoscopic resection and ablation techniques, which present a high potential for reducing the cancer risk in dysplasia-diagnosed patients) must be handled with monitoring methods and improved dysplasia state detection [Shaheen et al. 2009, Johnston et al. 2005, Overholt, Panjehpour e Halberg 2003]. Computer-aided analysis of early cancerous tissue figures as an essential tool for intensive research in the past years. The prediction of peritoneal metastasis in gastric cancer patients [Mirniaharikandehei et al. 2021] and the automated diagnosis of breast cancer in mammograms using a Convolutional Neural Network (CNN) [Tsochatzidis et al. 2021] are a few examples of recent works that make use of machine learning techniques in the context of medical imaging. The identification of early cancer in Barrett's esophagus has also been a subject of concern in the recent years [Souza Jr. et al. 2018]. Recently conducted works evaluated the use of handcrafted features of endoscopic images based on texture and color [Souza Jr. et al. 2018]. In contrast, others assessed the well-known CNN to identify the disease automatically. Recent works proposed by Souza Jr. et al. [Souza Jr. et al. 2018, Souza Jr. et al. 2019, Souza Jr. et al. 2017, Souza Jr. et al. 2018, Souza Jr. et al. 2017, Souza Jr. et al. 2020, Souza Jr. et al. 2021], Mendel et al. [Mendel et al. 2017], Ebigbo et al. [Ebigbo et al. 2020, Ebigbo et al. 2019], de Groof et al. [de Groof et al. 2020], van der Putten [van der Putten et al. 2020], Ma et al. [Ma et al. 2020], Ellis et al. [Ellis et al. 2020] and Passos et al. [Passos et al. 2019] are some examples that make use of artificial intelligence (AI) techniques for automatic diagnosis.

The works conducted by Souza Jr. et al. [Souza Jr. et al. 2018, Souza Jr. et al. 2019, Souza Jr. et al. 2017, Souza Jr. et al. 2018, Souza Jr. et al. 2017, Souza Jr. et al. 2020, Souza Jr. et al. 2021], and Passos et al. [Passos et al. 2019] assessed the use of image representation techniques to describe and classify adenocarcinoma and Barrett's esophagus regions. For such, the use Speeded-up Robust Features (SURF) [Bay et al. 2008] and Scale-Invariant Feature Transform (SIFT) [Lowe 2004], combined to Support Vector Machines (SVMs), Optimum Path-Forest (OPF) [Papa, Falcão e Suzuki 2009, Papa et al. 2012], and the infinite Restricted Boltzmann Machine (iRBM) [Peng, Gao e Li 2018], were considered to provide the injured region prediction. Mendel et al. [Mendel et al. 2017] and Ebigbo et al. [Ebigbo et al. 2020, Ebigbo et al. 2019] introduced for the first time the use of deep learning techniques to classify expert-

annotated esophagus samples presenting adenocarcinoma and Barrett's esophagus, in a real-time analysis employing a ResNet-based DeepLabV3+ approach with transfer learning, while Souza et al. [Souza Jr. et al. 2020, Souza Jr. et al. 2021] extended such a work by introducing the use of Generative Adversarial Networks to the evaluation of early adenocarcinoma detection in Barrett's esophagus context. In the study proposed by de Groof et al. [de Groof et al. 2020], a hybrid ResNet-UNet architecture was proposed for a real-time detection of early neoplasia in BE-diagnosed patients, while the works proposed by van der Putten et al. [van der Putten et al. 2020, van der Putten et al. 2020] aimed at (i) achieving the same detection based on the combination of features (principal tissue-of-interest dimension encoded with the majority of useful information, such as constrast and homogeneity) and conventional machine learning techniques applied to in-vivo volumetric laser endomicroscopy samples, and (ii) detecting and localizating esophagus' cancerous regions employing a multi-stage domain-specific pre-training technique to white-light endoscopy samples. Deep learning approaches applied to the evaluation and automatic identification of neoplasia regions in endoscopic images continue to figure an important research field.

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

Regarding the early-cancer detection in BE, the transparency is related not only to intellectual curiosity but also to risks and responsibilities intrinsic to the prediction output [Xie, Gao e Chen 2019, Cassel e Jameton 1981]. Unfortunately, the black-box nature of the deep learning techniques is still unresolved, not completely describable and presents a not trivial interpretation, leading to poorly understood decisions [Tjoa e Guan 2019]. Especially, the question of the relationship of the suspicious region for the physician and the most important regions for the computer-based decision is of interest.

Current research suggests different methods and frameworks in the computational interpretation of CNN, making the explainable artificial intelligence (XAI) a hotspot field to be followed by the ML community. The visual explanation proposed by XAI algorithms tracks the work process of deep learning techniques in visible ways, illustrating the learning process that supports its final outputs. The visual interpretation achieved by XAI techniques provides guiding posts for understanding not only correctness but errors behind the CNN learning process [Doshi-Velez e Kim 2017, Tonekaboni et al. 2019, Lapuschkin et al. 2019]. Many works have explored explainability in the medical field, such as the evaluation and analysis

of sentiment with applications in medicine proposed by Zucco et al. [Zucco et al. 2018], in which systematic methodologies were assessed to develop explainable Clinical Decision Support Systems. The evaluation of available interpretation models of cancer in chest radiography images has been proposed by Kallianos et al. [Kallianos et al. 2019], who attested the lack of effective and quantitative methods to cope with such a task. Lamy et al. [Lamy et al. 2019] proposed a qualitative interpretation and detection of breast cancer in mammographic images using a case-based reasoning approach with visual outcomes. The classification of melanoma in hypertrophic cardiomyopathy-diagnosed images was conducted by Codella et al. [Codella et al. 2018], with interpretation based on a evidence-based classification using CNN features and kNN search and comparison of non-expert and automatic classifications (baseline and proposed method) using Area Under Curve similarity metrics (0.772 and 0.874, respectively).

Even with the progress related to the interpretation of deep learning decisions, there is still a long way to go in terms of interpretability, assessment, and criteria definition (in regard of notions of “interpretability", “explainability"along with “reliability"and “trustworthiness") [Ribeiro, Singh e Guestrin 2016, Gilpin et al. 2018, Barredo-Arrieta et al. 2020].

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

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

The remainder of this work is organized as follows. Section 5.2 presents a brief theoretical background concerning XAI and the techniques used in the work, while Section 9.3 describes

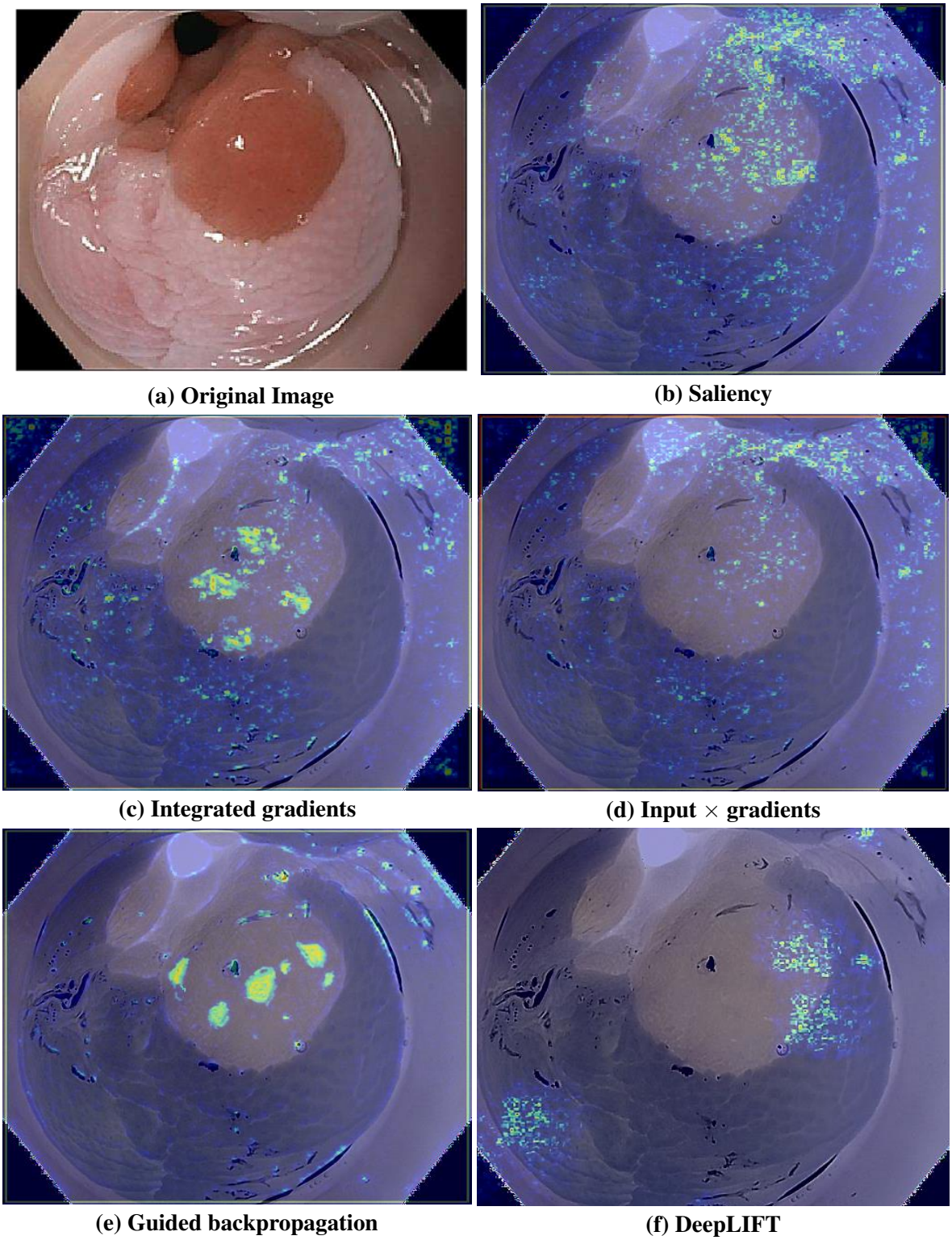
the methodology and the proposed method. Finally, Sections 9.4 and 9.5 state the experiments and results, as well as discussion and conclusions, respectively.

10.2 Explanable Artificial Intelligence

As long as autonomous machines and black-box algorithms make decisions entrusted to human knowledge, explaining themselves becomes necessary. Even considering the success in a wide range of tasks, including advertising, movie and book recommendations, medical assistance, and so on, there is in general mistrust about such black-box results. As the employment of black-box ML models has currently increased, to make important predictions in critical contexts, the demand for transparency has also raised from stakeholders in AI [Preece et al. 2018]. This may be justified by the fact that some output decisions are not clearly justifiable, legitimate, or with poor behavior details [Barredo-Arrieta et al. 2020].

The explanations behind the model's output decision are crucial in several areas that require high precision and figure experts requesting far more information from the model than just a binary prediction without extra support for such diagnosis [Tjoa e Guan 2019]. Improvements in understanding ML systems can lead to a better definition of its parameters, helping to ensure impartiality in decision-making, i.e., to detect and correctly explain the bias for training sets and tasks. Such improvements make the entire model's generalization more robust by highlighting potential adversarial and intrinsic problems that could harm prediction and evaluation. The explanation may be achieved by expressing which features are meaningful in the output inference [Barredo-Arrieta et al. 2020].

Especially in the medical domain, it is crucial that the interpretation of ML decisions is correlated to the human interpretation. Based on a previously trained ML model, the prediction and interpretation of new samples rely on their propagation through the model, with pixel impact visualizations (PIV) that may be based on layer, neuron, or prediction evaluations. This work highlights five different PIV techniques employed as tool for understanding the behavior behind CNN architecture decisions: saliency (SAL), guided backpropagation (GBP), integrated gradients (IGR), input $\times$ gradients (IXG) and DeepLIFT (DLF). Figure 10.1 illustrates the PIVs as heatmaps provided by each technique for an individual endoscopic instance.



(a) Original Image **(b) Saliency**

(c) Integrated gradients **(d) Input $\times$ gradients**

(e) Guided backpropagation **(f) DeepLIFT**

Figure 10.1: Explanable AI heatmaps based on (a) Original image: (b) saliency, (c) integrated gradients, (d) input $\times$ gradients, (e) guided backpropagation and, (f) DeepLIFT. The attributes' colors range from blue (not discriminative) to white-green (discriminative), and are related to their impact on the target's class prediction.

10.2.1 Saliency

Saliency methods, firstly proposed by Simonyan et al. [Simonyan, Vedaldi e Zisserman 2014], perform algorithm explanations by assigning values that reflect the importance of input components in their contribution to the output prediction. These values could be part of probabilities, heatmaps, or super-pixels. For such a method, having a previously trained deep model, a given class's spatial support is calculated for a classified image using a single backpropagation pass through a classification step [Simonyan, Vedaldi e Zisserman 2014].

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

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

10.2.2 Guided Backpropagation

The guided backpropagation technique [Springenberg et al. 2015] modifies the traditional backpropagation performed through the network to achieve inversion backward through a layer by zeroing negative signals from both the output or the input. To visualize the most activating image part for a specific high-level neuron (related to the highest value in the corresponding feature map), the guided backpropagation method performs a deconvolution backward pass, where negative values of either input or output are set to zero. The inversion-based method backpropagates the signal through the layers and still makes use of saliency maps for the activated signal visualization [Springenberg et al. 2015].

For deconvolution [Zeiler e Fergus 2013], the CNN data flow is inverted starting from a neuron activation in a higher layer down to the input image. Deconvolution and backpropagation mainly differ in the way the rectified linear unit (ReLU) is handeled [Springenberg et al. 2015]. As an output, a reconstructed image shows the region that presents the strongest

influence in activating a neuron [Springenberg et al. 2015].

Guided backpropagation proposes the combination of both deconvolution and backpropagation concepts, masking out only the values in which at least one is negative rather than masking out values related to negative entries of the top gradient (deconvolution) or bottom data (backpropagation). This means the corresponding neuron activation of the higher layers may be visualized, even though they present a decrease in their activation [Springenberg et al. 2015].

10.2.3 Integrated Gradients and Input $\times$ Gradients

While saliency and guided backpropagation focus on the score gradients regarding the input image, input $\times$ gradients also takes the input value into account. The attributes are computed by means of a pixel-wise multiplication of the input value and the gradient for a specific pixel [Shrikumar, Greenside e Kundaje 2017].

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

Therefore, the integrated gradients technique [Sundararajan, Taly e Yan 2017] not just compute the gradient once for each input pixel, but instead, for a fixed input image, a sequence of m intensity downscaled versions is generated applying a multiplication by $\frac{R}{m}$ $(R = 1, ..., m)$. This sequence simulates the stepwise vanishing of the signal at each pixel position. Then, the integrated gradients method sums up all the gradients related to the images of the sequence. Finally, and similar to input $\times$ gradients, this aggregated gradient is multiplied by the pixel intensity over all image positions, respectively.

10.2.4 DeepLIFT

Deep Learning Important Features (DeepLIFT), proposed by Shrikumar [Shrikumar, Greenside e Kundaje 2017], is a further development of integrated gradients. In such a method, the contributions of an input pixel to the output score are measured in relation to a reference image. This reference image should describe the unimportant background, which can be a constant zero image as a first choice. Given that reference, the output difference-from-reference value is explained in terms of the inputs difference-from-reference value. These contributions are split into positive and negative parts and backpropagated to all the neurons down to the input. Referring to a background baseline enables DeepLIFT to reveal dependencies among input pixels.

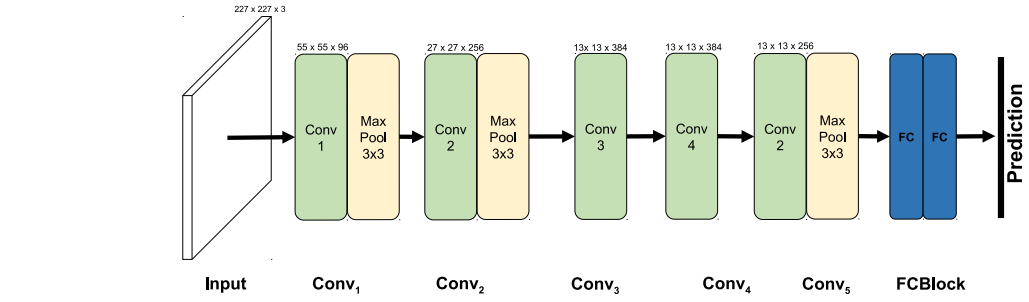
10.3 Methods and Material

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

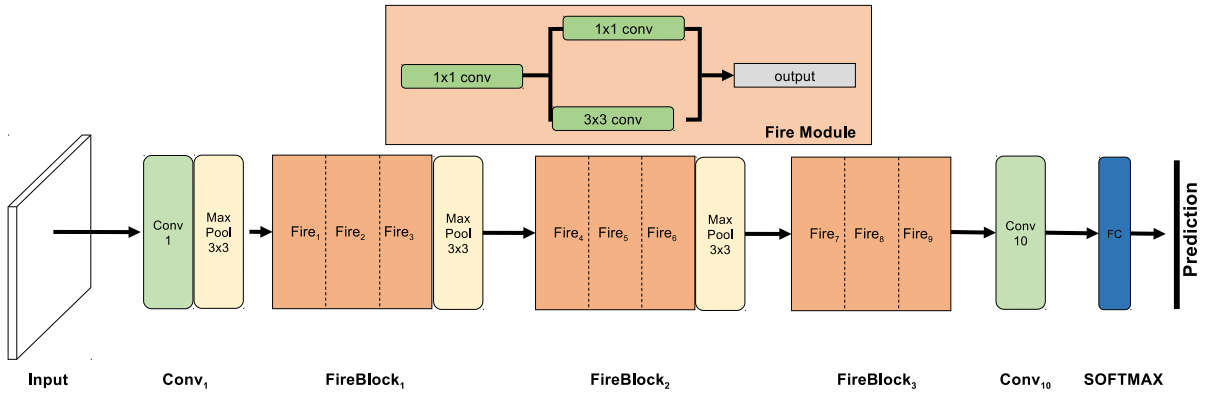
10.3.1 Method

As mentioned earlier, one of the primary contributions of this work is to provide an interpretability evaluation of positive samples in cancer diagnosed images using XAI techniques and a comparison of segmentation outputs based on the experts' annotations. To fulfill that purpose, we considered four different CNN architectures as models to be trained and validated for the XAI prediction interpretation, e.g., AlexNet, SqueezeNet, ResNet50, and VGG16, illustrated in Figure 10.2. Such architectures were considered due to their extensive usage in the literature, but any other model could also be used. With several models, a more robust and cohesive interpretation of deep networks could be delivered, aiming to understand the critical regions of BE and adenocarcinoma context. Moreover, it is imperative to assess how different CNN architectures can deal with the problem addressed in this work and whether they express or not meaningful regions for early-stage adenocarcinoma prediction during the classification step.

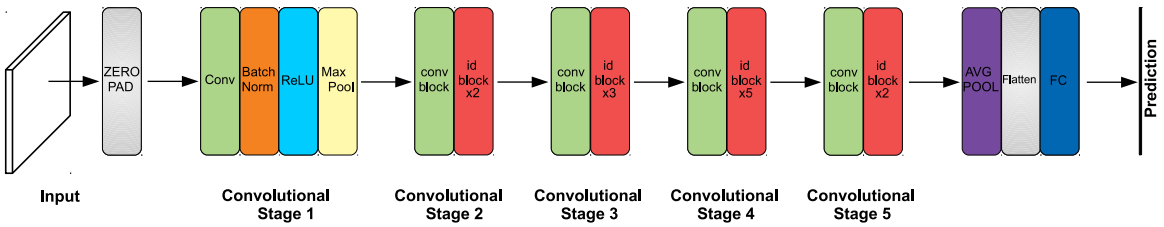
Algorithm 1 summarizes the approach proposed in this study to quantitatively evaluate the XAI techniques. The output prediction of each CNN model (after performing training and testing tasks) is provided based on two different validation protocols, i.e., the leave-one-patient-out (LOPO-CV) or the 20-fold (lines 3-5). Further, for all true positive (TP) and false negative (FN) samples (inner loop in lines 6-12), the XAI heatmaps are calculated using five different XAI techniques: saliency, guided backpropagation, integrated gradients, input $\times$ gradients, and DeepLIFT (lines 6-7). Considering that the XAI output is the pixel attribution of each evaluated sample, such attributes are normalized (line 8) for computing the OTSU threshold [Yu et al. 2010] and producing a segmentation mask to be compared with the respective experts' annotation (lines 9-10). Three agreement measures are then employed over such manual and automatic segmentations: Cohen-Kappa (CK) [McHugh 2012], intersection-over-union (IoU) and pixel accuracy (PA) (percentage of segmented pixels inside the expert's annotated area) (lines 11-12).



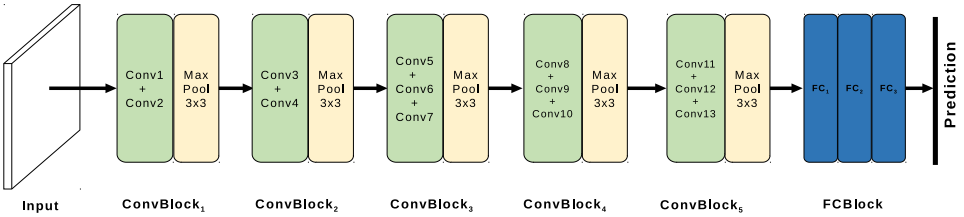
(a) AlexNet



(b) SqueezeNet



(c) ResNet50



(d) VGG16

Figure 10.2: Illustration of the selected models to perform the prediction interpretation: (a) AlexNet, (b) SqueezeNet, (c) ResNet50, and (d) VGG16.

Algorithm 1: Proposed method's algorithm.

Result: Agreement Measures: CK, IoU and PA

Initialization: protocol definition (LOPO or 20-FOLD);
repeat
  Definition of training and testing sets;
  Train CNN architectures;
  class ← Classify testing samples;
  for *img as TP and FN samples of class* **do**
    $attributes_{img}$ ← *img* interpretation using XAI techniques (saliency, gpb, igr, ixg, and dlf);
    $attributes_{nz}$ ← normalization of $attributes_{img}$;
    threshold ← OTSU thresold of $attributes_{nz}$;
    $attributes_{bz}$ ← binarization of $attributes_{nz}$ based on threshold;
    agreement_measures ← segmentation_comparison($attributes_{bz}$, $GroundTruth_{img}$), with segmentation_comparison ∈
    (Cohen Kappa, Intersection over Union and Pixel Accuracy);
  end
until *protocol's size*;

10.3.2 Datasets

Two high-definition white-light endoscopic datasets were used for performing an in-depth analysis of the proposed approach. The first dataset is composed of endoscopic examinations provided by the University Hospital Augsburg, Medizinische Klinik III, Germany. The dataset comprises a total of 76 endoscopic images captured from different BE-diagnosed patients, in which 42 present only BE and 34 present BE and early-stage adenocarcinoma. One physician manually annotated the cancerous biopsy-diagnosed images. Figure 10.3 displays some images from the Augsburg dataset labeled as positive to adenocarcinoma.

The second dataset is composed of images from a benchmark dataset available at the "MICCAI 2015 EndoVis Challenge"[1], and called "MICCAI". Such a dataset was published to encourage researchers to conduct studies for differentiating BE and early-cancerous images concerning how similar they look. Comprising 100 endoscopic images of the lower esophagus, the samples of such dataset were captured from 39 individuals. The MICCAI dataset presents 22 samples diagnosed as BE and 17 diagnosed as early-stage esophageal adenocarcinoma. Each patient figures a different amount of endoscopic images for this dataset, ranging from one to a maximum of eight. The dataset presents, in total, 50 images displaying cancerous tissue areas and 50 images showing BE disease. Five different experts have individually annotated suspicious regions in cancerous samples. Figure 10.4 depicts some samples diagnosed as positive to

[1] https://endovissub-barrett.grand-challenge.org/home/

adenocarcinoma from MICCAI dataset and their five respective experts' annotations.

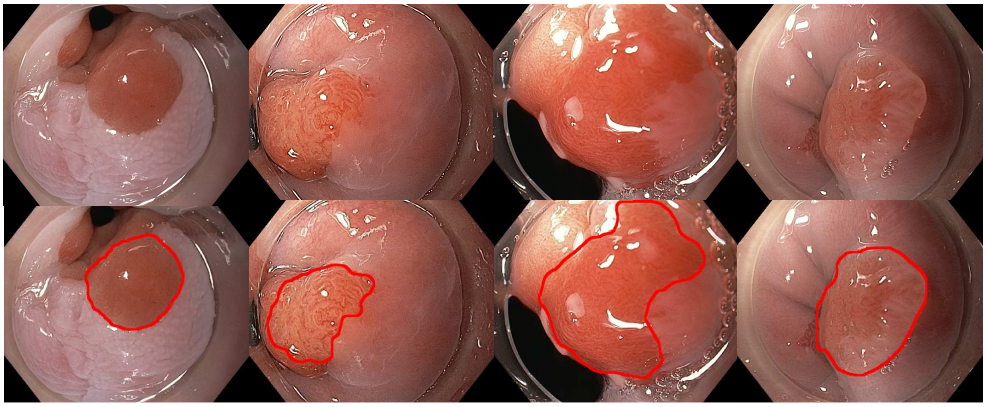


Figure 10.3: Some positive-to-adenocarcinoma images from the Augsburg dataset and their respective delineation.

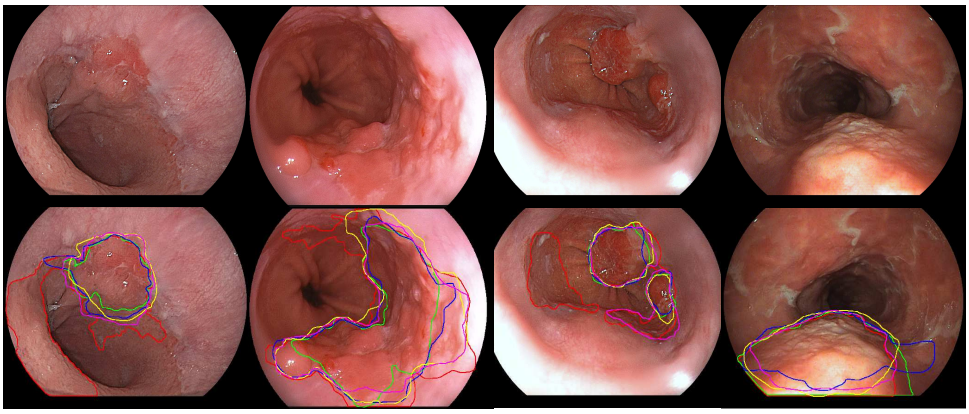


Figure 10.4: Some positive-to-adenocarcinoma images from the MICCAI dataset and their respective delineation.

10.3.3 Experimental Setup

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

10.3.3.1 Deep Model Definition

To cope with the model generation step, four different CNN architectures were evaluated, i.e., AlexNet [Krizhevsky, Sutskever e Hinton 2017], SqueezeNet [Iandola et al. 2016], ResNet50 [He et al. 2016] and VGG16 [Simonyan e Zisserman 2014]. The main rationale behind such choices are: (i) to evaluate the accuracy, sensitivity and specificity of current and trending CNN architectures in the context of BE and adenocarcinoma diagnosis and; (ii) to understand

the input discriminative parts learned by each CNN architecture and further assess their agreement with the human-expert annotation of early-cancerous regions in positive samples. All networks' parameters were kept constant, with batch size of 4, and a learning rate of 0.001. The Adam optimizer [Kingma e Ba 2015] was used with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Adam optimization has been widely employed in several classification tasks, due to its computationally efficience, little memory requirement, invariance to diagonal rescaling of the gradients, and well adaptation to problems that are large in terms of data and/or parameters. Also, such an optimization technique presents an intuitive interpretation and does not typically require intense tuning, being appropriate for non-stationary objectives and problems with very noisy and/or sparse gradients [Souza Jr. et al. 2020, Souza Jr. et al. 2021, Kingma e Ba 2015]. Therefore, we adopted standard values for all CNN's model parameters, defined empirically, considering the main scope was to assess the connection between the computational model and human interpretations in the identification of early cancer in BE samples.

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

10.3.3.2 Explanable Artificial Intelligence Evaluation

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

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

The XAI techniques present as the output the evaluation of each pixel in the input samples classified as TP or FN for the target prediction, i.e., for each pixel, attributions are calculated with values correlated to its impact in the predicted class. A zero image was used as standard

baseline to the techniques that are based on the first baseline assumption (integrated gradients and DeepLIFT). After the XAI heatmap calculation, a min-max normalization was performed for each image. However, the histograms of the pixel values for each image and XAI method are different. Therefore, and in a way to define the best binarization threshold for each sample, the OTSU threshold [Yu et al. 2010] was calculated to differentiate between meaningful and non-meaningful attributes. This binarized output can then be compared to the experts' ground truth.

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

Last but not least, the correlation between the segmentation measures and the sensitivity rates of each CNN architecture was performed employing the Spearman's Correlation Test [Lovie 1995]. With such a test, the agreement between computational and human segmentations in the correct classification of cancerous samples can be highlighted, sharply increasing the trustworthiness related to the CNN interpretation step.

10.4 Experimental Results

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

$$S = \frac{TP}{TP+FN} \cdot 100, \tag{10.1}$$

$$P = \frac{TN}{TN+FP} \cdot 100, \tag{10.2}$$

and

$$A = \frac{TP+TN}{TP+TN+FP+FN} \cdot 100, \tag{10.3}$$

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

Additionally, a statistical evaluation using the signed-rank Wilcoxon test [Wilcoxon 1945] was considered for comparison purposes between the segmentation measures over the XAI

interpretation results.

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

10.4.1 Results on Augsburg Dataset

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

10.4.1.1 Classification

Table 10.1 presents the mean classification results related to the Augsburg dataset. Regarding the 20-fold cross-validation approach, one can draw the following conclusions: (i) VGG16 model presented the highest accuracy ($83.37\% \pm 26.72\%$) and sensitivity ($80.59\% \pm 27.56\%$) results; and (ii) AlexNet classifier provided the best specificity mean result ($86.19\% \pm 22.59\%$). Concerning the LOPO-CV approach results, the following outcomes could be observed: (i) VGG16 model presented the highest accuracy ($84.37\% \pm 22.93\%$) and specificity ($86.62\% \pm 19.54\%$) results and; (ii) ResNet50 presented the best sensitivity rate ($87.05\% \pm 8.84\%$) among all configurations. The LOPO-CV results present higher accuracy, sensitivity, and specificity outcomes in most experimental cases due to the training sets for such protocols comprising more samples in a “patient-based"delineation for the classification step.

Table 10.1: Mean classification rates and time-consuming for the training task considering both 20-fold and LOPO-CV validation protocols for the Augsburg dataset. The best results for each protocol are highlighted in bold, and the best overall result for each rate is marked with a $\star$ symbol.

Protocol	Architecture	Accuracy	Sensitivity	Specificity	Time (min)
20-fold	AlexNet	0.80 ± 0.26	0.72 ± 0.29	$\mathbf{0.86 \pm 0.23}$	111.09 ± 19.13
	SqueezeNet	0.73 ± 0.38	0.67 ± 0.40	0.80 ± 0.19	175.23 ± 18.36
	ResNet50	0.76 ± 0.21	0.75 ± 0.02	0.77 ± 0.21	221.39 ± 20.33
	VGG16	$\mathbf{0.82 \pm 0.27}$	$\mathbf{0.81 \pm 0.28}$	0.84 ± 0.28	137.14 ± 15.58
LOPO-CV	AlexNet	0.83 ± 0.23	0.81 ± 0.29	0.84 ± 0.20	139.20 ± 19.25
	SqueezeNet	0.79 ± 0.11	0.73 ± 0.11	0.84 ± 0.12	201.15 ± 21.37
	ResNet50	0.84 ± 0.10	$\star\mathbf{0.87 \pm 0.09}$	0.84 ± 0.11	244.66 ± 19.17
	VGG16	$\star\mathbf{0.84 \pm 0.12}$	0.83 ± 0.14	$\star\mathbf{0.87 \pm 0.11}$	152.21 ± 12.11

10.4.1.2 XAI Interpretation

Table 10.2 presents the best results related to the XAI interpretation of the Augsburg dataset when SAL, GBP, IGR, IXG and DLF were applied to the model interpretation task, and CK, IoU, and PA measures were applied to assess the computational and human segmentation agreement[2]. In the evaluation of the 20-fold cross-validation approach, the saliency-based technique, when compared to the experts' segmentations using all three agreement measures, provided the best results for every conducted experiment, outperforming the agreement observed in all remaining XAI techniques. In addition, the agreement between computational and human interpretations for TP samples was statistically higher for all the presented results (Table 10.2) compared to the FN ones. Concerning CK, IoU, and PA measures, the best results were obtained evaluating TP inputs of ResNet50 architecture, i.e., 0.332 ± 0.023, 0.258 ± 0.023, and 0.590 ± 0.151, respectively.

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

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

10.4.2 Results on MICCAI Dataset

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

10.4.2.1 Classification

Table 10.3 presents the mean results related to the MICCAI dataset. Regarding the 20-fold cross-validation approach, one can draw the following conclusions: (i) VGG16 model presen-

[2]For the sake of clarity, we only displayed the results of the best XAI technique. The same is applied to the results presented in section 10.4.2.2

Table 10.2: CK, IoU, and PA mean values for the best XAI interpretation output of 20-fold and LOPO-CV validations over the Augsburg dataset. The best results for each protocol are highlighted in bold, and the best overall result for each masure is marked with a $\star$ symbol.

Protocol	Architecture	Prediction	Best XAI technique	CK	IoU	PA
20-fold	AlexNet	TP	SAL	0.30 ± 0.03	0.25 ± 0.03	0.59 ± 0.04
		FN	SAL	0.14 ± 0.02	0.09 ± 0.09	0.34 ± 0.10
	SqueezeNet	TP	SAL	0.33 ± 0.03	0.24 ± 0.06	0.51 ± 0.15
		FN	SAL	0.11 ± 0.03	0.08 ± 0.01	0.30 ± 0.03
	ResNet50	TP	SAL	**0.33 ± 0.02**	**$\star$0.26 ± 0.01**	**0.59 ± 0.15**
		FN	SAL	0.20 ± 0.03	0.13 ± 0.02	0.42 ± 0.06
	VGG16	TP	SAL	0.25 ± 0.07	0.19 ± 0.04	0.52 ± 0.11
		FN	SAL	0.18 ± 0.03	0.12 ± 0.01	0.28 ± 0.03
LOPO-CV	AlexNet	TP	SAL	0.27 ± 0.05	0.26 ± 0.06	0.61 ± 0.10
		FN	SAL	0.14 ± 0.06	0.15 ± 0.02	0.38 ± 0.05
	SqueezeNet	TP	SAL	0.33 ± 0.06	0.25 ± 0.04	**$\star$0.62 ± 0.07**
		FN	SAL	0.10 ± 0.04	0.08 ± 0.03	0.35 ± 0.12
	ResNet50	TP	SAL	0.32 ± 0.08	0.24 ± 0.08	0.57 ± 0.13
		FN	SAL	0.24 ± 0.07	0.15 ± 0.04	0.39 ± 0.04
	VGG16	TP	SAL	**$\star$0.36 ± 0.09**	**0.25 ± 0.06**	0.55 ± 0.12
		FN	SAL	0.20 ± 0.09	0.13 ± 0.02	0.31 ± 0.03

ted the highest accuracy ($83.73\% \pm 23.83\%$) and specificity ($85.16\% \pm 38.74\%$) results; and (ii) ResNet50 classifier provided the best sensitivity mean value ($88.77\% \pm 11.75\%$). Concerning the LOPO-CV approach results, the following outcomes could be observed: (i) ResNet50 architecture presented the highest accuracy ($86.55\% \pm 11.63\%$) and sensitivity ($88.51\% \pm 7.84\%$) results; and (ii) VGG16 classifier showed the best specificity rate ($88.95\% \pm 10.21\%$) among all configurations. The LOPO-CV results present higher accuracy, sensitivity, and specificity outcomes in most experimental cases due to the training sets for such protocols comprising more samples in a “patient-based"classification goal.

10.4.2.2 XAI Interpretation

Table 10.4 presents the best results related to the XAI interpretation of MICCAI dataset using the five XAI evaluation techniques and the segmentation comparison measures for TP and FN classified inputs of each CNN architecture. Regarding the 20-fold cross-validation approach, the saliency technique provided the best agreement results for every conducted experiment once again, outperforming the results observed in all remaining XAI techniques. It is also important to highlight that, for the TP samples, the obtained agreement between computational and human segmentations was statistically higher for all the experiments when compared to the FN ones. Concerning CK and IoU measures, the best results were obtained in the inter-

Table 10.3: Mean classification rates and time-consuming for the training task of both 20-fold and LOPO-CV validation protocols for the MICCAI dataset. The best results for each protocol are highlighted in bold, and the best overall result for each rate is marked with a $\star$ symbol.

Protocol	Architecture	Accuracy	Sensitivity	Specificity	Time (min)
20-fold	AlexNet	0.81 ± 0.17	0.82 ± 0.17	0.81 ± 0.19	119.45 ± 21.11
	SqueezeNet	0.77 ± 0.20	0.75 ± 0.15	0.78 ± 0.32	182.13 ± 22.27
	ResNet50	0.81 ± 0.14	$\star$**0.89 ± 0.12**	0.72 ± 0.20	237.58 ± 19.18
	VGG16	**0.84 ± 0.24**	0.78 ± 0.26	**0.85 ± 0.39**	144.53 ± 18.02
LOPO-CV	AlexNet	0.85 ± 0.05	0.82 ± 0.10	0.89 ± 0.09	146.33 ± 20.46
	SqueezeNet	0.88 ± 0.07	0.86 ± 0.09	0.89 ± 0.08	206.39 ± 20.44
	ResNet50	$\star$**0.87 ± 0.12**	**0.89 ± 0.08**	0.844 ± 0.113	251.05 ± 17.57
	VGG16	0.86 ± 0.05	0.85 ± 0.11	$\star$**0.89 ± 0.10**	156.17 ± 14.28

pretation of TP inputs using VGG16's model (0.311 ± 0.039 and 0.293 ± 0.014), while the best PA measure was obtained in the interpretation of TP samples predicted by the ResNet50 model (0.582 ± 0.081).

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

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

10.4.3 Correlation Test

The correlation test refers to the sensitivity value and the final interpretation presented by the use of XAI techniques. For such a task, the agreement measures and sensitivities of both TP and FN classified-and-interpreted samples were considered in the evaluation of each CNN

Table 10.4: CK, IoU and PA mean values for the best XAI interpretation output of 20-fold and LOPO-CV validations of the MICCAI dataset. The best results for each protocol are highlighted in bold, and the best overall result for each measure is marked with a $\star$ symbol.

Protocol	Architecture	Prediction	Best XAI technique	CK	IoU	PA
20-fold	AlexNet	TP	SAL	0.31 ± 0.03	0.29 ± 0.01	0.51 ± 0.12
		FN	SAL	0.23 ± 0.02	0.11 ± 0.02	0.19 ± 0.07
	SqueezeNet	TP	SAL	0.26 ± 0.02	0.20 ± 0.01	0.57 ± 0.10
		FN	SAL	0.20 ± 0.04	0.09 ± 0.04	0.19 ± 0.08
	ResNet50	TP	SAL	0.28 ± 0.02	0.22 ± 0.04	$\mathbf{0.58 \pm 0.08}$
		FN	SAL	0.13 ± 0.11	0.08 ± 0.04	0.12 ± 0.07
	VGG16	TP	SAL	$\mathbf{0.31 \pm 0.04}$	$\mathbf{0.29 \pm 0.01}$	0.51 ± 0.05
		FN	SAL	0.11 ± 0.07	0.07 ± 0.01	0.19 ± 0.05
LOPO-CV	AlexNet	TP	SAL	$\star\mathbf{0.32 \pm 0.06}$	$\star\mathbf{0.32 \pm 0.03}$	0.63 ± 0.06
		FN	SAL	0.26 ± 0.04	0.11 ± 0.03	0.24 ± 0.06
	SqueezeNet	TP	SAL	0.28 ± 0.06	0.22 ± 0.05	$\star\mathbf{0.64 \pm 0.13}$
		FN	SAL	0.26 ± 0.01	0.12 ± 0.04	0.27 ± 0.09
	ResNet50	TP	SAL	0.30 ± 0.03	0.19 ± 0.02	0.47 ± 0.14
		FN	SAL	0.18 ± 0.07	0.11 ± 0.05	0.24 ± 0.05
	VGG16	TP	GBP	0.32 ± 0.03	0.30 ± 0.02	0.56 ± 0.14
		FN	SAL	0.18 ± 0.03	0.10 ± 0.05	0.23 ± 0.10

model, taking its best protocol result into account. Table 10.5 presents the results of Spearman's correlation test for the very best results achieved for each CNN architecture in the evaluation of both datasets, with bold values meaning the highest achieved correlation between sensitivity and agreement measure.

After interpreting TP and FN samples by applying the methodology, we could observe that saliency technique was clearly more related to the experts' annotations than the others. Such a method provides the heatmap based on the calculated gradients of the target class, and for almost every experimental delineation, presented more attributes accorded to the experts' annotation region. As a result, all CK, IoU, and PA measures were higher for saliency maps, suggesting that this technique may work better than the remaining ones when dealing with observation and description of similar tissues of different natures, as BE and early-cancer are presented in the endoscopic instances. Besides, one can observe Figures 10.5 and 10.6, in which the best interpretation outputs of TP samples (from the best XAI technique) are presented for each CNN architecture. When analyzing such information, the attributes' incidence inside the physicians' delineations may be highlighted, even the techniques showing up that there were still relevant parts for the correct classification of positive samples outside of the agreement regions. Still, when comparing CK, IoU and PA measures observed for TP and FN samples in all experiments, FN-diagnosed segmentations presented lower values, corroborating the insights

Table 10.5: Spearman's correlation test among the best-obtained results of interpretation of each CNN architecture and validation protocol. The best results for each dataset are highlighted in bold, and the best overall result for each measure is marked with a $\star$ symbol.

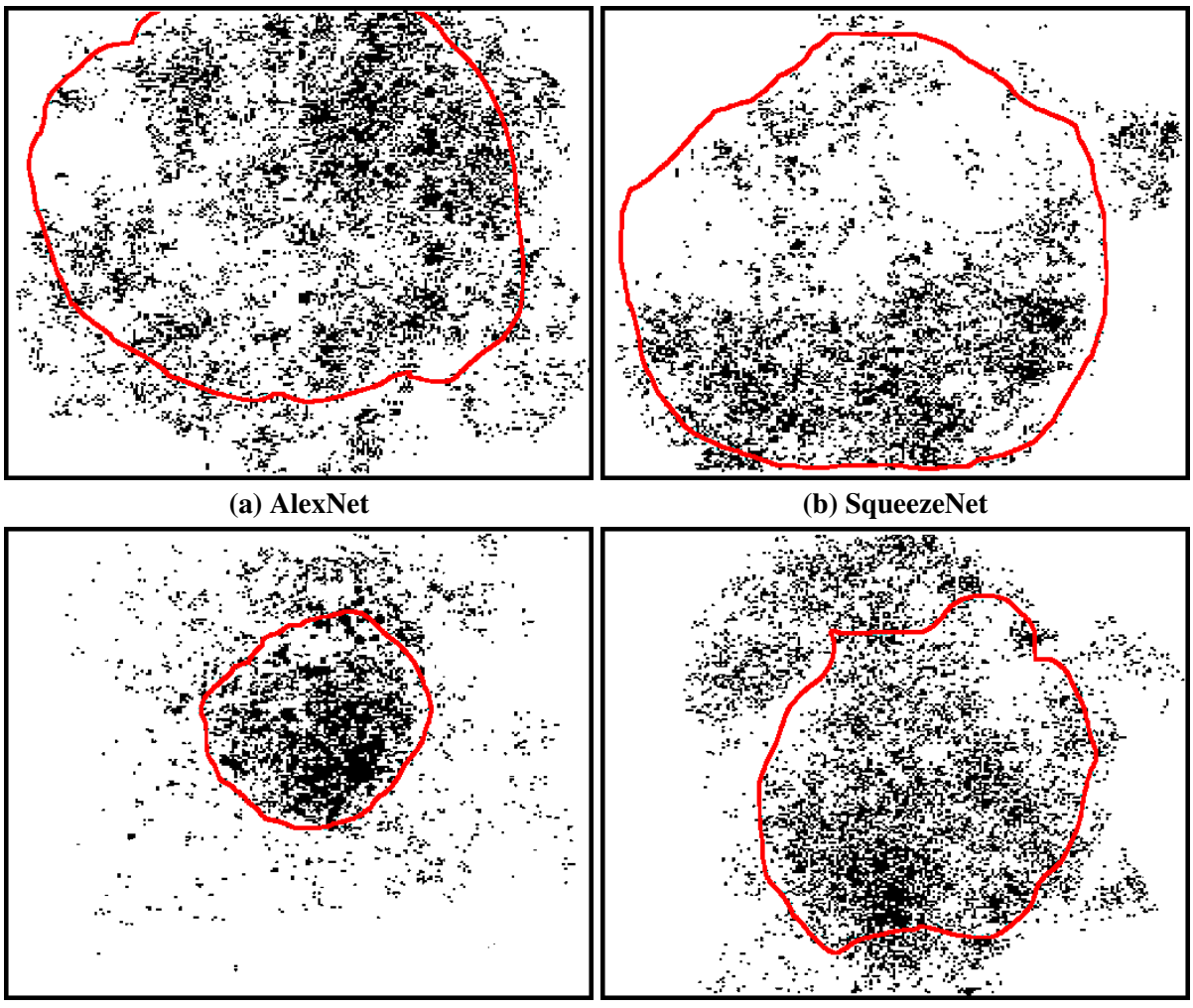
Dataset	CNN	Best Protocol	Best XAI Technique	Sperman's Correlation Test		
				Sensitivity × CK	Sensitivity × IoU	Sensitivity × PA
Augsburg	AlexNet	LOPO CV	SAL	0.53	0.51	0.62
	SqueezeNet	LOPO CV	SAL	0.48	0.52	0.57
	ResNet50	20-fold CV	SAL	**0.62**	**0.63**	0.61
	VGG16	LOPO CV	SAL	0.44	0.55	**0.63**
MICCAI	AlexNet	LOPO CV	SAL	0.51	0.61	0.71
	SqueezeNet	LOPO CV	SAL	0.41	0.66	0.68
	ResNet50	20-fold CV	SAL	0.61	$\star$**0.67**	$\star$**0.72**
	VGG16	LOPO CV	GBP	$\star$**0.63**	0.64	0.69

about the correlation among correct classification and region of interest agreement.

10.5 Discussion and Conclusions

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

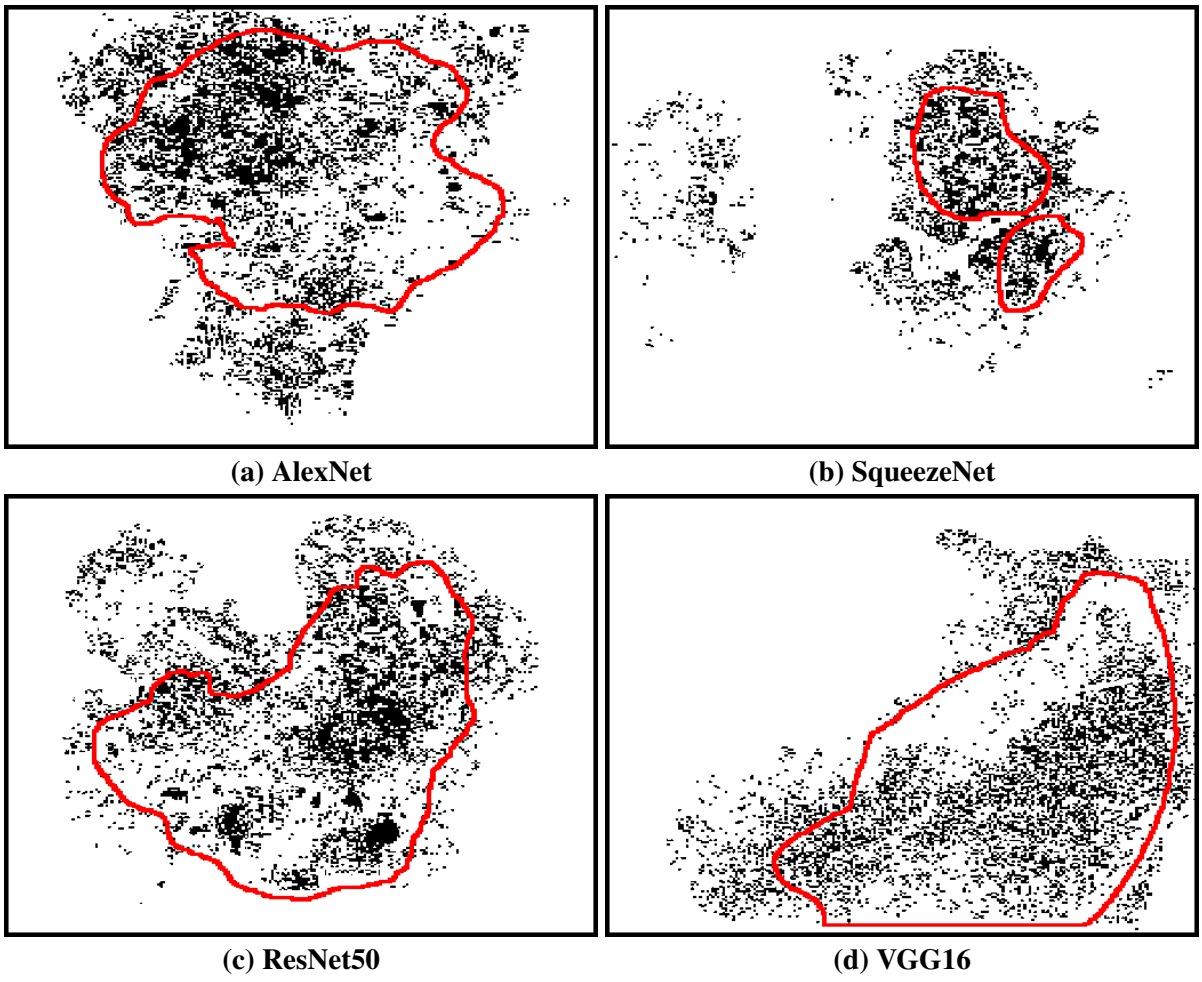
We could not observe any previous study proposing the interpretation of deep learning models in the BE and adenocarcinoma context to provide visual insights into the learning process. In this work, we fostered the research toward such tasks by introducing an interpretation of CNN classification based on XAI techniques, in both qualitative and quantitative assessment



(a) AlexNet **(b) SqueezeNet**

(c) ResNet50 **(d) VGG16**

Figure 10.5: Computational segmentation of TP samples (black) and their respective ground truth delineated area (red) over (a) AlexNet, (b) SqueezeNet, (c) ResNet50, and (d) VGG architectures for Augsburg dataset. Every segmented sample is related to the best XAI interpretation technique obtained for the respective CNN architecture.



(a) AlexNet **(b) SqueezeNet**

(c) ResNet50 **(d) VGG16**

Figure 10.6: Computational segmentation of TP samples (black) and their respective ground truth delineated area (red) over (a) AlexNet, (b) SqueezeNet, (c) ResNet50, and (d) VGG architectures for MICCAI dataset. Every segmented sample is related to the best XAI interpretation technique obtained for the respective CNN architecture.

of it. Thus, we could extend some recently proposed works over similar database and protocol [Souza Jr. et al. 2018, Souza Jr. et al. 2019, Souza Jr. et al. 2017, Souza Jr. et al. 2018, Souza Jr. et al. 2017]. Besides, we highlight three works that proposed the use of some interpretation technique to light up the learning process behing the deep model generalizaion. The first one, proposed by Gu et al. [Gu, Su e Zhao 2020], employs the use of extreme gradient boosting to predict breast cancer and case-based reasoning to explain the computational decisions. The second one, conducted by Moncada-Torres et al. [Moncada-Torres et al. 2021], designed a system for interpreting breast cancer prediction based on several ML models and Shapley Additive exPlanation. The last one from Sabol et al. [Sabol et al. 2020] showed improvements in the accountability in decision-making of colorectal cancer comparing CNN model outcomes from a novel XAI model that, besides the classification, presents: (i) a semantical explanation, (ii) a visualization of the training image most responsible for a given prediction, and (iii) a visualization of training images of other types of tissues to explain the decision. All the aforementioned works show promising interpretation of ML model decisions. Obviously, the XAI application to explain, in details, the ML decisions is crucial, but the proper quantitative evalulation is necesssary to make such an interpretation robust enough, not only relying on the visual evaluation of expert's and computational outcomes. Therefore, our method proposes an interpretation completely based on the quantitative correlation of manual and automatic explanations of the decisions, not only relying on visual insights of its outcomes.

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

The Spearman's correlation test was conducted to understand if the computational-segmented images really presented correlation to the sensitivity results for both TP and FN interpretations. As one can observe in Table 10.5, the achieved relationship for all experimental sets lay

on positive moderate to positive strong correlation among sensitivity and agreement measures (moderate is the range within ± 0.30 and ± 0.49, and strong is the range within ± 0.5 and ± 1) [Spearman Rank Correlation Coefficient 2008].

In an in-depth analysis of the correlation results, one can observe that, with positive moderate and strong correlations, as long as the number of interpreted pixels inside the experts' annotations increases for the correct classification, the sensitivity result also increases. Considering the highest correlation result (i.e., MICCAI dataset for PA agreement measure in ResNet50 interpretation), one can observe that such outcome suggests how important the agreement of computational and human region definitions is for the correct classification of positive-to-cancer samples in the evaluated context. In addition, the correlation values for TP and FN were quite far from each other. Moreover, the same behavior can be observed for the remaining correlation results. For the higher ones, the difference between TP and FN measures was also higher, showing that the sensitivity increasing may also be related to lower FN agreement within human-annotated and computational-annotated regions. The same outcome could be observed in many other correlation assessments. However, some high sensitivity values did not present strong correlation to the segmentation measures (but moderate), suggesting that not only the agreement region is important for the correct classification of positive samples, but also the attributes highlighted on the outside (see Figures 10.5 and 10.6). Furthermore, even presenting satisfactory human-and-computation agreement in the TP evaluation sets, a deeper look into such defined attributes could be performed to find, perhaps, more discriminative regions for cancerous instance sampling. The fuzzy-regions, defined as areas in which the experts' annotations do not agree in the manual definition of the ground truth, may be considered as a potential discriminative region for the attribute definition.

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

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

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

- The saliency technique, based on the interpretation of input's gradients, achieved the best results, suggesting promising behavior in the interpretation of cancerous tissues.
- The proposed hypothesis about "how related are the computational and human learnings for BE and adenocarcinoma context?"could be answered after the conducted correlation experiments, in a conclusion that, yes, the experts' annotated regions present from moderate to strong correlation to the correct classification of cancerous samples using black-boxes models, even though outside regions defined as important by such deep learning architectures also present relevance for the correct and incorrect predictions. Moreover, we conclude that the FN-classified samples always presented lower agreement of important regions with experts' annotations, corroborating the importance of such delineations in the computational learning and classification processes.

Similarly to the results obtained in this work, the study proposed by Souza Jr et al. [Souza Jr. et al. 2019] aimed at understanding the impact of the handcrafted-feature localization on the class prediction of cancerous-tissue over BE samples. For such, the authors evaluated the position and amount of features inside the cancerous region, concluding that the higher the number of features inside such a region, the higher the model's capability of correctly predicting cancerous samples. Considering the nature of object detection techniques, such as SURF and SIFT (assessed in the mentioned work), it is extremely important not only to perform the same evaluation for CNN architectures (considering the high challenge in solving the black-box learning process) but to highlight that the same interpretation could be achieved, suggesting once again the importance of defining the correct cancerous region for its correct description, learning and classification.

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

10.6 Chapter's Considerations

This Chapter continues the work related to the evaluation of deep-learnable features proposed in the previous ones, in addition to the introduction of an interpretation technique for visually understanding the behavior of deep models' feature description while generalizing BE

and adenocarcinoma context. To cope with such a task, after training deep models for the positive identification of cancer, we conducted experiments to interpret, using XAI techniques, the learning behavior behind the CNN decision, finally comparing spatial influences of such a prediction with experts' annotations of the same input samples.

First, several CNN architectures, such as VGG-16 and ResNet-50, were evaluated for the correct identification of early-cancerous tissue. After training, a deep model could be generalized, and the interpretation step would occur. Such an interpretation was conducted by applying XAI methods that, having pre-trained models, highlight important layers, neurons, or pixels according to the observation of the learning process the net performed for its final decision. As a first attempt, we focused on visually observing the learning behavior of deep architectures as a way to correlate it to the understanding that the experts obtain during surveillance.

Then, the experimental delineation was composed of training and testing CNN models, followed by the application of XAI techniques that provide the visual interpretation of the learning process, finally enabling the comparison of computational and human impactful areas of cancer-label decision. This design promotes substantial contribution, making capable the spatial observation of deep-learnable features elected as discriminative in the adenocarcinoma representation. We propose the use of four CNN architectures, AlexNet, SqueezeNet, ResNet-50, and VGG-16, for the model's training, and five XAI methods, saliency, integrated gradients, input $\times$ gradients, guided back-propagation, and deep-lift, for the interpretation of impactful pixels. To perform the correlation test, first, the annotation provided by experts and CNN architectures were compared using segmentation agreement measures, i.e., cohen kappa, intersection over union, and pixel accuracy, for further application of spearman's test.

From the obtained in this Chapter, we proposed a qualitative, quantitative, and effective technique for observing the learning behavior performed by computational methods. As we questioned in previous Chapters, do computers learn the same information the experts' do for its decision of cancer or not cancer classification regarding BE context? Obviously, the experts' observation relies on visual insights, while CNN models perform convolutional-encoding processes for describing, analyzing, and generalizing information. However, from the observed in our experiments, XAI techniques positively highlight, in combination with correlation tests, that what computers learn is directly related to what experts see for the correct identification of cancerous regions in esophagus endoscopies. The correlation test has shown moderate to strong correlation in the agreement of cancerous regions from humans' and computers' segmentations and higher cancer identification accuracy, corroborating our insights about the importance of both techniques in the correct classification we aim to deliver.

Finally, the remarkable outcomes we achieved in this Chapter guided the proposed in the next ones, in which a deeper understanding of neural networks would be proposed so the best deeply-learnable features could be selected for properly composing a method that encodes both human insights and computational learning for enhancing the classification of early cancer in BE examinations.

Chapter 11

LAYER-SELECTIVE DEEP REPRESENTATION TO IMPROVE ESOPHAGEAL CANCER CLASSIFICATION

To propose an interpretation of the CNN black-box nature based on the layer-output evaluation, the study submitted to “Nature Biomedical Engineering"coped with such a task by applying a two-step training to the classification of early cancer in Barrett’s esophagus samples.

11.1 Introduction

The last decades witnessed a world technological revolution, which established new paradigms and completely changed the way most processes were performed. Despite the computational power growth and the internet’s communication progress, artificial intelligence (AI) and machine learning-based approaches assume a protagonist role, executing tasks once considered too complex to be automated. Besides, these techniques can also perform many tasks once thought-out too dangerous, expensive, time-consuming, and annoying to be executed by humans.

Several science and knowledge fields have been overly favored by machine learning (ML) approaches, mentioning medicine amid the top ones. In this context, one can refer to applications for dementia [Zhou et al. 2020], breast mass [Ribeiro et al. 2015], and Parkinson’s disease [Passos et al. 2018] identification, among others. Despite the success of traditional ML methods, deep learning approaches recently imposed a new standard, providing paramount results in virtually any segment of medicine, e.g., exudate detection in fundus images [Khojasteh et al. 2019, Atasoy et al. 2012] and neoplasia identification in patients with Barrett’s esophagus

(BE) [de Groof et al. 2020, Hong, Park e Park 2017, Ghatwary, Zolgharni e Ye 2019, Ebigbo et al. 2020], to cite a few.

BE is a disease that attacks the lower part of the esophagus, inducing changes in the mucosa's cells. In most cases, the cause of the disease is related to obesity and smoking [Lagergren e Lagergren 2010], while the condition's remission depends on an early diagnosis [Dent 2011, Sharma et al. 2016, Phoa et al. 2016]. Moreover, late diagnosis or neglected treatment may lead to more complicated scenarios, with risks of cancer development and even death [Shaheen et al. 2009, Johnston et al. 2005].

Such a demand for BE analysis and early cancer detection associated with deep learning models' classification power flourished in several works. [Mendel et al. 2017], for instance, proposed a deep learning approach based on Convolutional Neural Networks (CNN) for BE analysis, extended by [Horie et al. 2019] and Souza Jr. et al. [Souza Jr. et al. 2021, Souza Jr. et al. 2020], who proposed similar approaches for esophageal cancer detection. Other works [Souza Jr. et al. 2018, Souza Jr. et al. 2017, Hassan e Haque 2015] aimed to use features extracted from endoscopic images for the classification of Barrett's esophagus and adenocarcinoma. Furthermore, Souza et al. [Souza Jr. et al. 2017, Souza Jr. et al. 2018, Souza Jr. et al. 2019] conducted several studies comparing different approaches for handcrafted-feature extraction in the context of BE and adenocarcinoma distinction. Aside from classification rates, many works have also stressed the importance of prediction's understanding and interpretability. Concerning the problem of early cancer detection in BE, as well as most of the medical and legal issues, transparency is not only about intellectual curiosity but also the risks and responsibilities intrinsic to the prediction's output [Xie, Gao e Chen 2019, Cassel e Jameton 1981]. Unfortunately, most of the deep learning-based techniques yield a black-box-nature approach, which result's interpretation does not denote a trivial task, and the decision process lacks meanings [Tjoa e Guan 2019]. Hence, a new trend in intelligent approaches emerged for predictions' explanation, the so-called explainable artificial intelligence (XAI) [Lamy et al. 2019] and the multistep training models [de Groof et al. 2020].

Such a requirement is essential for legal and ethics issues [Rudin 2019], healthcare and criminal justice [Hacker et al. 2020], and financial risk assessment [Ma e Lv 2019]. Regarding medical purposes, one can find XAI applications for breast cancer detection [Lamy et al. 2019], interpretable classification of Alzheimer's disease pathologies [Tang et al. 2019], and heart-attack prediction [Aghamohammadi et al. 2019], among others. Moreover, [Holzinger et al. 2017] outlined some topics related to the subject, while [Tjoa e Guan 2019] compiled several related approaches in a survey. Such XAI evaluation comprises a proper interpretation

of the learning process related to the training and generalization of deep learning models [Lamy et al. 2019]. Such interpretation may be related to pixel-wise visualization, layer-wise, and neuron-wise activation interpretations [Souza Jr. et al. 2021]. The last two approaches clearly present less intuitive understanding because they are not directly related to image regions. However, the evaluation of the training process by means of the layers and neurons' behavior introduces a barely explored area. Instead of displaying visual insights of the input over the deep model, the understanding considers a more abstract resource, i.e., the model presents a different interpretation concerning the behavior of each layer or neuron outcome.

By providing a learning explanation, a multistep approach allied to an XAI-inspired evaluation may enhance the classification rates and may help to interpret the training process once several steps may address different problems described by the same assessed context. Consequently, the multistep training field [de Groof et al. 2020] may be considered for enhancing and filtering the training behavior, once some steps may increase the samples' positive accuracy. In this context, recent works [Toth e Brath 2007, Szegedy et al. 2015, Santana et al. 2019] propose the application of multistep training methods to enhance the classification of medical instances by calibrating data and pre-fine-tuning the learning process. Regarding BE and adenocarcinoma problem, [de Groof et al. 2020] conducted experiments for detecting neoplasia with higher accuracy through a multistep training and validation system based on a hybrid deep learning model composed of ResNet and U-Net concepts.

By performing learning processes based on several levels of training, testing, and description, the interpretation of inner CNN models may be helped by local analysis encouraged by modular approaches proposed by multistep and selective methods. The work conducted by Imai et al. [Imai, Kawai e Nobuhara 2020] started the concept of knowledge growing based on the selection of pre-trained weights. The aforementioned method, named PathNet, automatically selects pre-trained modules in modular networks to improve the accuracy during the fine-tuning process. The authors achieved up to 10% of accuracy improvement over the classification of popular datasets such as CIFAR-100 and SVHN. The selective methods can also be employed in different levels of abstraction, as the work proposed by [Geifman e El-Yaniv 2017], in which a model is designed based on a risk function that selects the optimal coverage rate during the generalization process. As the learning progresses, instances are rejected to ensure optimal coverage. The authors achieved high accuracy results from the experiments conducted over CIFAR and imageNet datasets, close to 99% with 60% of test coverage. Considering such works and the current importance of assessing the model-behavior during a CNN training process based on non-trivial interpretation, and knowing that no work published to date addressed such a problem for BE and adenocarcinoma in a multistep layer-evaluation, this paper introduces three

main contributions:

- to propose a CNN-layer-wise interpretation model capable of presenting layer-fashion insights regarding the model's attention mechanisms;
- to introduce a layer-wise two-step approach to BE and adenocarcinoma context evaluation;
- to support the literature regarding both BE and multistep-training-related-researches, proposing a novel method and exposing new insights about intrinsic patterns observed over BE and adenocarcinoma endoscopic images.

The remainder of the pa per is presented as follows. Section 11.2 introduces the main concepts regarding ResNet-50 model's learning, while Section 11.3 proposes a novel architecture based on a CNN-layer evaluation. Further, Sections 11.4 and 11.6 describe the experimental results and conclusions, respectively. Last but not least, future work is also discussed.

11.2 Theoretical Background

This section presents a brief theoretical background about learning interpretation and CNN to provide a base knowledge to be followed by the proposed experimental approach of this manuscript.

11.2.1 Model's Learning Improvement

As the employment of black-box ML models has currently increased to make significant predictions in critical contexts, the demand for transparency has also increased from stakeholders in AI [Preece et al. 2018]. Such a demand regards not clearly justifiable or legitimate output decisions, also presenting details of deficient behavior of ML models [Barredo-Arrieta et al. 2020].

The understanding of the model's behavior is crucial in several areas, such as medicine, that requires high precision and requests more information from the model than just a final prediction without extra support for the diagnosis [Tjoa e Guan 2019]. Improvements in understanding ML systems lead to a better definition of their parameters, ensuring impartiality in the decision-making process. Such improvements make the entire model more robust by highlighting potential adversarial and intrinsic problems that could harm prediction [Barredo-Arrieta

et al. 2020]. Understanding how the network behaves in several aspects may also improve the correct classification of instances as long as deep networks tend to become deeper to generalize more information, making such understanding even more complicated due to the large amount of information encoded through the layers [Szegedy et al. 2017].

The explainability in the medical field has been explored in many works, such as the evaluation of cancer in chest radiography images [Zucco et al. 2018], the detection and interpretation of breast cancer in mammographic images [Lamy et al. 2019], and the classification of melanoma in hypertrophic cardiomyopathy-diagnosed images [Codella et al. 2018]. Even with the progress related to the interpretation of deep learning decisions, a long way must be overtaken in terms of interpretability, assessment, and criteria definition (in regard to notions of “interpretability", “explainability"along with “reliability"and “trustworthiness") [Ribeiro, Singh e Guestrin 2016, Gilpin et al. 2018, Barredo-Arrieta et al. 2020].

Along with the explainability, which aims at interpreting the deep model’s decisions, the proper definition of models and their parameters, as well as their deepness, are essential to compose trustworthy, robust, and precise networks. The multistep field targets the improvement of classification rates by proposing several steps in the training process, refining the parameter calculation and model’s generalization at each module [Toth e Brath 2007, Szegedy et al. 2015, Santana et al. 2019]. In this context, the generalization may decrease the network’s depth at each step once several optimization tasks are performed in the training process. Still, different problems in the same circumstances can be coped at each step in a “divide to conquer"learning generalization process [de Groof et al. 2020, Szegedy et al. 2017].

Therefore, ML models must provide, along with their predictions, a set of human-capable tools for interpreting and understanding the evaluated context, even though such decisions are strongly related to the network’s capability to express the learned information. Based on a previously trained ML model, the prediction and interpretation of new samples rely on their propagation through the model that may be based on layers, neurons, or prediction evaluations. Considering that most XAI methods for CNN interpretations are based on pixel-backpropagation-techniques, which display the model’s activation in a pixel-fashion way through input samples [Barredo-Arrieta et al. 2020], the evaluation and interpretation of layer and neuron-based methods need more related-knowledge concerning CNN architecture, deepness, and neuron-configuration. This is explained by the fact that the output of a layer-or-neuron interpretation technique is not properly a visual insight, but the behavior of a current layer, neuron or group of neurons. The learning process performed by CNN models describes a hard task to be understood once the convolutions, poolings, and activations related to such a task may “deconstruct"the in-

put in a non-linear way.

The decomposition, local-evaluation, and understanding of layers may provide a key evaluation we use in a multistep-task learning process. Such characteristics may also suggest an important approach towards evaluating the learning process besides increasing the accuracy rates. Even though some state-of-the-art works proposed to use the CNN training process as a nested-step approach to enhance the accuracy rate [de Groof et al. 2020, Ismail et al. 2019], none proposed using multistep training to evaluate any learning process besides enhancing accuracy result.

As such, this work highlights two approaches for the layer-based enhancement of a ResNet-50 model based on addition and concatenation of convolutional layer outputs, aiming at understanding the positive aspects through the learning process, later explained in Section 11.3.

11.2.2 Convolutional Neural Networks - ResNet50

Residual Networks (ResNets) [He et al. 2016] stand for a family of deep models proposed to deal with the degradation of gradients in deeper architectures. In such an approach, sets of layers are employed to fit a residual mapping $F(\boldsymbol{x})$, described as follows:

$$F(\boldsymbol{x}) = H(\boldsymbol{x}) - \boldsymbol{x}, \tag{11.1}$$

where $H(\boldsymbol{x})$ represents the underlying mapping and $\boldsymbol{x}$ denotes the input image. In a nutshell, since optimizing the residual mapping is more accessible than the original mapping $H(\boldsymbol{x})$, which is recast into $F(\boldsymbol{x}) + \boldsymbol{x}$, we can explicitly expect the stacked layers to approximate the residual function instead of the original. The residual learning is adopted for each bottleneck, i.e., a building block composed of a few stacked layers, defined as:

$$\boldsymbol{y} = F(\boldsymbol{x}, \{\mathbf{W}_i\}) + \boldsymbol{x}, \tag{11.2}$$

where $\boldsymbol{y}$ stands for the output vector and $\{W_i\}$ stands for the set filters in a block. Additionally, the function $F(\boldsymbol{x}, \{W_i\})$ formalizes the residual mapping, while the element-wise addition of $\boldsymbol{x}$ in the Equation (11.2) represents a shortcut connection. If the dimension of $\boldsymbol{x}$ and $\boldsymbol{y}$ does not fit, a projection shortcut allows dimension reduction with the help of the linear projection matrix W_s:

$$\boldsymbol{y} = F(\boldsymbol{x}, \{\mathbf{W}_i\}) + W_s\boldsymbol{x}. \tag{11.3}$$

Figure 11.1 depicts the pipeline of residual learning.

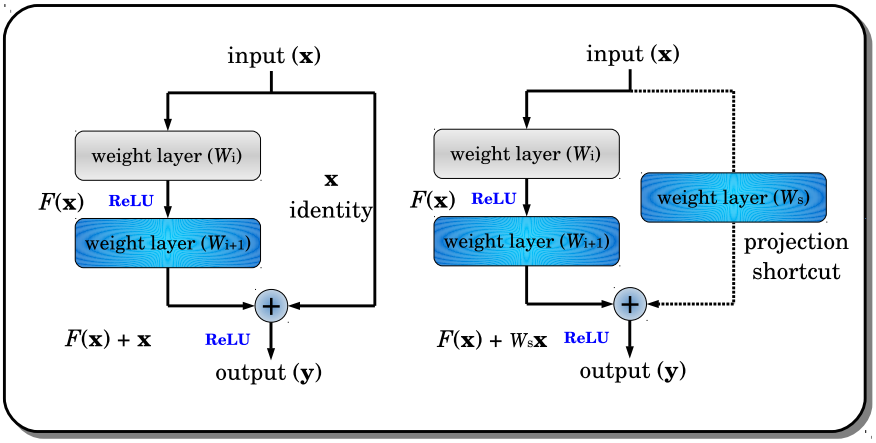


Figure 11.1: ResNet-50 bottleneck blocks: the left block denotes the identity shortcut, and the right block stands for projection shortcut. Notice the latter is considered when the dimension of x is different from the shortcut output size.

An attractive feature concerning convolutional neural networks regards the employment of a transfer learning approach, which denotes the adaptation of parameters learned for a specific task to a more general domain, capable of applying for different contexts and datasets [Yosinski et al. 2014]. Besides, the new model can be fine-tuned using the well-known backpropagation learning algorithm. In a nutshell, one can replace the base model's input and output layers by new ones from the target model and then train the network once again. Therefore, we opted to employ ResNet-50 as the base model, which was pre-trained for object detection tasks over the ImageNet 2012 [Russakovsky et al. 2015], which comprises 1.28 million images from $1,000$ classes.

Figure 11.2 depicts a ResNet-50 architecture composed of five convolutional main stages. As one can observe, some of these stages comprise a different number of identity blocks (id block) stacked together one after another. Each id block is a standard ResNet-50 block of transformations composed of two-dimensional convolutions, batch normalization, and ReLU activation processes. To name an id block, its input and output dimensions must match, requiring no transformation in the shortcut between them. On the other hand, the convolutional blocks' (conv block) input and output dimensions do not match, demanding transformations to perform the shortcut, usually employing two-dimensional convolutions and batch normalizations. Along with the entire ResNet-50 architecture, several transformations are performed in each convolutional main stage, starting with zero-padding and a 3×3 convolution, and ending up with a fully connected layer that may define the number of classes for the problem to be evaluated using a softmax activation function. The regional property is lost as long the architecture goes deeper, once the local representation is encoded into high-dimensional information.

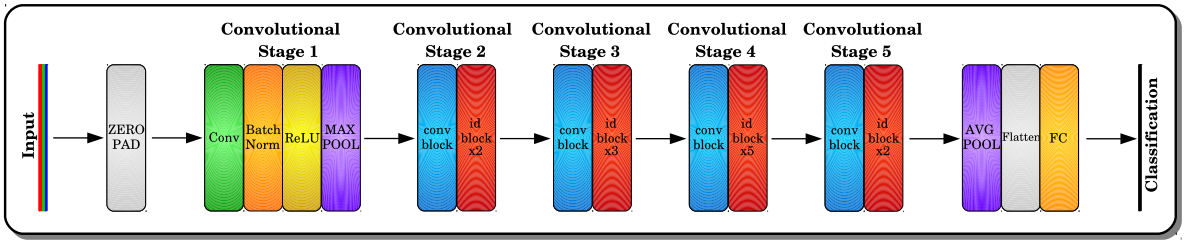


Figure 11.2: ResNet-50 architecture: the deep-neural network presents five main stages, composed of two-dimensional convolutions, batch normalizations, ReLU activations, poolings (max and average) for decreasing the output sizes, flatten (to reshape the output to one dimension), and finally, a dense layer to redefine the number of classes of the problem. Many inner transformations are applied in CONV and ID blocks, turning the architecture deeper and significantly increasing the number of parameters and filters applied over the learning process. As long as the architecture goes deeper, more regional information is encoded into high-dimensional representations, losing its local property. Identity blocks are repeated two, three, five, and two times in stages two, three, four, and five, respectively.

11.3 Methodology

This section presents the datasets, the proposed approach and experimental delineation.

11.3.1 Datasets

Two high-definition white-light endoscopic datasets were used for performing an in-depth analysis of the proposed method. The first comprises endoscopic examinations provided by the University Hospital Augsburg, Medizinische Klinik III, Germany. The dataset includes a total of 76 endoscopic images captured from different BE-diagnosed patients, in which 42 presented only BE and 34 presented BE and early-stage adenocarcinoma. One physician manually annotated the cancerous biopsy-diagnosed images. Figure 11.3 displays some images from the Augsburg dataset labeled as positive to adenocarcinoma.

The second dataset is composed of images from a benchmark dataset available at the “MICCAI 2015 EndoVis Challenge"[1], hereinafter called “MICCAI", published to encourage researchers to conduct studies for differentiating BE and early-cancerous images. Comprising 100 endoscopic images of the lower esophagus, the samples of such dataset were captured from 39 individuals. The MICCAI dataset presents 22 samples diagnosed as BE and 17 diagnosed as early-stage esophageal adenocarcinoma. Each patient figures a different amount of endoscopic images for this dataset, ranging from one to a maximum of eight. The dataset presents, in total, 50 images displaying cancerous tissue areas and 50 images showing BE disease. Five different experts have individually annotated suspicious regions in cancerous samples. Figure 11.4 de-

[1] https://endovissub-barrett.grand-challenge.org/home/

picts some instances diagnosed as positive to adenocarcinoma from MICCAI dataset and their five respective experts' annotations. For all images of both datasets the ground truth class was confirmed by histopathologic evaluation of a biopsy. The manual delineations are not used for our approach.

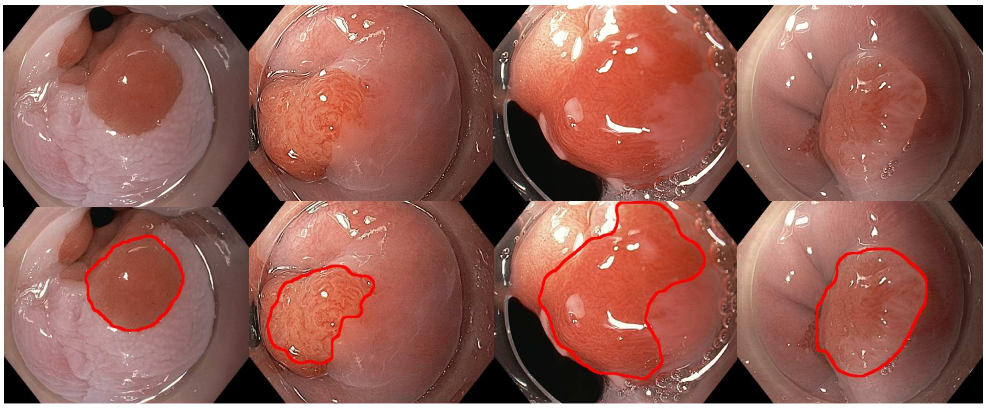


Figure 11.3: Some images positive to adenocarcinoma from the Augsburg dataset and their respective delineation.

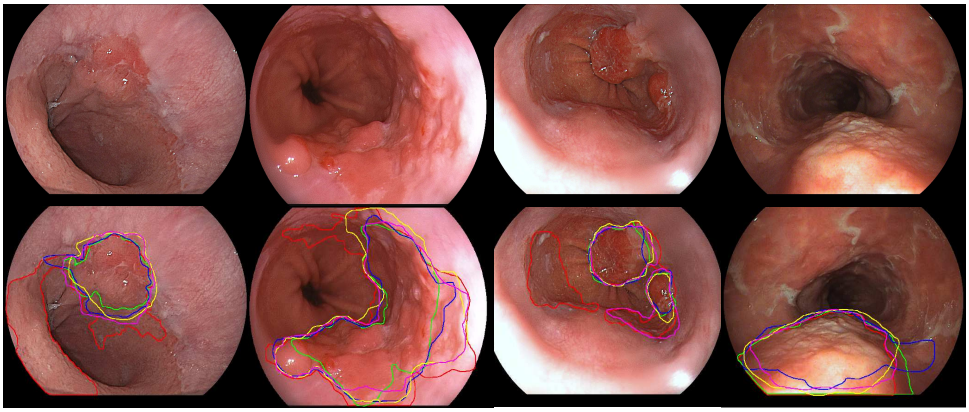


Figure 11.4: Some images positive to adenocarcinoma from the MICCAI dataset and their respective delineation.

11.3.2 Layer-Selective Deep Representations

Considering the ResNet-50 network, the experimental delineation is based on changes in its main architecture to assess each layer's impact over training and classification. First of all, each layer was evaluated by adding a fully connected layer to its output, so the training process was based on each layer's training, instead of evaluating the impact of all the layers through the deep convolutional network. As an output, each layer's accuracy value can be obtained, bringing insights about how deep the network should be. In short, from each convolutional layers that compose the entire ResNet-50 architecture, a fully connected layer is added to its

output, changing the training target from one to n fully connected layers. The convolutional layers are present inside several stages, as depicted in Figure 11.2, and are defined as layers in which convolution transformation are applied. Therefore, each layer inside the ResNet-50 model that comprises a convolution was employed in the process. A model based on n convolution layer outputs is obtained from such a training step, so the classification step aims to define the label of some input sample for each one of the ResNet-50 trained layers for the entire network. In the end, instead of presenting one accuracy output for such evaluated input, n accuracy measures related to each trained layer are calculated. Figure 11.5 illustrates the layer-wise training step.

Further, the layer-wise accuracies were ranked to define the most discriminative ones for the next step of the model, i.e., the shortcut-to-output step. In this context, for every layer i, where $i \in \{1,2,\cdots,n\}$, a new model based on ResNet-50 is defined with a shortcut from layer i to the model's output, and a fully connected layer is added on the top of the model, aiming at classifying BE and adenocarcinoma samples. This shortcut is performed based on addition and concatenation transformations respectively, configuring two different approaches to be evaluated. In a nutshell, $m < n$ different models are selected and evaluated, each with a different layer-to-output shortcut related to the best m layers calculated in the layer-wise step and associated with the layer-wise accuracies. Such a structure has the purpose of defining the best layer accuracies related to the layer-wise step, determining the proper layers for building the selective ResNet-50 models with custom shortcuts in the shortcut-to-output step. Each model's training is based on the main output loss, configuring a conventional training in this step. Figure 11.6 extends its interpretation, addressing the layers' organization to compose the best layer-wise-based models in the shortcut-to-output step.

The layer-wise step was performed over 350 epochs, while each shortcut-to-output step was evaluated over 250 epochs. As a matter of comparison with state-of-the-art approaches, 80% of the data were randomly selected for training purposes, while the remaining 20% was selected for testing (including both layer-wise and shortcut-to-output proposed steps). To keep the fairness related to blind training and testing steps, from such 80% of the training samples, 40% were randomly selected for the layer-wise training step, and the remaining 40% were used for the shortcut-to-output training task. The same was conducted over the testing tasks for both method's steps, with 10% of samples randomly selected for testing over layer-wise step and 10% chosen for testing the best models provided in the shortcut-to-output samples.

Indeed, two standalone training and test processes are performed for both layer-wise and layer-to-output steps. The first step's main goal is to define the most discriminative convolu-

tional layers concerning BE and AD classification based on the fine-tuning of each FC weight defined as outputs of the multi-output convolutional model. With the best convolutional layers in hands, the second step aims to determine the shortcuts of such layers to the CNN's main output, building several learners influenced by the best layers in the model's definition and classification. Once both steps are performed in a standalone fashion, none of the weights are shared during their training or testing tasks. The concept of a "pre-evaluation"is established in a multi-training and testing approach to enhance the main classification outcome and to refine the CNN encoded information related to the convolutional deepness addressed in the model computation.

For comparison purposes, a standard ResNet-50 model with a fully connected layer attached to its output and pre-trained with imageNet was trained and classified with the same amount of samples used for training and testing each shortcut-to-output model. The experiments were conducted: (i) 20-fold cross-validation (20-fold CV) and (ii) leave-one-out cross-validation (LOO CV) protocols.

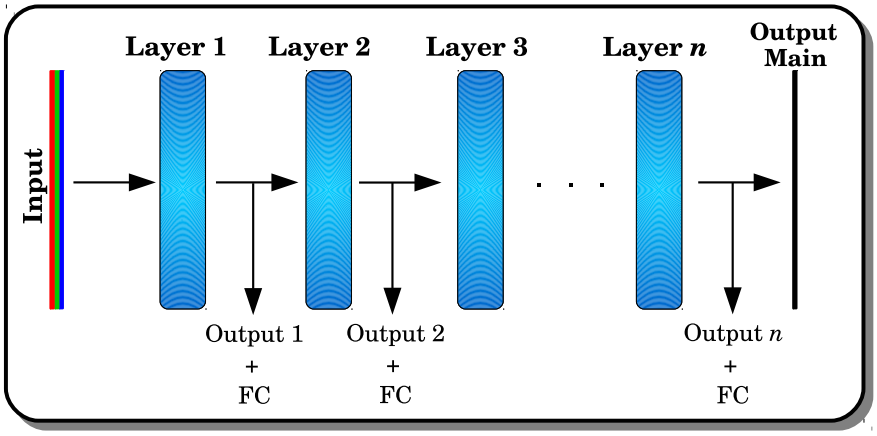


Figure 11.5: Layer-wise step: a base model composed of n convolutional and fully connected layers (FC). While the training step consists of adjusting each layer's weight based on the loss function at the main output, the testing phase comprises the classification performed by each FC consistent with Figure 11.2.

11.4 Experimental Results

This section presents the experiments used to evaluate the proposed methodology. The first round of experiments aimed at evaluating the accuracy of each CNN-layer in the layer-wise step, followed by the evaluation of the layer-to-output step also using the sensitivity (S) and specificity (P), and accuracy (A) rates.

For the sake of keeping the results clear, the maximum number of CNN models comprising

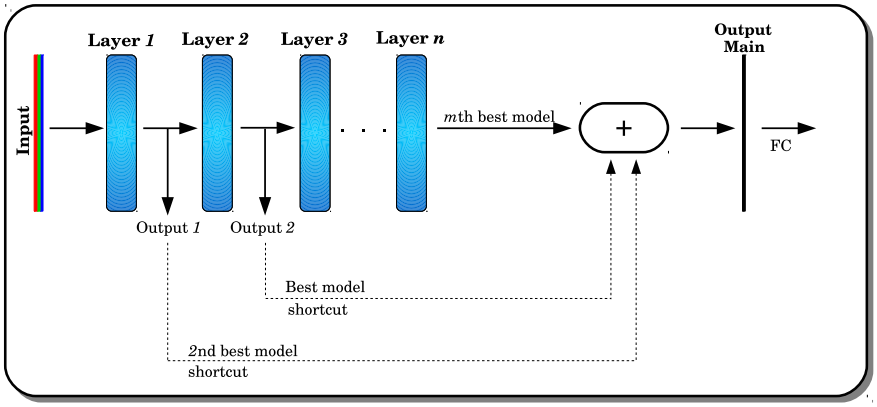


Figure 11.6: Best layers for the shortcut-to-output step: m ($m < n$) layers evaluated in the layer-wise step define the best ResNet-50 models that are trained and classified with the implementation of a shortcut from the layer i to the output. As one can observe, for each layer considered in the best-accuracies, a brand new and independent ResNet-50 model is defined.

the shortcut-to-output step, described by m, was set to five. Therefore, the five best convolution layers belonging to all five ResNet-50 stages were calculated in the layer-wise step, from which up to five best models included in the shortcut-to-output step.

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

Table 11.1: Shortcut-to-output mean accuracy results concerning Augsburg and MICCAI datasets. Best models 1 to 5 mean the best model outcomes related to the best-layers-accuracies calculated during the layer-wise step and defining the five best models of the shortcut-to-output step. The statistically similar results are highlighted in bold, while the best overall result is marked with $\star$.

Protocol	Dataset	Shortcut	Best-model-1	Best-model-2	Best-model-3	Best-model-4	Best-model-5	Baseline
20-fold CV	Augsburg	Concat	0.82 ± 0.03	0.79 ± 0.03	0.77 ± 0.04	0.76 ± 0.03	0.74 ± 0.02	0.69 ± 0.16
		Add	0.79 ± 0.03	0.77 ± 0.03	0.76 ± 0.04	0.75 ± 0.03	0.73 ± 0.03	
	MICCAI	Concat	$\mathbf{0.86 \pm 0.02}$	0.84 ± 0.03	0.82 ± 0.03	0.81 ± 0.02	0.78 ± 0.05	0.76 ± 0.11
		Add	0.85 ± 0.03	0.81 ± 0.02	0.79 ± 0.05	0.77 ± 0.03	0.76 ± 0.05	
LOO CV	Augsburg	Concat	0.84 ± 0.06	0.82 ± 0.02	0.81 ± 0.07	0.80 ± 0.05	0.76 ± 0.04	0.80 ± 0.11
		Add	0.85 ± 0.08	0.84 ± 0.09	0.81 ± 0.06	0.80 ± 0.02	0.79 ± 0.05	
	MICCAI	Concat	$\star\mathbf{0.90 \pm 0.05}$	$\mathbf{0.89 \pm 0.08}$	$\mathbf{0.87 \pm 0.06}$	0.85 ± 0.07	0.83 ± 0.04	0.84 ± 0.15
		Add	$\mathbf{0.88 \pm 0.05}$	$\mathbf{0.87 \pm 0.04}$	0.86 ± 0.09	0.85 ± 0.04	0.85 ± 0.04	

11.4.1 Classification Results

After training the layers' outputs belonging to ResNet-50 architecture in the layer-wise step, the best layers can be obtained for each of the cross-validation runs. A further training and classification were performed in the layer-to-output step, building a shortcut from the i^{th} layer to the CNN output to compose the flatten and final dense layer. Such layer-to-output task was conducted over two different shortcut approaches, i.e., based on concatenation and addition of i^{th} layer's output and CNN's output, and considering only the best layers obtained in the

Table 11.2: Layer-wise mean frequency (%) concerning Augsburg and MICCAI datasets for the 20-fold CV protocol. Best-layers 1 to 5 stand for the best layers selected during the layer-wise step and defining the five best models of the shortcut-to-output step. CS represents the ResNet-50 architecture's Convolutional Stages. In a nutshell, this table provides the frequency at which each ResNet-50 Convolutional Stage appeared in selecting each best five layers in the layer-wise step. The best overall result of each dataset is marked with symbol $\star$, while the best result of each approach is in bold.

Dataset	Shortcut	Convolutional Layer	Best-layer-1	Best-layer-2	Best-layer-3	Best-layer-4	Best-layer-5	Best Block
Augsburg	Concat	CS_1	0%	0%	0%	0%	20%	1
		CS_2	0%	0%	5%	15%	15%	4
		CS_3	20%	35%	40%	30%	15%	4
		CS_4	$\star$**80%**	60%	55%	40%	25%	4
		CS_5	0%	5%	0%	0%	25%	1
	Add	CS_1	0%	5%	10%	0%	25%	1
		CS_2	15%	0%	35%	15%	30%	3
		CS_3	15%	50%	35%	40%	30%	4
		CS_4	**70%**	45%	15%	40%	15%	3
		CS_5	0%	0%	5%	5%	0%	1
MICCAI	Concat	CS_1	0%	0%	0%	15%	25%	1
		CS_2	0%	0%	0%	25%	30%	2
		CS_3	10%	15%	50%	20%	15%	4
		CS_4	$\star$**90%**	75%	50%	30%	20%	6
		CS_5	0%	10%	0%	10%	5%	2
	Add	CS_1	0%	0%	0%	5%	25%	1
		CS_2	10%	15%	35%	0%	15%	2
		CS_3	25%	40%	35%	45%	5%	3
		CS_4	**75%**	40%	25%	50%	25%	4
		CS_5	0%	5%	5%	0%	30%	1

layer-wise step, presented over 5-best outcomes (from the best five layers selected during the layer-wise step). In other words, each best model result is presented, from the very-best to the fifth-best. The layer-to-output classification results for both Augsburg and MICCAI datasets can be observed in Table 11.1, in comparison to the plain CNN classification results (Baseline) also presented and based on the same protocol.

As one can observe, for the 20-fold CV approach, the accuracy results of the shortcut-to-output step decreased from Best model "1"to "5", i.e., using information from only one intermediate layer ($m = 1$) is the best scenario. The best result for Augsburg dataset, i.e., the value of 0.82 ± 0.03, was obtained in the concatenation-based shortcut approach. Concerning MICCAI dataset, the best accuracy results were also achieved in the concatenation approach, with a mean value of 0.86 ± 0.02. It is important to highlight that the results of all five best models outperformed the final classification results using the standard ResNet-50 architecture as a baseline for this classification approach.

A similar behavior recognized in the previous approach could also be observed concerning the LOO CV approach, with a tendency of accuracy decreasing from "1"to "5"best-models. For the Augsburg dataset, the best accuracy result was obtained in the add-based shortcut approach, with a value of 0.85 ± 0.08, while for the MICCAI dataset, the best accuracy result was achieved

in the concatenation approach, with a mean value of 0.90 ± 0.05.

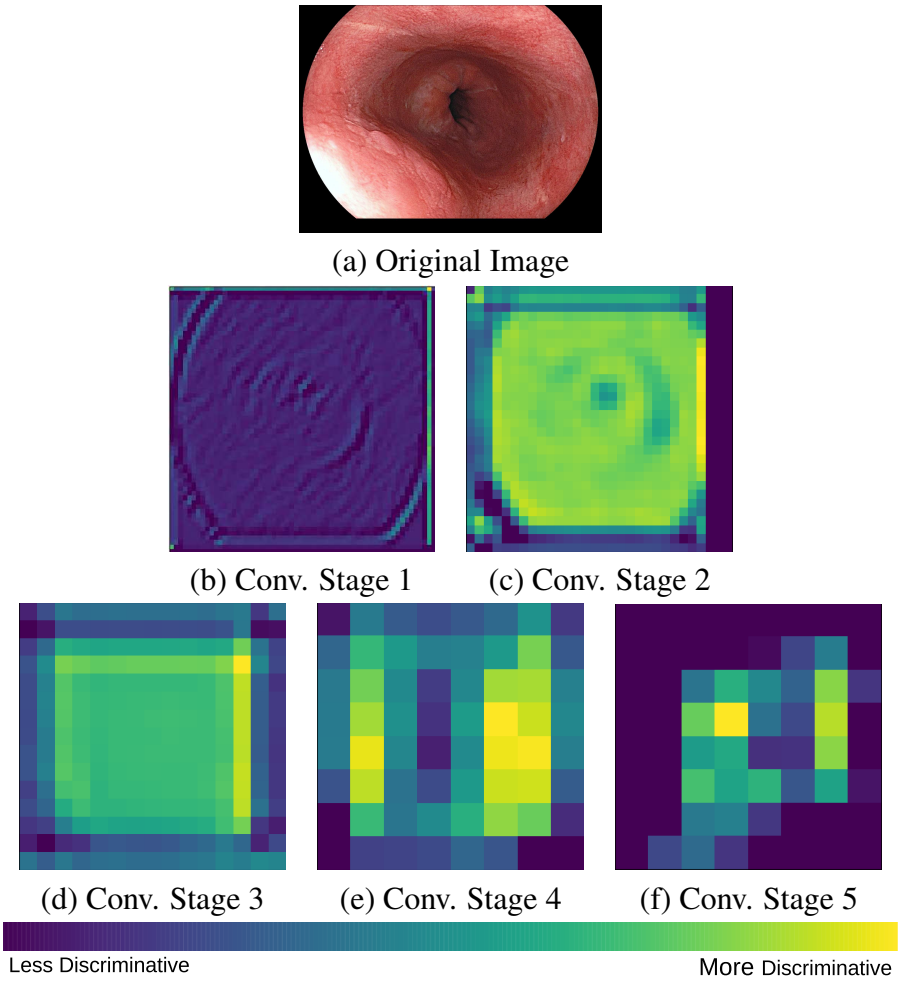


Figure 11.7: Convolutional stage outputs: the regional information is encoded through the deep convolutional stages of ResNet-50. The images presented in *viridis* color palette illustrate the encoded information level related to each ResNet-50 stage (Figure 11.2). Such images comprise the average value of all feature maps as output of a convolutional layer concerning each main stage. Notice values range from dark blue (not discriminative) to green (discriminative). Even with less local-visualizing information, the generalization achieved in deeper convolutional stages comprises high discriminative features for the classification task.

For each trial of the cross-validation approach, different layers belonging to different convolutional blocks may be defined as the best ones. Therefore, each convolutional block's frequency to be selected as one of the five best layers in the layer-wise step over all the 20 runs is presented in Table 11.2. Notice that ResNet-50 presents five main convolutional layers (convolutional stages) in its architecture, and each main convolutional layer is composed of sets of inner layers, building its deep-convolutional nature. The "Best Block"column relates the convolutional transformation inside each convolutional stage block that presented the best accuracy result when selected for the best models' definition during the layer-wise step of the model. Each stage block is composed of several convolutions, poolings, and activations, and such a co-

lumn expresses the inner convolutional transformation output that presented the best accuracy for such a stage block. As one can observe, for all experiments, convolutional stage number 4 (CS_4) always presented the highest frequency in the Best-model-1, and also high frequency in Best-model-2 results, followed by convolutional stages 3 and 2, in both dataset analysis. As long as the convolutional transformations are applied to the input image, less local information can be observed, but more class meaningful information may be comprised, as one can observe in Figure 11.7.

To calculate each pixel impact for all convolutional stages, one may consider that deeper layers inside deep models present lower resolution but a higher number of channels. Due to that, images belonging to deeper layers tend to lose their region features in encoded information presented by the high-channel dimensionality. For the presented in Tables 11.1 and 11.2, Figure 11.7 illustrates an input interpreted by a convolutional stage selected for a best-model. Considering the wide range of channels, a mean channel was calculated, and the impact of each pixel over all the channels could be achieved. The images are presented in *virids* color pallet, highlighting the impact of the pixels over the mean-chanel transformed inputs.

11.5 Discussion

We could not observe any previous study proposing the evaluation of deep learning models in the BE and adenocarcinoma context to provide layer-learning-insights into the model's generalization process. In this work, we fostered the research toward such tasks by introducing a CNN layer-wise technique for classification purposes inspired by multistep evaluation in a quantitative assessment. We outperformed the results obtained in some recent classification works with a similar database and protocol [Souza Jr. et al. 2018, Souza Jr. et al. 2019, Souza Jr. et al. 2017, Souza Jr. et al. 2018, Souza Jr. et al. 2017]. Table 11.3 presents a more detailed comparison of the current results against a fair selection of recent state-of-the-art works in which a similar protocol or dataset was employed.

With the wide application of ML techniques in the medical field, the decision's interpretation seems to be a must-do task. As long as the results are improved, the black-box generalization, learning process, and evaluation of samples must be understood and described, driving insights into how the problem was dealt with. Along with the CNN model, tools can interpret training and validation processes in the way of correct and incorrect classification assessment. For ResNet-50, a powerful CNN architecture presenting very promising results through the years and different fields, a better understanding of the impact of its deep-nature in the evalua-

Table 11.3: Comparison against state-of-the-art works with the application of similar evaluation protocols.

Dataset	Method	Data Augmentation	Images	Protocol	Outcome
MICCAI	[van der Sommen et al. 2016]	No	Selected Region of Interest	Leave-one-out cross-validation	0.86 Sensitivity 0.87 Specificity
	[Ghatwary, Zolgharni e Ye 2019]	No	Downsized Images	Leave-one-out cross-validation	0.96 Sensitivity 0.92 Specificity
	[Ghatwary, Ye e Zolgharni 2019]	Yes	Selected Region of Interest	Leave-one-out cross-validation	0.95 Recall 0.91 Precision
	[van der Putten et al. 2020]	Yes	Preprocessed Images	5-fold cross-validation	0.88 Accuracy 0.93 Sensitivity 0.83 Specificity
	[Souza Jr. et al. 2017]	No	Full-images	20-fold cross-validation	0.74 Accuracy 0.73 Sensitivity 0.78 Specificity
	[Souza Jr. et al. 2019]	No	Full-mages	20-fold cross-validation	0.79 Accuracy 0.82 Sensitivity 0.76 Specificity
	[Souza Jr. et al. 2018]	No	Full-images	20-fold cross-validation	0.69 Accuracy 0.61 Sensitiviy 0.71 Specificity
	[Passos et al. 2019]	No	Full-images	20-fold cross-validation	0.67 Accuracy 0.58 Sensitivity 0.77 Specificity
	[Souza Jr. et al. 2020]	Yes	Full-images	20-fold cross-validation	0.85 Accuracy 0.88 Sensitivity 0.82 Specificity
	Proposed Method	No	Full-images	20-fold cross-validation	**0.86 Accuracy** **0.85 Sensitivity** **0.90 Specificity**
	Proposed Method	No	Full-images	Leave-one-out cross-validation	**0.90 Accuracy** **0.88 Sensitivity** **0.94 Specificity**
Augsburg	[Souza Jr. et al. 2019]	No	Full-images	20-fold cross-validation	0.73 Accuracy 0.71 Sensitivity 0.75 Specificity
	[Souza Jr. et al. 2018]	No	Full-images	20-fold cross-validation	0.65 Accuracy 0.64 Sensitivity 0.66 Specificity
	[Souza Jr. et al. 2020]	Yes	Full-images	20-fold cross-validation	0.83 Accuracy 0.80 Sensitivity 0.86 Specificity
	Proposed Method	No	Full-images	20-fold cross-validation	**0.82 Accuracy** **0.79 Sensitivity** **0.86 Specificity**
	Proposed Method	No	Full-images	Leave-one-out cross-validation	**0.85 Accuracy** **0.83 Sensitivity** **0.89 Specificity**

tion of cancerous samples seems legit, considering many convolutional stages are related to the performed learning process. As a matter of interpretation of layer-learning impact, the output of each convolutional layer that comprises ResNet-50's architecture was evaluated to the final output.

In a close observation of the improved results regarding shallow ResNet architectures instead of its fully-deep version for defining the best deepness level by selecting the best convolutional layers and blocks, one can presume the architecture deepness severely influences the optimal classification result. Due to that, an ablation study was conducted to assess other Res-

Net architectures and ensure that better results can be achieved when properly selecting the best convolution level within the networks. For such, ResNet-18, ResNet-34, and Resnet-101 were employed as deep networks in the ablation set of experiments, presented in Figure 11.8. The same comparison protocol was applied during the ablation study, comparing the best models with the baseline of the respective architectures, represented by its standard architecture pre-trained over imageNet and with no layer evaluation applied to it. As one can observe, shallow architectures, as ResNet-18, required more of its convolutional layers to achieve the best results, very close to the baseline one. From ResNet-34 to ResNet-101, the best results could be achieved by applying our layer-selective method, with outcomes that outperformed the baseline mostly by reducing the impact of deeper convolutional blocks in the composition of the last feature maps, i.e., for ResNet-34, the best results lied on convolution block 4, while for ResNet-101, the best results were obtained by highlighting convolutional block 3 impact on the final feature map output. The improvements obtained applying the proposed layer-selective method over ResNet architectures can also be observed, in visual terms, in Figure 11.8.

Table 11.4: Ablation study of ResNet architectures using the layer-selective method in the 20-fold CV protocol. The statistically similar results are highlighted in bold, while the best overall result of each ResNet architecture is marked with symbol $\star$.

Architecture	Dataset	Shortcut	Best-model-1	Best-model-2	Best-model-3	Best-model-4	Best-model-5	Baseline
ResNet-18	Augsburg	Concat	0.65 ± 0.02	0.64 ± 0.04	0.64 ± 0.05	0.62 ± 0.03	0.62 ± 0.04	0.61 ± 0.08
		Add	0.66 ± 0.04	0.65 ± 0.05	0.63 ± 0.03	0.62 ± 0.07	0.61 ± 0.04	
	MICCAI	Concat	$\star 0.77 \pm 0.04$	0.75 ± 0.02	0.74 ± 0.04	0.72 ± 0.06	0.71 ± 0.04	0.71 ± 0.09
		Add	0.76 ± 0.06	0.75 ± 0.05	0.72 ± 0.04	0.72 ± 0.05	0.71 ± 0.09	
ResNet-34	Augsburg	Concat	0.70 ± 0.06	0.67 ± 0.04	0.66 ± 0.05	0.64 ± 0.04	0.63 ± 0.11	0.63 ± 0.11
		Add	0.68 ± 0.05	0.67 ± 0.04	0.65 ± 0.02	0.63 ± 0.11	0.62 ± 0.09	
	MICCAI	Concat	$\star 0.81 \pm 0.04$	0.80 ± 0.02	0.78 ± 0.05	0.76 ± 0.03	0.74 ± 0.07	0.74 ± 0.06
		Add	0.79 ± 0.06	0.78 ± 0.08	0.77 ± 0.03	0.74 ± 0.06	0.72 ± 0.09	
ResNet-101	Augsburg	Concat	0.83 ± 0.05	0.81 ± 0.05	0.79 ± 0.05	0.77 ± 0.12	0.76 ± 0.16	0.75 ± 0.07
		Add	0.81 ± 0.08	0.79 ± 0.06	0.78 ± 0.09	0.76 ± 0.15	0.75 ± 0.07	
	MICCAI	Concat	$\star\mathbf{0.88 \pm 0.09}$	$\mathbf{0.87 \pm 0.11}$	$\mathbf{0.86 \pm 0.06}$	$\mathbf{0.85 \pm 0.12}$	0.83 ± 0.10	0.82 ± 0.11
		Add	$\mathbf{0.86 \pm 0.03}$	$\mathbf{0.85 \pm 0.11}$	$\mathbf{0.85 \pm 0.12}$	0.83 ± 0.15	0.83 ± 0.13	

Concerning the time consumed to classify samples using the proposed method, one can observe that due to keeping the original CNN architecture but only proposing a new shortcut (that does not change the architecture itself), our proposed method does not increase the original number of parameters during the learning process, maintaining the same time as the baseline model. In comparison to the work proposed by [Hou et al. 2021], which achieved a refresh rate of ≈ 60 frames per second during the classification task for the core model (using ResNet50 architecture), our can achieve a refresh rate value of ≈ 75 frames per second. In addition, our method's training process presents a modular nature, requiring significantly fewer epochs to outperform the baseline model's result (and not only for the core architecture but also for other ones presented in 11.4), aforementioned in Section 11.3. Still, regarding the classification

results, our method presents competitive ones compared to the literature, achieving sensitivity and specificity extremely close to the ones obtained by [Hou et al. 2021], but in a model designed to keep the original CNN structure with fast classification outcome without adding extra and not necessary parameters that can increase its complexity.

In an in-depth analysis of the results presented in section 11.4, one can observe that inner convolutional blocks provided the highest accuracy results for BE and adenocarcinoma identification. The very best results, presented over best-model-1 classification, have always shown a convolutional layer from ResNet-50 stage 4. This outcome suggests an interesting interpretation, once all five best layers also outperformed the respective Baseline classification. For BE and adenocarcinoma context, one can observe that the entire ResNet-50 architecture is not necessary to achieve promising results. Still, it may harm the classification rates with layer outputs that do not present relevant meaning. Considering that convolutional stage 4 presented the very best results, followed by stages 3 and 2, one can conclude that not only the regional information is important (as observed in [Souza Jr. et al. 2019]), but also the convolutional-encoded information (Figure 11.7). ResNet-50 architecture presents a high generalization ability. However, the layer interpretation highlights relevant insights about BE and adenocarcinoma identification, suggesting that less convolutional steps must be applied once a shallow architecture could outperform its results.

Indeed, a close look at Figure 11.7 may raise the following question: do humans and CNNs observe and understand the same set of information regarding BE and AD context? Given such an assumption, one should consider that the processing applied to the input samples in the model's generalization change the image nature in the long term, encoding information from all processes that compose the net (convolution, normalizations, and so on). Even needing the starting layers to obtain the best results, the outcome also suggests that some critical encoded information related to all performed transformations is essential to enhance the classification rate. Therefore, our result shows that the most relevant CNN information (the best classification result) is not exactly the related to local-and-visual input information, once the most discriminative block, CS_4, regarding our results, encodes information to a less visual interpretative level. Hence, the visual representation of early cancerous tissue is encoded through the network layers [Souza Jr. et al. 2018, Souza Jr. et al. 2017, Souza Jr. et al. 2019], but at some level of abstraction, it can be incorporated into human knowledge to improve even more the classification of cancerous samples.

The shortcut task was based on concatenation and addition transformations. For both datasets, the concatenation approach presented the highest results, and this may be explained by the

fact that concatenation needed less output processing than the addition one. For adding some layer output to ResNet-50 final output, pooling and convolutional processes must be performed, while for concatenating, only the pooling task must be done. These processes of convolution and pooling must be performed for matching the outputs' dimensions, and enable the proper encoding of information we aim at achieving. With less processing, the layer output may present less changes to its original value, and observing the experimental results, may achieve higher accuracy results.

Regarding the evaluation of both datasets, Augsburg dataset achieved results lower than the MICCAI outcomes, reinforcing that the evaluation and interpretation of such a dataset are more challenging. Even though still satisfactory outcomes could be achieved, outperforming the state-of-the-art results by presenting a layer interpretation that enhances the adenocarcinoma classification of cancerous instances for BE images.

11.6 Conclusion

In this paper, we dealt with computer-assisted Barrett's esophagus and adenocarcinoma identification, interpretation, and comparison by means of deeply-learnable features computed using ResNet-50 network and its inner layers' outputs.

The main achieved contributions of the study are presented as follows:

- The interpretation of black-box generalization in BE endoscopic images based on a two-step layer-wise method showed up to be promising, presenting meaningful insights about how deep a CNN architecture should be to present trustworthy accuracy outputs;
- The shortcut based on concatenation showed up better results than the ones based on addition, but still figures as an interesting way to cope with the task of highlight more meaningful layers in a CNN architecture;
- The convolutional layer 4, as the most discriminative one, highlights how important is the combination of regional and meaningful features in the learning process of BE and adenocarcinoma, reinforcing that CNN architectures may present strong generalization ability, harming up the correct classification when too deep networks are employed;

Regarding future work, we aim to consider a more-discriminative layer combination for the shortcut step, proposing an evaluation that keeps combining promising layer blocks to improve even more the classification rates. Also, a more in-depth layer interpretation may be employed

by evaluating most-discriminative layer neurons to define the most interesting layers to be evaluated.

11.7 Chapter's Considerations

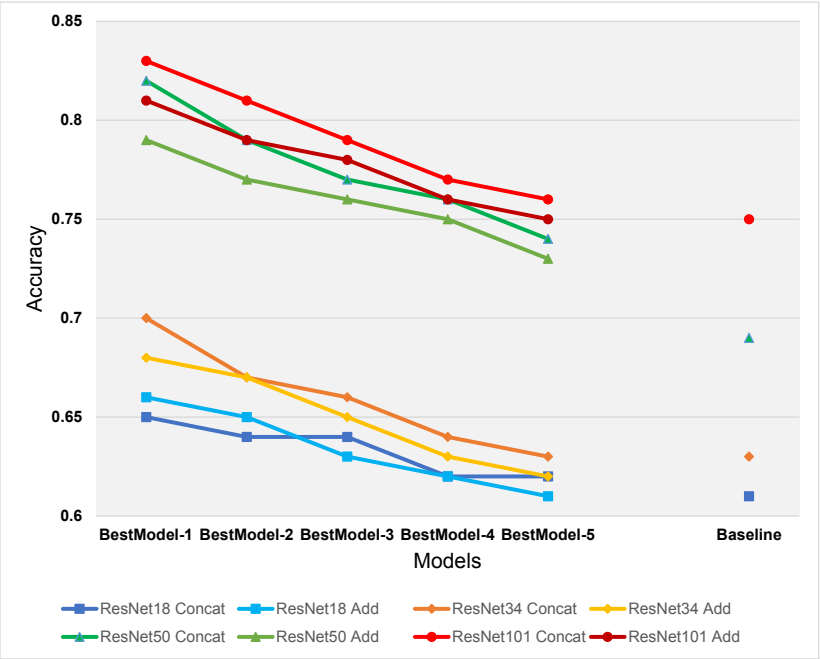
After evaluating the potential of CNN models in addressing BE and adenocarcinoma distinction, and corroborating its benefits when applied to the studied context, this Chapter's main goal is to propose a new XAI-inspired method of interpreting deep architecture's learning by extensivelly analysing layers' information during training and testing tasks. Such a contribution would promote understanding across deep networks' decisions by understanding the transformations performed over each convolutional layer of its configuration. Finally, by observing the discriminative power of each layer, the early cancer detection rate could be enhanced by imposing more impact of such a layer on the final descriptor computing, using shortcuts present in residual networks.

Residual deep networks, such as ResNet-based ones, are composed of weights sharing within its configuration, called skip connections or more popularly shortcuts, to avoid gradient vanishing and accuracy saturation. It is very well-known that deep convolutional nets change the spatial nature of inputs due to convolutional and nonlinear transformations throughout its training step. Hence, this manuscript adopted residual nets, such as ResNet-50, to understand how impactful layers within the network are in the correct classification of esophageal cancer tissue. By changing the natural linear workflow of classification performed by CNN models, we propose a two-stage architecture in which the first one trains every convolutional layer that composes the deep architecture, for observing the most discriminative ones for each class of interest; and the second one adds a new skip connection of meaningful layers detected in first step to the final composition of deeply-learnable features, for further performing the global training and testing process themselves.

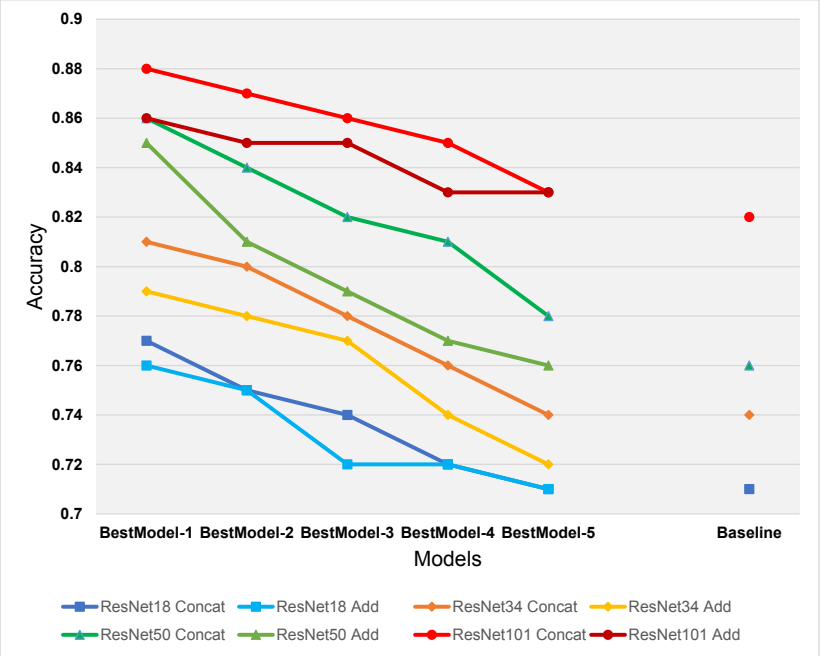
In a clear way of highlighting how important is enconding the spatial information to achieve promising prediction results, from the obtained we could observe that, indeed, the local information is encoded through deep architectures so global description is represented and enables a more accurate definition of adenocarcinoma in BE images. Although this is extremely promising in computational terms, we observed that no visual meaning could be observed from the most discriminative layers, the ones that provided the very best classification results. Our results, presented in sections above, suggest an interpretation of the CNN black-box nature based on the interpretation of layers behavior among CNN architectures, and we successfully delive-

red such an artifact. In fact, the transformation CNN imposes to input samples harms the visual interpretation of its learning process. However, it is possible to observe that, at some point within the layers, the network stops encoding discriminative information to its global representation of the problem, as our results highlight. Deep networks are important, as spatial information, for the correct detection of cancerous regions in BE samples, and as a BoVW-inspired process of encoding more and more information at each layer, CNN models can misunderstand regions as humans do at some point of its generalization.

Our conclusion, by the end of this Chapter, is that spatial information is essential during deeply-learnable features' generalization, and a global-and-encoded representation is performed when deep models are employed to solve classification tasks. Hence, the deepness degree of a convolutional net must be very well observed and analyzed to promote the best overall prediction, as some non-discriminative regions may harm the CNN representation at some levels of deep architectures, as we showed in the results of such Chapter. Finally, after extensively evaluating handcrafted features, deeply-learnable features, and their relation to spatial information in correlation to human insights, for the next and final Chapter, we aim at incorporating all the remarkable achievements we presented until this point, proposing the actual fusion of human-based and computation-based features, once we satisfactorily proved both esophageal feature-representation natures are meaningful for the correct identification of esophageal cancer in BE.



(a) Augsburg



(b) MICCAI

Figure 11.8: Ablation study of the proposed layer-selective method concerning several ResNet architectures. An evident decreasing behavior can be observed over all evaluated architectures from the best models 1 to 5, mostly all the time outperforming the respective baseline accuracy.

Chapter 12

DeepCraftFuse: Handcrafted and Deeply Learnable Features Work Better Together for Esophageal Cancer Detection in Patients with Barrett's Esophagus

To propose a final compilation of all the studied techniques and methods in the identification of early cancer in BE patients, a deep architecture based on high-level regional information and low-level global encoding is proposed, promoting the fusion of handcrafted features and deeply-learnable features. Such a study was submitted to "MICCAI 2022", closing the current research across the evaluation and proper combination of features to enhance the correct identification of esophageal cancer in endoscopic examinations.

12.1 Introduction

Machine learning has shaped how we address tasks considered too complex to be automated. The rising of the deep learning concept helped considerably in such a context, for information extracted from data did not require human intervention mostly. As time went by, scientists realized that the performance of deep nets did not improve to a certain extent, either by the lack of labeled data or their learning capacities were fulfilled.

Some applications require sharp and accurate outputs, for lives depend on it. Computer-assisted medical diagnosis is paramount when dealing with either large amounts of exams or too complex cases (e.g., a rare or improbable disease). In this paper, we are interested in identifying

adenocarcinoma in patients affected by Barrett's esophagus (BE), a disease located at the lower part of the esophagus, inducing changes in the mucosa's cells. In most cases, it is related to obesity and smoking [Lagergren e Lagergren 2010], with remission depending on an early diagnosis [Dent 2011, Sharma et al. 2016, Phoa et al. 2016]. Late diagnosis or neglected treatment leads to more complicated situations, with risks of cancer development and even death [Shaheen et al. 2009, Johnston et al. 2005].

Physicians often face the challenge that esophageal cancer is frequently misdiagnosed as Barrett's esophagus, decreasing the chances of starting early treatment. Some recent works aimed at better understanding the relations between esophageal cancer and BE using endoscopic data. Ghatwary et al. [Ghatwary, Zolgharni e Ye 2019, Ghatwary, Ye e Zolgharni 2019] employed object detection techniques, such as Regional-based CNN, Fast Regional-based CNN, Single-shot Multibox Detector, and Gabor filtering with a VGG-16-based backbone for feature extraction, achieving remarkable results. Hou et al. [Hou et al. 2021] described esophageal early-cancerous tissue using an attentive hierarchical aggregation mechanism, where features aggregate information from adjacent layers from deep models. The idea is to describe the learning meaning and representation capabilities progressively. Early cancer identification was also considered by van der Putten et al. [van der Putten et al. 2020] with a multi-stage learning strategy for classifying and localizing injured tissues in BE and adenocarcinoma samples. The proposed architecture was based on U-Net and trained at different transfer learning stages to compute fine-grained features and accurately describe cancerous regions.

All studies mentioned above employ deep learning methods to cope with the early cancer identification in patients affected by BE, but the models do not use prior knowledge from humans. Although state-of-the-art works have shown promising results using deep nets to distinguish BE from adenocarcinoma, could human knowledge somehow contribute to such a process? We are focused on addressing this question in the manuscript. We are aware of previous works that attempted at a similar idea to some extent [Souza Jr. et al. 2021, Souza Jr. et al. 2020], but they did not blend information learned by deep models with humans' knowledge.

This manuscript proposes *DeepCraftFuse*, which extends deep architectures to combine handcrafted and deeply-learnable features in two modules called *HCF-module* and *DLF-module*, followed by the *FeatFusion-module* where features from the two previous modules are encoded together to make predictions jointly. DeepCraftFuse figures the following advantages: (i) high-level spatial features and low-level semantic context provided by handcrafted and deeply-learnable features can be effectively captured; (ii) it does not require severe changes to original deep architectures but provides simple branches to reuse their flow; and (iii) it provides state-

of-the-art results.

To the best of our knowledge, DeepCraftFuse is the only in-built architecture that gathers deeply learnable and handcrafted features to improve esophageal cancer identification in patients affected by Barrett's esophagus. The proposed approach figures the high-generalization capabilities of deep models and can encode high-specialized knowledge provided by experts, represented here by key points that object detectors found relevant. Findings enlightened by Souza et al. [Souza Jr. et al. 2019] showed that such key points concentrate in the cancerous regions annotated by experts. Experiments demonstrate that DeepCraftFuse outperforms several other state-of-the-art techniques in two datasets composed of endoscopic images.

12.2 Proposed Method

DeepCraftFuse consists of three modules, two parallel and one sequential, that process information differently:

1. *DLF-module* (Deeply-learnable Features): it uses the layers' receptive fields to encode features from a global perspective;
2. *HCF-module* (Handcrafted Features): it recovers local information from convolutions performed with the deep architecture using any suitable object detection approach; and
3. *Feat-Fuse-module*: it combines the local information the HCF-module provides with the global information the DLF-module delivers.

Figure 12.1 depicts the proposed workflow, where a different color represents each module. We say that *DLF* and *HCF* modules work in parallel, for there is no need to wait for the entire feature extraction flow of the backbone to compute the activation maps. After processing the information flow in a particular convolutional layer, they are forwarded to the *HCF-module*.

Features provided by the first two modules feed the FeatFuse-module to compute the final loss. There are two main benefits our DeepCraftFuse architecture offers: (i) first, the combination of local and global information to assist cancer identification, and (iii) to exploit simultaneously prior knowledge from experts and features learned automatically.

DLF-module. Usually, hundreds of layers in deep architectures end up in features that encode global context at different levels of abstraction. The *DFL-module* allows any deep architecture for such a purpose, provided we have a fully connected layer followed by a softmax

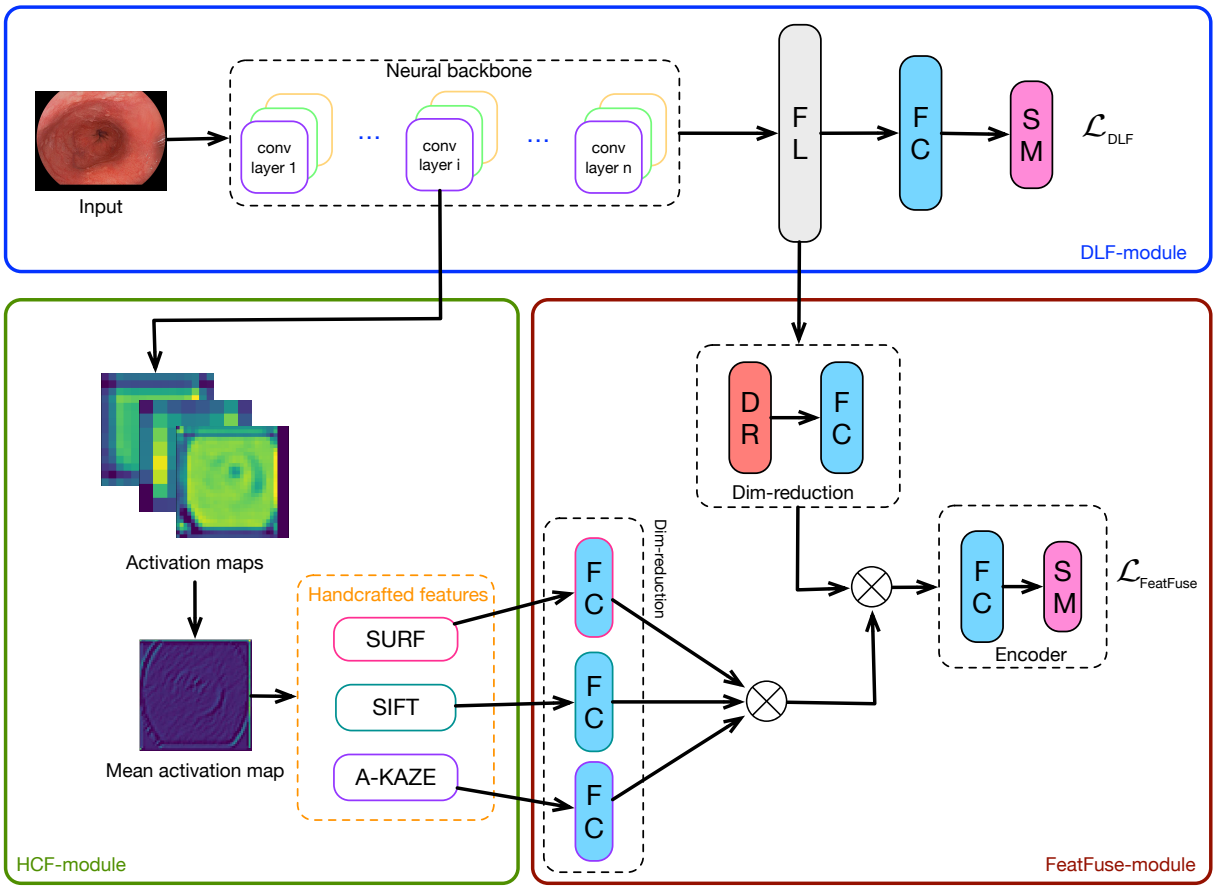


Figure 12.1: DeepCraftFuse architecture (best viewed in color) and its three main modules. FL = flattened layer, FC = fully connected layer, DR = dropout layer, SM = softmax layer, and the symbol $\otimes$ denotes the concatenation operation.

layer for evaluating purposes. This module is in charge of learning features to differentiate positive or negative patients from esophageal cancer using the well-known binary cross-entropy loss function $\mathscr{L}_{DFL}$. Moreover, the *DLF-module* is trivial, not presenting any changes in the backbone used for feature extraction.

HCF-module. For each convolutional layer (*conv layer*) of the backbone used in the *DLF-module*, we take its activations (i.e., feature maps) to compute a mean activation map. Object detectors then extract features. In this work, we considered SURF (Speeded Up Robust Features) [Bay et al. 2008], SIFT (Scale-invariant Feature Transform) [Lowe 2004], and A-KAZE (Accelerated KAZE) [Alcantarilla, Nuevo e Bartoli 2013] to extract key points from the activation map. Nonetheless, any other image description approach may work here. SURF, SIFT and A-KAZE outputs are feature vectors of size 64, 128, and 61, respectively.

FeatFuse-module. To effectively combine features from the deep model and the ones from object detector techniques, we propose a new FeatFuse-module, depicted in the red module in Figure 12.1. The module first reshapes dimensions (*Dim-reduction*) to avoid bias towards either

features coming from *DLF-module* or *HCF-module*. Further on, an *Encoder* concatenates the reduced outputs. The distillation process takes place for the final loss computation, represented by $\mathscr{L}_{\text{FINAL}}$.

Loss Function. The entire model is trained end-to-end with a binary cross-entropy loss function. We elaborate a composite final loss function to improve the gradient flow by adding the term $\mathscr{L}_{\text{DLF}}$. Tue final training loss balances the global knowledge the *DLF-module* encodes with the more localized information the *HCF-module* provides. The final loss function is computed as follows:

$$\mathscr{L}_{\text{FINAL}} = \alpha \mathscr{L}_{\text{DLF}} + \mathscr{L}_{\text{FeatFuse}}, \tag{12.1}$$

where α is a tunable hyperparameter that controls the influence of the global information the deep model provides.

12.3 Experiments and Results

Datasets. Two high-definition white-light endoscopic datasets were employed to evaluate the proposed method. The first dataset comprises examinations from the University Hospital Augsburg, Medizinische Klinik III, Germany, containing 76 endoscopic images captured from different patients affected by BE, of which 42 are negative BE and 34 are positive to esophageal cancer. One physician manually annotated the dataset. The second set of images belongs to a benchmark dataset available at the "MICCAI 2015 EndoVis Challenge"[1], hereinafter denoted as "MICCAI". The dataset comprises 100 endoscopic images of the lower esophagus captured from 39 individuals positive to BE, of which 22 are patients negative to cancer and 17 are positive to early-stage esophageal cancer. Five different physicians have annotated the cancerous samples. Figure 12.2 depicts exemplars from Augsburg and MICCAI datasets.

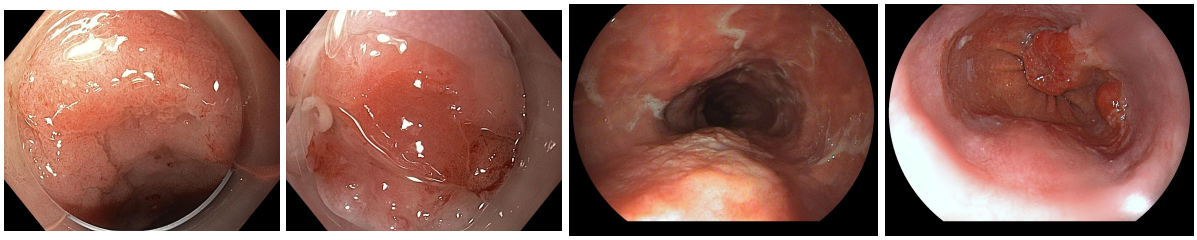


Figure 12.2: Images positive to cancer from Augsburg (first two from the left) and MICCAI (first two from the right) datasets.

[1] https://endovissub-barrett.grand-challenge.org

Evaluation Measures. We employed four well-known metrics to evaluate the proposed approach: Sensitivity (S), Specificity (P), Accuracy (A), and F1-Score (F1). The experimental setup also comprises a statistical evaluation using Wilcoxon's signed-rank test [Wilcoxon 1945] with a significance of 5%. Statistically similar results (based on measure A) are highlighted in bold, and a $\star$ marks the best overall result.

Experimental Delineation. We used a ResNet-50 pre-trained on the ImageNet dataset as a backbone in the DLF-module. DeepCraftFuse employs the Adam optimizer with a learning rate of 10^{-4}, $\beta = 0.5$, weight decay of 0.5, and momentum of 0.999 for fine-tuning ResNet-50 weights in the DLF-module and also to learn the weights in the Dim-reduction and Encoder modules. This second-stage training uses 500 epochs with a batch size of 16, with 80% of samples composing the training set and the remaining serving as the testbed. The training set also comprises data augmentation, with new images generated by rotation, horizontal flipping, and zooming. Each operation adds new images at a rate of 10% of the training set size, with rotation degrees chosen at random within the interval $\{1^\circ, 359^\circ\}$, and zooming rates chosen at random up to 50% of the image size[2]. For statistical analysis, we considered a 20-fold cross-validation approach.

Implementation Details. The experiments employ a computer with 96 GB RAM and an NVIDIA TitanX® Graphics card of 12 *GB* RAM. The implementation used the TensorFlow 2.+ framework.

Classification Results. DeepCraftFuse focuses on three main aspects: feature learning (*DLF-module*), handcrafted feature extraction (*HCF-module*), and feature fusion (*FeatFuse-module*). Since we used ResNet-50 as the backbone, we evaluated DeepCraftFuse on each convolutional layer individually. Table 12.1 presents the outcomes over Augsburg and MICCAI datasets.

The results confirm that initial layers of deep architectures maintain some degree of visual information from the input data, enabling its description using object detectors. The introduction of handcrafted features can gradually increase the correct identification of cancerous regions when calculated from initial convolutional blocks. This behavior is observed for both datasets, with information extracted from the first convolutional layer allowing the best results. It is also a consensus that using all object detectors is more appropriate. According to Souza et al. [Souza Jr. et al. 2017, Souza Jr. et al. 2017, Souza Jr. et al. 2019, Souza Jr. et al. 2018], it is of the great importance to adequately describe the handcrafted features once they must be spatially correlated to the expert-annotated area to influence the classification of cancerous tissue

[2]We used such a zooming rate upper boundary to avoid missing important details of the esophagus area.

Table 12.1: DeepCraftFuse classification results: (a) Augsburg and (b) MICCAI datasets.

Conv Block	SURF	SIFT	A-KAZE	DLF	S	P	A	F1
1	✓	x	x	x	0.75	0.82	0.80	0.79
	✓	✓	x	x	0.81	0.84	0.83	0.82
	✓	✓	✓	x	**0.85**	**0.89**	**0.87**	**0.86**
	✓	✓	✓	✓	⋆**0.86**	**0.89**	**0.87**	**0.87**
2	✓	x	x	x	0.73	0.79	0.77	0.75
	✓	✓	x	x	0.80	0.85	0.83	0.82
	✓	✓	✓	x	0.83	0.85	0.84	0.84
	✓	✓	✓	✓	**0.84**	**0.89**	**0.86**	**0.85**
3	✓	x	x	x	0.71	0.76	0.74	0.73
	✓	✓	x	x	0.70	0.77	0.75	0.72
	✓	✓	✓	x	0.69	0.74	0.72	0.72
	✓	✓	✓	✓	0.73	0.79	0.77	0.75
4	✓	x	x	x	0.70	0.75	0.74	0.72
	✓	✓	x	x	0.70	0.76	0.74	0.73
	✓	✓	✓	x	0.70	0.73	0.72	0.71
	✓	✓	✓	✓	0.72	0.78	0.76	0.74
5	✓	x	x	x	0.68	0.73	0.72	0.71
	✓	✓	x	x	0.68	0.71	0.70	0.70
	✓	✓	✓	x	0.66	0.70	0.68	0.67
	✓	✓	✓	✓	0.70	0.74	0.73	0.72
Baseline	x	x	x	x	0.71	0.79	0.76	0.73

(a) Augsburg

Conv Block	SURF	SIFT	A-KAZE	DLF	S	P	A	F1
1	✓	x	x	x	0.80	0.83	0.82	0.81
	✓	✓	x	x	0.86	0.90	0.88	0.87
	✓	✓	✓	x	**0.90**	**0.93**	**0.92**	**0.91**
	✓	✓	✓	✓	⋆**0.93**	**0.95**	**0.94**	**0.93**
2	✓	x	x	x	0.78	0.82	0.81	0.80
	✓	✓	x	x	0.82	0.87	0.85	0.83
	✓	✓	✓	x	0.85	0.91	0.88	0.86
	✓	✓	✓	✓	0.87	0.91	0.90	0.89
3	✓	x	x	x	0.75	0.81	0.79	0.77
	✓	✓	x	x	0.75	0.83	0.81	0.79
	✓	✓	✓	x	0.75	0.83	0.81	0.79
	✓	✓	✓	✓	0.77	0.85	0.83	0.80
4	✓	x	x	x	0.70	0.77	0.76	0.73
	✓	✓	x	x	0.70	0.75	0.74	0.72
	✓	✓	✓	x	0.68	0.74	0.72	0.71
	✓	✓	✓	✓	0.73	0.81	0.79	0.77
5	✓	x	x	x	0.70	0.73	0.72	0.71
	✓	✓	x	x	0.69	0.72	0.71	0.70
	✓	✓	✓	x	0.67	0.72	0.70	0.69
	✓	✓	✓	✓	0.71	0.79	0.77	0.74
Baseline	x	x	x	x	0.74	0.81	0.80	0.78

(b) MICCAI

meaningfully.

Impact of Different Architectures. Different architectures allow us to draw more robust conclusions. Besides ResNet-50, we also consider ResNet-18, VGG-16, and DenseNet architectures pre-trained on ImageNet as backbones for the *DLF-module*. The same evaluation protocol and hyperparameters are adopted, besides a number of epochs of 350. Table 12.2 presents the outcomes for Augsburg and MICCAI datasets. Similar behavior as in Table 12.1 is observed, i.e., the first convolutional block allows the more accurate results. DeepCraftFuse outperforms all scenarios' baselines (i.e., using the *DLF-module* only). DenseNet provided the best results (slightly better than ResNet-50), but at an expensive training step.

Table 12.2: Effect of different architectures for Augsburg and MICCAI datasets.

Backbone	Best Conv Block	S	P	A	F1
ResNet-18	1	0.70	0.75	0.73	0.72
	Baseline	0.61	0.66	0.64	0.63
ResNet-50	1	**0.86**	**0.89**	**0.87**	**0.87**
	Baseline	0.71	0.79	0.76	0.73
VGG-16	1	0.84	0.87	0.86	0.85
	Baseline	0.74	0.79	0.77	0.76
DenseNet	1	⋆**0.85**	**0.90**	**0.89**	**0.87**
	Baseline	0.78	0.82	0.80	0.79

(a) Augsburg

Backbone	Best Conv Block	S	P	A	F1
ResNet-18	1	0.74	0.79	0.77	0.76
	Baseline	0.65	0.69	0.67	0.66
ResNet-50	1	**0.93**	**0.95**	**0.94**	**0.93**
	Baseline	0.74	0.81	0.80	0.78
VGG-16	1	0.89	0.92	0.90	0.90
	Baseline	0.79	0.83	0.81	0.70
DenseNet	1	⋆**0.93**	**0.96**	**0.95**	**0.94**
	Baseline	0.81	0.85	0.83	0.82

(b) MICCAI

State-of-the-art Comparison. We gathered recently and somehow similar works for comparison purposes against DeepCreaftFuse. Some do not employ the same protocol, but they serve as a basis to evaluate the robustness of the proposed approach. Table 12.3 presents such comparison, for which DeepCraftFuse outperforms all techniques in the two datasets.

Table 12.3: Comparison against state-of-the-art techniques.

Dataset	Method	Year	Input	Protocol	Outcome
MICCAI	van der Sommen et al. [van der Sommen et al. 2016]	2017	Selected Region of Interest	Leave-one-out cross-validation	0.86 Sensitivity 0.87 Specificity
	Souza Jr. et al. [Souza Jr. et al. 2017]	2017	Full-images	20-fold cross-validation	0.74 Accuracy 0.73 Sensitivity 0.78 Specificity
	Mendel et al. [Mendel et al. 2017]	2017	Patches	Leave-one-patient-out cross-validation	0.94 Sensitivity 0.88 Specificity
	Riel et al. [van Riel et al. 2018]	2018	Downsized Images	Leave-one-patient-out cross-validation	0.82 Sensitivity 0.80 Specificity
	Souza Jr. et al. [Souza Jr. et al. 2018]	2018	Full-images	20-fold cross-validation	0.69 Accuracy 0.61 Sensitiviy 0.71 Specificity
	Ghatwary et al. [Ghatwary, Zolgharni e Ye 2019]	2019	Downsized Images	Leave-one-out cross-validation	0.96 Sensitivity 0.92 Specificity
	Ghatwary et al. [Ghatwary, Ye e Zolgharni 2019]	2019	Selected Region of Interest	Leave-one-out cross-validation	0.95 Recall 0.91 Precision
	Passos et al. [Passos et al. 2019]	2019	Full-images	20-fold cross-validation	0.67 Accuracy 0.58 Sensitivity 0.77 Specificity
	Ohmori et al. [Ohmori et al. 2019]	2019	Downsized Images	Leave-one-patient-out cross-validation	0.77 Accuracy 0.81 Sensitivity 0.73 Specificity
	van der Putten et al. [van der Putten et al. 2020]	2020	Preprocessed Images	5-fold cross-validation	0.88 Accuracy 0.93 Sensitivity 0.83 Specificity
	Souza Jr. et al. [Souza Jr. et al. 2019]	2020	Full-mages	20-fold cross-validation	0.79 Accuracy 0.82 Sensitivity 0.76 Specificity
	Souza Jr. et al. [Souza Jr. et al. 2020]	2020	Full-images	20-fold cross-validation	0.85 Accuracy 0.88 Sensitivity 0.82 Specificity
	Hou et al. [Hou et al. 2021]	2021	Full-images	Leave-one-patient-out 5-fold cross-validation	0.93 F1-score 0.91 Sensitivity 0.94 Specificity
	Souza Jr. et al. [Souza Jr. et al. 2021]	2021	Full-images	Leave-one-patient-out 20-fold cross-validation	0.87 Accuracy 0.89 Sensitivity 0.84 Specificity
	Gehrung et al. [Gehrung et al. 2021]	2021	Downsized Images	Leave-one-patient-out	0.91 Accuracy 0.86 Sensitivity 0.95 Specificity
	DeepCraftFuse	2021	Full-images	20-fold cross-validation	**0.95 Accuracy** **0.93 Sensitivity** **0.96 Specificity**
Augsburg	Souza Jr. et al. [Souza Jr. et al. 2018]	2018	Full-images	20-fold cross-validation	0.65 Accuracy 0.64 Sensitivity 0.66 Specificity
	Ohmori et al. [Ohmori et al. 2019]	2019	Downsized Images	Leave-one-patient-out cross-validation	0.71 Accuracy 0.78 Sensitivity 0.65 Specificity
	Souza Jr. et al. [Souza Jr. et al. 2019]	2019	Full-images	20-fold cross-validation	0.73 Accuracy 0.71 Sensitivity 0.75 Specificity
	Souza Jr. et al. [Souza Jr. et al. 2020]	2020	Full-images	20-fold cross-validation	0.83 Accuracy 0.80 Sensitivity 0.86 Specificity
	Souza Jr. et al. [Souza Jr. et al. 2021]	2021	Full-images	Leave-one-patient-out 20-fold cross-validation	0.84 Accuracy 0.83 Sensitivity 0.87 Specificity
	Gehrung et al. [Gehrung et al. 2021]	2021	Downsized Images	Leave-one-patient-out	0.83 Accuracy 0.81 Sensitivity 0.86 Specificity
	DeepCraftFuse	2021	Full-images	20-fold cross-validation	**0.89 Accuracy** **0.85 Sensitivity** **0.90 Specificity**

Ablation Study. We conducted an ablation study to understand the impact of α in Equation 12.1, for it weights the impact of DLF-module in the final loss. We considered $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ using the same set of parameters defined in the previous experiment, but now performed over 250 epochs. Such a methodology lets us figure out how important the handcrafted features are concerning the ones learned by deep nets. Table 12.4 presents the ablation

outcomes, with $\alpha = 0.5$ leading to the more accurate results in all architectures and datasets. Moreover, it is evident a degradation in performance when features learned by the backbone are not used ($\alpha = 0$).

Table 12.4: Ablation study for the assessment of α parameter.

Backbone	α	S	P	A	F1
ResNet-18	0.0	0.65	0.70	0.68	0.67
	0.25	0.67	0.70	0.69	0.68
	0.50	0.68	0.72	0.71	0.70
	0.75	0.66	0.71	0.69	0.66
	1.0	0.66	0.72	0.69	0.68
	Baseline	0.61	0.66	0.64	0.63
ResNet-50	0.0	0.73	0.74	0.74	0.73
	0.25	0.75	0.83	0.80	0.78
	0.50	**0.82**	**0.86**	**0.85**	**0.84**
	0.75	0.80	0.86	0.84	0.83
	1.0	0.78	0.87	0.83	0.81
	Baseline	0.71	0.79	0.76	0.73
VGG-16	0.0	0.77	0.80	0.78	0.77
	0.25	0.79	0.82	0.81	0.80
	0.50	**0.82**	**0.84**	**0.84**	**0.83**
	0.75	0.79	0.81	0.80	0.80
	1.0	0.80	0.83	0.82	0.81
	Baseline	0.74	0.79	0.77	0.76
DenseNet	0.0	0.79	0.82	0.81	0.80
	0.25	0.80	0.84	0.83	0.82
	0.50	$\star$**0.83**	**0.86**	**0.85**	**0.84**
	0.75	0.81	0.85	0.84	0.82
	1.0	0.78	0.85	0.82	0.80
	Baseline	0.78	0.82	0.80	0.79

(a) Augsburg

Backbone	α	S	P	A	F1
ResNet-18	0.0	0.68	0.72	0.71	0.70
	0.25	0.70	0.73	0.72	0.71
	0.50	0.72	0.77	0.75	0.74
	0.75	0.71	0.74	0.73	0.72
	1.0	0.69	0.75	0.73	0.71
	Baseline	0.65	0.69	0.67	0.66
ResNet-50	0.0	0.84	0.87	0.86	0.85
	0.25	0.87	0.86	0.87	0.86
	0.50	**0.88**	**0.92**	**0.91**	**0.90**
	0.75	0.85	0.90	0.88	0.87
	1.0	0.81	0.89	0.86	0.84
	Baseline	0.74	0.81	0.80	0.78
VGG-16	0.0	0.80	0.83	0.82	0.81
	0.25	0.81	0.84	0.83	0.82
	0.5	0.85	0.86	0.86	0.85
	0.75	0.80	0.83	0.82	0.81
	1.0	0.82	0.84	0.84	0.83
	Baseline	0.79	0.83	0.81	0.70
DenseNet	0.0	0.83	0.87	0.85	0.85
	0.25	0.85	0.88	0.87	0.86
	0.50	$\star$**0.88**	**0.93**	**0.92**	**0.90**
	0.75	0.86	0.90	0.88	0.87
	1.0	0.84	0.89	0.87	0.86
	Baseline	0.81	0.85	0.83	0.82

(b) MICCAI

12.4 Conclusion

We present DeepCraftFuse, a novel approach for combining deep networks and object detectors to improve esophageal cancer detection in patients diagnosed with Barrett's esophagus. Our approach does not require severe changes in the backbone configurations and object detectors. DeepCraftFuse outperforms several state-of-the-art techniques in two datasets, including baselines designed in different scenarios. This work aims to bring a new perspective on using deep networks for feature extraction and information provided by experts, represented here by object detectors. Future works include the evaluation of DeepCraftFuse on other medical-related classification tasks.

12.5 Chapter's Considerations

To compile all the developed work proposed in previous Chapters, this one explored the possibility of combining the best achievements of using handcrafted features, deeply-learnable

features, and proper classifiers for the identification of early esophageal cancer.

From the observed in Chapters 3 to 7, the extensive evaluation of handcrafted features showed up its important relevance and relation to spatial insights provided by experts' knowledge in the correct identification of cancerous tissues. After, Chapters 8 to 11 extensively assessed the computation of fully-automated features based on deep learning models, their interpretation, and learning behavior in comparison to human observations of cancerous tissue definition and localization. Moreover, this Chapter depicts all of the best knowledge this research could build, reuniting high-level spatial information of esophageal cancer with low-level global information deep learning models encode in the correct description of cancerous tissue.

To cope with the fusion we propose, we present a new architecture called DeepCraftFuse, that uses deep learning models' inner layers information for calculating spatial-and-relevant information of the esophagus, while the prediction process is guided by a self-distillation mechanism that re-uses the global encoded generalization that CNN models provide. By the use of object detector techniques for the spatial description and CNN models for the global representation, DeepCraftFuse promotes a fusion process that not only presents a novel way of cancer identification but significantly enhances its accuracy in qualitative and quantitative ways, once the natures of both features could be clearly studied and explained through this entire thesis.

This final Chapter aggregates the knowledge we could compute through the study period, answering several questions about "how impactful could be the use of spatial and global information in the correct classification of esophageal cancer?; how could the learning process provided by deep learning models be combined with the human's knowledge described by computational techniques?; and moreover, it could be possible to combine handcrafted and deep learning features, taking advantage of the best of each description, for enhancing the identification of early cancer in BE patients?". The answers, described in detail all over this Chapter, represent the relevance of this research for the medical and computational communities, once we could satisfactorily propose an end-to-end model that effectively employs the best of human and computational representations of a context, highlighting its advantages in an interpretable and accurate unified method for early esophageal cancer prediction.

Chapter 13

CONCLUSIONS AND FUTURE WORK

The present text was organized into thirteen chapters, described as follows: the introduction describing the research context and the motivation and main contributions related to the proposed subject as Chapter 1. Chapter 2 briefly presentes the theoretical background regarding the objective of the research in a compilation of the most important works related to the research subject. Chapter 3 presented the first proposed experiment in which the regions of adenocarcinoma were evaluated by the feature dimension reduction of SURF descriptors, while Chapters 4, 5, and 6 present the progressive evaluation of the adenocarcinoma and BE context by the introduction of, respectively, the OPF classifier and SIFT technique for the feature calculation, the color co-occurrence matrices in the feature extraction step, and the unsupervised OPF and A-KAZE features in the feature extraction, in a concept of visual-words generalization. Closing the handcrafterd features evaluation, Chapter 7 applied iRBM for Barret's Esophagus and adenocarcinoma detection. Every chapter regarding the evaluation of handcrafted features were already published in high impactful conferences and journals.

The results obtained from Chapters 2 to 7 are related to handcrafted features and confirm the promising research area proposed and followed during the first research period. Some techniques for the "handcrafted features"evaluation of the problem were proposed, executed, and published. The current results present an important understanding of the behavior of the handcrafted features for the BE and adenocarcinoma context evaluation. As long as these features are obtained based on the representation of experts' knowledge, it is important to realize how such description may further contribute to the correct definition of early cancer in the esophagus. Several classifiers also present an important understanding of the features' behavior and ways of evaluating them. The observation of handcrafted features was essential to understand its relation to human knowledge and how impactful its correct localization is to the proper definition of cancer in BE-diagnosed patients. Moreover, satisfactory outcomes could be achieved,

describing how important is the representation of experts' insights for the correct cancer definition in our experiments, leading the next steps across the evaluation of features extracted using deep learning techniques.

With the presented from Chapters 2 to 7, we could positively assess the specific hypothesis concerning the importance of handcrafted features in the correct prediction of early cancer in the esophagus, observing clear spatial correspondence in the computational feature extraction and the experts' cancer delineation for achieving promising results. Such achievements could be obtained by the evaluation of several handcrafted features approaches, always combined with different ML classifiers that highlighted the discriminative potential of each description method. The assessments concerning the handcrafted features description allowed us to understand its behavior in quantitative ways, totally related to the observed by the experts during early-cancer surveillance, according to the assumption we first made with the hypothesis.

Following the research studies, chapters 8 to 12 focused on the evaluation of fully-automated systems for BE and adenocarcinoma identification based on deep learning techniques. Chapter 8 introduced the use of GAN-technique for promoting a robust and high-quality synthetic image data augmentation for BE and adenocarcinoma context for further classification, accepted for publication at the journal "Computers in Biology and Medicine". Chapter 9 continued the work proposed in Chapter 8 by proposing a GAN-hyperparameters fine-tuning using meta-heuristic techniques. Chapter 10 applied XAI techniques to lighten up discriminative regions belonging to the model generalization and classification of esophagus-cancerous samples, while the last Chapter, 11, focused on enhancing the classification rates based on the pre-training of CNN layer outputs, providing an understanding of learning behavior to be encoded into the correct identification of samples. The results obtained in all the presented chapters confirm the promising research area proposed and followed during the research period, where some techniques for the "deeply-learnable features"evaluation of the problem were proposed and executed.

With the presented from Chapters 8 to 12, we could evaluate the specific hypothesis regarding the impact of deep learning techniques in the correct detection of early cancer in the esophagus. We were able to understand, in quantitative and qualitative ways, that the generalization performed during deep models' learning compromises visual insights of BE context, but the final abstraction they compute is highly discriminative in the task of detecting early esophageal cancer. Then, we could understand that the generalization performed by deep architectures encodes the information needed to describe and predict esophageal cancer properly. Nowadays, experts demand interpretative models that require deep models to highlight the behavior behind its learning and decision activities. Hence, as a further and required observation, essential as-

pects of interpretation arose to solve CNN models' black-box nature, and the XAI application was essential for us to express the prediction process of deep models we employed. Moreover, such an achievement is correlated to spatial insights in the feature description step, as it is for the manual definition performed when experts track the cancer tissue during surveillance.

Considering the main proposal of such a project, in which the handcrafted features would somehow be merged with the fully-automated features, the results observed in Chapter 12, that proposes the fusion of both description natures, present an important understanding of their behavior and interpretation concerning BE and adenocarcinoma context. As long as these features are obtained with and without experts' knowledge, it is important to realize how such descriptions may further contribute to each other in a models' generalization based on both musts, which we presented in Chapter 12. Hence, our research has brought remarkable contributions to both computational and medical fields, highlighting the importance of humans' insights for detecting early-esophageal cancer, the spatial importance and relevance in the calculation of computational features, which must be correlated to experts' regions to present proper meaning of correct cancer and BE definitions, and finally, how impactful can be the association of handcrafted and deeply-learnable features in the representation of cancerous esophageal tissues, significantly enhancing the correct classification of cancerous samples but still carrying important spatial and human-interpretable information to its core processing.

Our main hypothesis could be fully evaluated with the proposal of several ML and image processing techniques to describe and identify cancer samples in BE images; the evaluation of humans' knowledge representation by the use of handcrafted features, showing up its importance to the classification's success; the relevance and meaning of applying deep learning techniques to enhance the classification of early esophageal cancer, highlighting not only its generalization potential but also its correlation to human's spatial interpretations of the same problem; and by the end, how important is the representation of esophageal cancer based on both handcrafted features, that provide high-level spatial information, and deeply-learnable features, that encode low-level global meaning, in an end-to-end model that comprises all the relevant factors of our proposals so far and highlights the complementary behavior we first assumed for both descriptions by presenting the very best results, in quantitative and qualitative ways, we could ever achieve.

Moreover, part of the evaluation of deep learning features conducted abroad at the Ostbayerische Technische Hochschule (OTH) Regensburg, led by the co-supervisor of the project, Professor Christoph Palm, and supported by ReMIC research team showed to be very fruitful, with several research papers proposed and executed, focused on the interpretation of deep le-

arning's behavior and video evaluation [Ebigbo et al. 2019, Ebigbo et al. 2020, Mendel et al. 2020].

13.1 Works developed during the study period

Table 13.1 presents the works produced during the study period, while Table 13.2 describes the schedule adopted during the study.

13.2 Future Works

In regard to the notable impact of such research, with remarkable production potential as the presented Chapters has shown, we propose its continuation by assessing new approaches related to the evaluation of esophageal cancer in endoscopic examinations:

- First, a semantic segmentation based on siamese neural nets is proposed to detect, in spatial means, the cancerous regions in BE samples.
- Second, we propose to extend or work of assessing the deepness of neural models by employing Neural Architecture Search techniques in the task of defining the best architectures for the correct identification of cancer in the esophagus area.
- Finally, we aim to evaluate the degree of ill tissue by employing the Few-Shot Learning. This will enable the correct definition of several cancer levels in the esophagus, even with the lack of data for solving such a problem.

Acknowledgments

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) - processes #2017/04847-9 and #2019/08605-5.

Name	Type	Qualis	Year	Status
Barrett's Esophagus Analysis Using SURF Features [Souza Jr. et al. 2017]	Conference	A3	2017	Published
Barrett's Esophagus Identification Using Optimum-Path Forest [Souza Jr. et al. 2017]	Conference	A3	2017	Published
A Survey on Barrett's Esphagus Analysis Using Machine Learning [Souza Jr. et al. 2018]	Journal	A2	2018	Published
Barrett's Esophagus Analysis Using Color Co-occurrence Matrices [Souza Jr. et al. 2018]	Conference	A3	2017	Published
Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma [Ebigbo et al. 2019]	Journal	N/A	2018	Published
Learning Visual Representations with Optimum-Path Forest and its Applications to Barrett's Esophagus and Adenocarcinoma Diagnosis [Souza Jr. et al. 2019]	Journal	A1	2019	Published
Barrett's Esophagus Analysis Using Infinity Restricted Boltzmann Machines [Passos et al. 2019]	Journal	A2	2018	Published
Artificial Intelligence in Gastrointestinal Endoscopy: a Review	Journal	A1	2019	Submitted
Real-time use of artificial intelligence in the evaluation of cancer in Barrett's oesophagus [Ebigbo et al. 2020]	Journal	N/A	2019	Published
Semi-Supervised Segmentation based on Error-Correcting Supervision [Mendel et al. 2020]	Conference	A1	2020	Published
Assisting Barrett's Esophagus Identification Using Endoscopic Data Augmentation Based on Generative Adversarial Networks [Souza Jr. et al. 2020]	Journal	A2	2020	Published
Fine-tuning Generative Adversarial Networks Using Metaheuristics: A Case Study [Souza Jr. et al. 2021] on Barrett's Esophagus Identification	Conference	A3	2021	Published
Convolutional Neural Networks for the Evaluation of Cancer in Barrett's esophagus: Explainable AI to lighten up the Black-Box [Souza Jr. et al. 2021]	Journal	A2	2020	Published
Layer-Selective Deep Representation to Improve Esophageal cancer classification	Journal	A1	2021	Submitted
DeepCraftFuse: Handcrafted and Deeply Learnable Features Work Better Together for Esophageal Cancer Detection in Patients with Barrett's Esophagus	Journal	A1	2021	Submitted

Table 13.1: Works developed during the study period.

Features	Techniques	Period
Handcrafted Features Evaluation	SURF features + SVM classifier (with BoVW using SVM)	From 2017/1 to 2018/2
	SURF and SIFT features + OPF, SVM and Bayes	
	AKAZE, SURF and SIFT features + OPF, SVM and Bayes	
	Color co-occurrence matrix features + OPF, SVM and Bayes classifiers (with BoVW using PCA)	
	SURF and SIFT feature optimization (using iRBM) + OPF, SVM and Bayes classifiers	
Deeply-Learnable Features	CNN evaluation + GAN augmented data + GAN-hyperparameter fine-tuning	From 2019/1 to 2019/2
	CNN + Semi-supervised segmentation (cancerous automatic segmentation)	From 2019/2 to 2020/1
	Pixel-wise Explanable Artificial Inteligence evaluation	2020/1
	Layer-wise Explanable Artificial Inteligence evaluation	From 2020/2 to 2021/1
	Semantic-segmentation of cancerous tissues based on symetric CNN classifiers	2021/2
Feature Fusion	Handcrafted + deeply learnable features fusion method	From 2021/2 to 2022/1

Table 13.2: Schedule of the research.

REFERENCES

Abrams, J. A. et al. Adherence to biopsy guidelines for barrett's esophagus surveillance in the community setting in the united states. *Clinical Gastroenterology and Hepatology*, v. 7, n. 7, p. 736–742, 2009. ISSN 1542-3565.

AFONSO, L. et al. Automatic visual dictionary generation through optimum-path forest clustering. In: *2012 19th IEEE International Conference on Image Processing*. [S.l.: s.n.], 2012. p. 1897–1900. ISSN 1522-4880.

Aghamohammadi, M. et al. Predicting heart attack through explainable artificial intelligence. In: *International Conference on Computational Science (ICCS)*. [S.l.: s.n.], 2019. p. 633–645. ISSN 978-3-030-22741-8.

Alcantarilla, P.; Nuevo, J.; Bartoli, A. Fast explicit diffusion for accelerated features in nonlinear scale spaces. In: *British Machine Vision Conference (BMVC)*. [S.l.: s.n.], 2013. p. 13.1–13.11.

Allène, C. et al. Some links between extremum spanning forests, watersheds and min-cuts. *Image and Vision Computing*, v. 28, n. 10, p. 1460–1471, 2010. ISSN 0262-8856. Image Analysis and Mathematical Morphology.

Almond, L. M.; Barr, H. Advanced endoscopic imaging in barrett's oesophagus. *International Journal of Surgery*, v. 10, n. 5, p. 236 – 241, 2012. ISSN 1743-9191.

Alyafi, B.; Diaz, O.; Marti, R. DCGANs for realistic breast mass augmentation in x-ray mammography. In: *Medical Imaging 2020: Computer-Aided Diagnosis*. [S.l.: s.n.], 2020. v. 11314, p. 473 – 480.

Amorim, W. P. et al. Improving semi-supervised learning through optimum connectivity. *Pattern Recognition*, v. 60, n. C, p. 72—85, 2016. ISSN 0031-3203.

Atasoy, S. et al. Endoscopic video manifolds for targeted optical biopsy. *IEEE Transactions on Medical Imaging*, v. 31, n. 3, p. 637–653, 2012. ISSN 1558-254X.

Barredo-Arrieta, A. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, v. 58, p. 82–115, 2020. ISSN 1566-2535.

Bay, H. et al. Speeded-up robust features (SURF). *Computer Visision and Image Understanding*, v. 110, n. 3, p. 346–359, 2008. ISSN 1077-3142.

Boschetto, D.; Gambaretto, G.; Grisan, E. Automatic classification of endoscopic images for premalignant conditions of the esophagus. In: *Proceedings of SPIE Medical Imaging 2016: Biomedical Applications in Molecular, Structural, and Functional Imaging*. [S.l.: s.n.], 2016. v. 9788, p. 1–6.

Brock, A.; Donahue, J.; Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *CoRR*, abs/1809.11096, 2018.

Cao, Y. et al. Recent advances of generative adversarial networks in computer vision. *IEEE Access*, v. 7, p. 14985–15006, 2018. ISSN 2169-3536.

Cassel, C.; Jameton, A. Dementia in the elderly: an analysis of medical responsibility. *Annals of Intern Medicine*, v. 94, n. 6, p. 802–807, June 1981. ISSN 0003-4819.

Chan, D. K. et al. Breath testing for barrett's esophagus using exhaled volatile organic compound profiling with an electronic-nose device. *Gastroenterology*, p. 1–9, 2016.

Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, v. 2, n. 3, p. 1–27, 2011. ISSN 2157-6904.

Civicioglu, P. Backtracking search optimization algorithm for numerical optimization problems. *Applied Mathematics and Computation*, v. 219, n. 15, p. 8121–8144, 2013. ISSN 0096-3003.

Codella, N. C. F. et al. Collaborative human-ai (CHAI): evidence-based interpretable melanoma classification in dermoscopic images. *CoRR*, abs/1805.12234, 2018.

Côté, M.-A.; Larochelle, H. An infinite restricted boltzmann machine. *Neural computation*, v. 28, n. 7, p. 1265–1288, 2016. ISSN 0899-7667.

Csurka, G. et al. Visual categorization with bags of keypoints. In: *Proceedings of the Workshop on Statistical Learning in Computer Vision*. [S.l.: s.n.], 2004. p. 1–22.

Curvers, W. L.; Bergman, J. J. A new paradigm shift in endoscopy: From interpretation to automated image analysis? *Gastrointestinal Endoscopy*, v. 83, n. 1, p. 115–116, 2016.

Davies, R.; Twining, C.; Taylor, C. *Statistical Models of Shape: Optimisation and Evaluation*. 1. ed. [S.l.]: Springer Publishing Company, Incorporated, 2008. ISSN 978-1-84800-138-1.

de Groof, A. J. et al. Deep-learning system detects neoplasia in patients with Barrett's esophagus with higher accuracy than endoscopists in a Multistep Training and Validation study with benchmarking. *Gastroenterology*, v. 158, n. 4, p. 915–929.e4, 2020. ISSN 0016-5085.

Dent, J. Barrett's esophagus: a historical perspective, an update on core practicalities and predictions on future evolutions of management. *Journal of Gastroenterology and Hepatology*, v. 26, p. 11–30, 2011.

Diaz-Pinto, A. et al. Retinal image synthesis and semi-supervised learning for glaucoma assessment. *IEEE Transactions on Medical Imaging*, v. 38, n. 9, p. 2211–2218, 2019. ISSN 1558-254X.

Dickson, I. Deep-learning ai for neoplasia detection in barrett oesophagus. *Nature Reviews Gastroenterology & Hepatology*, v. 17, p. 66–67, 2019. ISSN 1759-5053.

Doman, K.; Konishi, T.; Mekada, Y. Lesion Image Synthesis using DCGANs for Metastatic Liver Cancer Detection. In: *Deep Learning in Medical Image Analysis*. [S.l.: s.n.], 2020. p. 95–106. ISBN 978-3-030-33127-6.

Doshi-Velez, F.; Kim, B. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv: Machine Learning*, 2017.

Du, Y. et al. DCGAN Based Data Generation for Process Monitoring. In: *2019 IEEE 8th Data Driven Control and Learning Systems Conference (DDCLS)*. [S.l.: s.n.], 2019. p. 410–415. ISBN 978-1-7281-1454-5.

Duchi, J.; Hazan, E.; Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, v. 12, n. null, p. 2121–2159, 2011. ISSN 1532-4435.

Ebigbo, A. et al. Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma. *Gut*, v. 68, n. 7, p. 1143–1145, 2019. ISSN 0017-5749.

Ebigbo, A. et al. Real-time use of artificial intelligence in the evaluation of cancer in Barrett's oesophagus. *Gut*, v. 69, n. 4, p. 615–616, 2020. ISSN 0017-5749.

Ebigbo, A. et al. Endoscopic prediction of submucosal invasion in barrett's cancer with the use of artificial intelligence: A pilot study. *Endoscopy*, v. 53, n. 9, p. 878–883, 2020. ISSN 0013-726X.

Ellis, R. et al. Impact of hybrid supervision approaches on the performance of artificial intelligence for the classification of chest radiographs. *Computers in Biology and Medicine*, v. 120, p. 103699, 2020. ISSN 0010-4825.

Falcão, A. X.; Stolfi, J.; Lotufo, R. A. The image foresting transform: theory, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 26, n. 1, p. 19–29, 2004. ISSN 1939-3539.

Fei-Fei, L.; Perona, P. A bayesian hierarchical model for learning natural scene categories. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. [S.l.: s.n.], 2005. v. 2, p. 524–531. ISSN 1063-6919.

Fiore, U. et al. Network anomaly detection with the restricted boltzmann machine. *Neurocomputing*, v. 122, p. 13–23, 2013. ISSN 0925-2312.

Frid-Adar, M. et al. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, v. 321, p. 321 – 331, 2018. ISSN 0925-2312.

Fujishiro, M. et al. Endoscopic submucosal dissection of esophageal squamous cell neoplasms. *Clinical Gastroenterology and Hepatology*, v. 4, n. 6, p. 688 – 694, 2006. ISSN 1542-3565.

Geem, Z. W. *Music-Inspired Harmony Search Algorithm: Theory and Applications*. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2009. ISSN 1860-949X.

Gehrung, M. et al. Triage-driven diagnosis of barrett's esophagus for early detection of esophageal adenocarcinoma using deep learning. *Nature Medicine*, v. 27, p. 833–841, 2021. ISSN 1546-170X.

Geifman, Y.; El-Yaniv, R. Selective classification for deep neural networks. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. [S.l.: s.n.], 2017. p. 4885–4894. ISBN 9781510860964.

Georgakopoulos, S. V. et al. Weakly-supervised convolutional learning for detection of inflammatory gastrointestinal lesions. *2015 IEEE International Conference on Imaging Systems and Techniques (IST)*, p. 510–514, 2016.

Ghatwary, N.; Ahmed, A.; Ye, X. Automated detection of barrett's esophagus using endoscopic images: A survey. In: *Proceedings os Medical Image Understanding and Analysis: 21st Annual Conference (MIUA)*. [S.l.: s.n.], 2017. p. 897–908. ISBN 978-3-319-60964-5.

Ghatwary, N.; Ye, X.; Zolgharni, M. Esophageal abnormality detection using densenet based faster r-cnn with gabor features. *IEEE Access*, v. 7, p. 84374–84385, 2019. ISSN 2169-3536.

Ghatwary, N.; Zolgharni, M.; Ye, X. Early esophageal adenocarcinoma detection using deep learning methods. *International Journal of Computer Assisted Radiology and Surgery*, v. 14, p. 611–621, 2019.

Gilpin, L. H. et al. Explaining explanations: An approach to evaluating interpretability of machine learning. *CoRR*, abs/1806.00069, 2018.

González-Castro, V. et al. Automatic rating of perivascular spaces in brain mri using bag of visual words. In: ____. *Proceedings of 13th International Conference Image Analysis and Recognition, in Memory of Mohamed Kamel*. [S.l.]: Springer International Publishing, 2016. (ICIAR, v. 9730), p. 642–649.

Goodfellow, I. J. et al. Generative adversarial nets. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14)*. [S.l.: s.n.], 2014. v. 2, p. 2672–2680.

Gotlieb, C. C.; Kreyszig, H. E. Texture descriptors based on co-occurrence matrices. *Computer Vision, Graphics, and Image Processing*, v. 51, n. 1, p. 70–86, 1990. ISSN 0734-189X.

Gu, D.; Su, K.; Zhao, H. A case-based ensemble learning system for explainable breast cancer recurrence prediction. *Artificial Intelligence in Medicine*, v. 107, p. 101858, 2020. ISSN 933–3657.

Hacker, P. et al. Explainable AI under contract and tort law: Legal incentives and technical challenges. *Artificial Intelligence and Law*, v. 28, p. 1–25, 2020.

Han, C. et al. Learning More with Less: Conditional PGGAN-based Data Augmentation for Brain Metastases Detection using Highly-Rough Annotation on MR Images. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*. [S.l.]: Association for Computing Machinery, 2019. p. 119–127. ISBN 9781450369763.

Han, C. et al. Combining Noise-to-Image and Image-to-Image GANs: Brain MR Image Augmentation for Tumor Detection. *IEEE Access*, v. 7, p. 156966–156977, 2019. ISSN 2169-3536.

Han, C. et al. Infinite Brain MR Images: PGGAN-Based Data Augmentation for Tumor Detection. In: *Neural Approaches to Dynamics of Signal Exchanges*. [S.l.]: Springer Singapore, 2020. v. 151, p. 291–303. ISSN 2190-3018.

Haralick, R.; Shanmugam, K.; Dinstein, I. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3, n. 6, p. 610–621, 1973. ISSN 0018-9472.

Hassan, A. R.; Haque, M. A. Computer-aided gastrointestinal hemorrhage detection in wireless capsule endoscopy videos. *Computers in Biology and Medicine*, v. 122, n. 3, p. 341–353, 2015. ISSN 0169-2607.

Hauta-Kasari, M. et al. Multi-spectral texture segmentation based on the spectral cooccurrence matrix. *Pattern Analysis and Applications*, v. 2, p. 275–284, 1999.

He, K. et al. Deep residual learning for image recognition. In: *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 770–778. ISSN 1063-6919.

Heusel, M. et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. [S.l.: s.n.], 2017. p. 6629–6640. ISBN 9781510860964.

Hinton, G. E.; Osindero, S.; Teh, Y.-W. A fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, v. 18, n. 7, p. 1527–1554, 2006. ISSN 0899-7667.

Holzinger, A. et al. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

Hong, J.; Park, B.-y.; Park, H. Convolutional neural network classifier for distinguishing barrett's esophagus and neoplasia endomicroscopy images. In: *Proceedings of 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. [S.l.: s.n.], 2017. p. 2892–2895. ISSN 1558-4615.

Hopkins, J. Barrett's esophagus: Introduction. *Gastroenterology & Hepatology*, 2008.

Horie, Y. et al. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. *Gastrointestinal Endoscopy*, v. 89, n. 1, p. 25–32, 2019.

Hou, W. et al. Early neoplasia identification in barrett's esophagus via attentive hierarchical aggregation and self-distillation. *Medical Image Analysis*, v. 72, p. 102092, 2021. ISSN 1361-8415.

Huo, Q.; Tang, G.; Zhang, F. Particle swarm optimization for great enhancement in semi-supervised retinal vessel segmentation with generative adversarial networks. In: *Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting*. [S.l.: s.n.], 2019. p. 112–120.

Iandola, F. N. et al. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and < 1mb model size. *CoRR*, abs/1602.07360, 2016.

Ilgner, J. et al. Colour texture analysis for quantitative laryngoscopy. *Acta Oto-Laryngologica*, v. 123, n. 6, p. 730–734, 2003.

Imai, S.; Kawai, S.; Nobuhara, H. Stepwise pathnet: a layer-by-layer knowledge-selection-based transfer learning algorithm. *Scientific Reports*, v. 10, n. 8132, 05 2020. ISSN 2045-2322.

Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML'15)*. [S.l.: s.n.], 2015. v. 37, p. 448–456. ISSN 0717-6163.

Ismail, N. H. et al. Multivariate multi-step deep learning time series approach in forecasting parkinson's disease future severity progression. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB'19)*. [S.l.: s.n.], 2019. p. 383–389. ISBN 9781450366663.

Jain, A.; Healey, G. A multiscale representation including opponent color features for texture recognition. *IEEE Transactions on Image Processing*, v. 7, n. 1, p. 124–128, 1998. ISSN 1941-0042.

Jin, D. et al. CT-Realistic Lung Nodule Simulation from 3D Conditional Generative Adversarial Networks for Robust Lung Segmentation. *CoRR*, abs/1806.04051, 2018.

Johnston, M. H. et al. Cryoablation of barrett's esophagus: A pilot study. *Gastrointestinal Endoscopy*, v. 62, p. 842–848, 2005.

Kallianos, K. et al. How far have we come? artificial intelligence for chest radiograph interpretation. *Clinical Radiology*, v. 74, n. 5, p. 338–345, 2019. ISSN 0009-9260.

Kandemir, M. et al. Digital pathology: Multiple instance learning can detect barrett's cancer. In: *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. [S.l.: s.n.], 2014. p. 1348–1351.

Karras, T. et al. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *CoRR*, abs/1710.10196, 2017.

Khojasteh, P. et al. Exudate detection in fundus images using deeply-learnable features. *Computers in Biology and Medicine*, v. 104, p. 62–69, 2019. ISSN 0010-4825.

Kim, D. D. et al. Generating Pedestrian Training Dataset Using DCGAN. In: *Proceedings of the 2019 3rd International Conference on Advances in Image Processing (ICAIP'19)*. [S.l.: s.n.], 2019. p. 1–4. ISBN 9781450376754.

Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Klomp, S. et al. Evaluation of image features and classification methods for barrett's cancer detection using vle imaging. *Proceedings of SPIE Medical Imaging 2017*, v. 10134, p. 101340D, 2017.

Koh, J. E. W. et al. Automated retinal health diagnosis using pyramid histogram of visual words and fisher vector techniques. *Computers in Biology and Medicine*, v. 92, p. 204–209, 2018. ISSN 0010-4825.

Koulaouzidis, A.; Iakovidis, D. K. "kid: Koulaouzidis-iakovidis database for capsule endoscopy.". 2015. Disponível em: <http://is-innovation.eu/kid/>.

Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, v. 60, n. 6, p. 84–90, 2017. ISSN 0001-0782.

Lagergren, J.; Lagergren, P. Oesophageal cancer. *BMJ*, BMJ Publishing Group Ltd, v. 341, 2010. ISSN 0959-8138.

Lamy, J.-B. et al. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artif Intell Med.*, v. 94, p. 42–53, 2019. ISSN 0933-3657.

Lapuschkin, S. et al. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nat Commun.*, v. 10, n. 1096, 2019. ISSN 2041-1723.

Larochelle, H.; Bengio, Y. Classification using discriminative restricted boltzmann machines. In: *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*. [S.l.: s.n.], 2008. p. 536–543. ISBN 9781605582054.

Larochelle, H. et al. Learning algorithms for the classification restricted boltzmann machine. *The Journal of Machine Learning Research*, v. 13, n. 1, p. 643–669, 2012. ISSN 1532-4435.

Lepage, C.; RACHET, B.; JOOSTE, V. Continuing rapid increase in esophageal adenocarcinoma in england and wales. *The American Journal of Gastroenterology*, v. 103, n. 11, p. 2694–2699, 2008.

Li, B.; Meng, M. Q.-H. Texture analysis for ulcer detection in capsule endoscopy images. *Image and Vision Computing*, v. 27, n. 9, p. 1336–1342, 2009. ISSN 0262-8856.

Li, C. et al. Multiple instance learning for computer aided detection and diagnosis of gastric cancer with dual-energy ct imaging. *Journal of Biomedical Informatics*, v. 57, p. 358–368, 2015. ISSN 1532-0464.

Lovie, A. D. Who discovered Spearman's rank correlation? *British Journal of Mathematical and Statistical Psychology*, v. 48, n. 2, p. 255–269, 1995.

Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, Kluwer Academic Publishers, Hingham, MA, USA, v. 60, n. 2, p. 91–110, 2004. ISSN 0920-5691.

Ma, B. et al. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Computers in Biology and Medicine*, v. 121, p. 103761, 2020. ISSN 0010-4825.

Ma, X.; Lv, S. Financial credit risk prediction in internet finance driven by machine learning. *Neural Computing and Applications*, v. 31, n. 12, p. 8359–8367, 2019.

McHugh, M. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, v. 22, n. 3, p. 276–82, 10 2012.

Mendel, R. et al. Barrett's esophagus Analysis using Convolutional Neural Networks. In: *Bildverarbeitung für die Medizin 2017*. [S.l.: s.n.], 2017. p. 80–85. ISBN 978-3-662-54345-0.

Mendel, R. et al. Semi-supervised segmentation based on error-correcting supervision. In: *Proceedngs of European Conference on Computer Vision (ECCV'20)*. [S.l.: s.n.], 2020. p. 141–157. ISBN 978-3-030-58526-6.

MICCAI 2015: 18th International Conference. *Medical Image Computing and Computer Assisted Interventions*, 2015. Disponível em: <https://endovissub-barrett.grand-challenge.org/>.

Middel, L.; Palm, C.; Erdt, M. Synthesis of Medical Images using Gans. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures*. [S.l.: s.n.], 2019. p. 125–134. ISBN 978-3-030-32689-0.

Mirniaharikandehei, S. et al. Applying a random projection algorithm to optimize machine learning model for predicting peritoneal metastasis in gastric cancer patients using ct images. *Computer Methods and Programs in Biomedicine*, v. 200, p. 105937, 2021. ISSN 0169-2607.

Moncada-Torres, A. et al. Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival. *Scientific Reports*, v. 11, p. 6968, 2021. ISSN 2045-2322.

Montufar, G.; Ay, N. Refinements of universal approximation results for deep belief networks and restricted boltzmann machines. *Neural Computation*, v. 23, n. 5, p. 1306–1319, 2011. ISSN 1530-888X.

Muldoon, T. et al. Evaluation of quantitative image analysis criteria for the high-resolution microendoscopic detection of neoplasia in barrett's esophagus. *Journal of Biomedical Optics*, v. 15, p. 026027, 2010.

Nakamura, R. Y. M. et al. Nature-inspired framework for hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing*, v. 52, n. 4, p. 2126–2137, 2014. ISSN 0196-2892.

Nancarrow, D. J. et al. Whole genome expression array profiling highlights differences in mucosal defense genes in barrett's esophagus and esophageal adenocarcinoma. *PLOS ONE*, v. 6, n. 7, p. 1–15, 2011.

Ohmori, M. et al. Endoscopic detection and differentiation of esophageal lesions using a deep neural network. *Gastroitestinal Edoscopy*, v. 91, n. 2, p. 301–309.e1, 2019.

OpenCV. *Open Source Computer Vision Library*. 2015. `https://github.com/itseez/opencv`.

Overholt, B. F.; PANJEHPOUR, M.; HALBERG, D. L. Photodynamic therapy for barrett's esophagus with dysplasia and/or early stage carcinoma: long-term results. *Gastrointestinal Endoscopy*, v. 58, n. 2, p. 183–188, 2003.

Palm, C. Color texture classification by integrative co-occurrence matrices. *Pattern Recognition*, v. 37, n. 5, p. 965–976, 2004. ISSN 0031-3203.

Palm, C.; Lehmann, T. M. Classification of color textures by gabor filtering. *Medicine Graphics & Vision*, v. 11, n. 2/3, p. 195–219, 2002. ISSN 1230-0535.

Pan, W. et al. Identification of barrett's esophagus in endoscopic images using deep learning. *BMC Gastroenterology*, v. 21, n. 479, 12 2021. ISSN 1471-230X.

Papa, J. P.; Falcão, A. X.; Suzuki, C. T. N. Supervised pattern classification based on optimum-path forest. *International Journal of Imaging Systems and Technology*, v. 19, n. 2, p. 120–131, 2009. ISSN 0899-9457.

Papa, J. P.; Falcão, C. T. N. S. A. X. LibOPF: A library for the design of optimum-path forest classifiers. Software version 2.1 available at http://www.ic.unicamp.br/ afalcao/libopf/index.html.

Papa, J. P. et al. Efficient supervised optimum-path forest classification for large datasets. *Pattern Recognition*, v. 45, n. 1, p. 512–520, 2012. ISSN 0031-3203.

Papa, J. P.; Fernandes, S. E. N.; Falcão, A. X. Optimum-path forest based on k-connectivity: Theory and applications. *Pattern Recognition Letters*, v. 87, n. 1, p. 117–126, 2017. ISSN 0167-8655.

Papa, J. P.; Rocha, A. R. Image categorization through optimum path forest and visual words. In: *Proceedings of the 18th IEEE International Conference on Image Processing*. [S.l.: s.n.], 2011. p. 3525–3528. ISSN 1522-4880.

Papa, J. P. et al. On the model selection of bernoulli restricted boltzmann machines through harmony search. In: *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'15)*. [S.l.: s.n.], 2015. p. 1449–1450. ISBN 9781450334884.

Papa, J. P. et al. Model selection for discriminative restricted boltzmann machines through meta-heuristic techniques. *Journal of Computational Science*, v. 9, p. 14–18, 2015. ISSN 1877-7503.

Papa, J. P. et al. Libopt: An open-source platform for fast prototyping soft optimization techniques. *arXiv*, arXiv:1704.05174, 2017.

Papa, J. P.; Scheirer, W.; Cox, D. D. Fine-tuning deep belief networks using harmony search. *Applied Soft Computing*, v. 46, p. 875–885, 2016. ISSN 1568-4946.

Papa, J. P. et al. Optimum path forest classifier applied to laryngeal pathology detection. In: *2008 15th International Conference on Systems, Signals and Image Processing*. [S.l.: s.n.], 2008. p. 249–252. ISSN 2157-8672.

Paschos, G. Perceptually uniform color spaces for color texture analysis: An empirical evaluation. *IEEE Transactions on Image Processing*, v. 10, n. 6, p. 932–937, 2001. ISSN 1941-0042.

Passos, L. A. et al. Barrett's esophagus analysis using infinity restricted boltzmann machines. *Journal of Visual Communication and Image Representation*, v. 59, p. 475–485, 2019. ISSN 1047-3203.

Passos, L. A.; Papa, J. P. Fine-tuning infinity restricted boltzmann machines. In: *Electronic Proceedings of the 30th Conference on Graphics, Patterns and Images (SIBGRAPI'17)*. [S.l.: s.n.], 2017. ISSN 2377-5416.

Passos, L. A.; Papa, J. P. A metaheuristic-driven approach to fine-tune deep boltzmann machines. *Applied Soft Computing*, v. 97, p. 105717, 2019. ISSN 1568-4946.

Passos, L. A. et al. Parkinson disease identification using residual networks and optimum-path forest. In: *2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. [S.l.: s.n.], 2018. p. 000325–000330. ISBN 978-1-5386-4640-3.

Passos, L. A.; Rodrigues, D.; Papa, J. P. Quaternion-based backtracking search optimization algorithm. In: *2019 IEEE Congress on Evolutionary Computation (CEC)*. [S.l.: s.n.], 2019. p. 3014–3021. ISBN 978-1-7281-2153-6.

Pech, O. et al. Confocal laser endomicroscopy for in vivo diagnosis of early squamous cell carcinoma in the esophagus. *Clinical Gastroenterology and Hepatology*, v. 6, n. 1, p. 89–94, 2008. ISSN 1542-3565.

Peng, X.; Gao, X.; Li, X. On better training the infinite Restricted Boltzmann Machines. *Machine Learning*, v. 107, n. 6, p. 943–968, 2018.

Peng, X. et al. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, v. 150, p. 109–125, 2016. ISSN 1077-3142.

Phoa, K. N. et al. Multimodality endoscopic eradication for neoplastic barrett oesophagus: results of an european multicentre study (euro-ii). *Gut*, v. 65, n. 4, p. 555–562, 2016. ISSN 0017-5749.

Pisani, R. J. et al. Toward satellite-based land cover classification through optimum-path forest. *IEEE Transactions on Geoscience and Remote Sensing*, v. 52, n. 10, p. 6075–6085, 2014. ISSN 1558-0644.

Preece, A. D. et al. Stakeholders in Explainable AI. *CoRR*, abs/1810.00184, 2018.

Pu, W. et al. Targeted bisulfite sequencing identified a panel of dna methylation-based biomarkers for esophageal squamous cell carcinoma (escc). *Clinical Epigenetics*, v. 9, n. 1, p. 129, 2017. ISSN 1868-7083.

Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv*, abs/1511.06434, 2015.

Rajan, P. et al. Automated diagnosis of barrett's esophagus with endoscopic images. In: ____. *World Congress on Medical Physics and Biomedical Engineering, September 7 - 12, 2009, Munich, Germany: Vol. 25/4 Image Processing, Biosignal Processing, Modelling and Simulation, Biomechanics*. [S.l.]: Springer, 2010. p. 2189–2192. ISBN 978-3-642-03882-2.

Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. New York, NY, USA: [s.n.], 2016. p. 1135–1144. ISBN 9781450342322.

Ribeiro, P. B. et al. Unsupervised breast masses classification through optimum-path forest. In: *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*. [S.l.: s.n.], 2015. p. 238–243. ISSN 2372-9198.

Riel, S. V. et al. Automatic detection of early esophageal cancer with cnns using transfer learning. *IEEE International Conference on Image Processing (ICIP)*, p. 1383–1387, 2018. ISSN 2381-8549.

Rocha, L. M.; Cappabianco, F. A. M.; Falcão, A. X. Data clustering as an optimum-path forest problem with applications in image analysis. *International Journal of Imaging Systems and Technology*, v. 19, n. 2, p. 50–68, 2009.

Rodrigues, D. et al. Adaptive improved flower pollination algorithm for global optimization. In: *Nature-Inspired Computation in Data Mining and Machine Learning*. [S.l.]: Springer, 2020. p. 1–21. ISBN 978-3-030-28553-1.

Rodrigues, D.; Yang, X. S.; Papa, J. P. Fine-tuning deep belief networks using cuckoo search. In: *Bio-Inspired Computation and Applications in Image Processing*. [S.l.]: Academic Press, 2016. p. 47–59. ISBN 978-0-12-804536-7.

Rodrigues, D. et al. Recent advances in swarm intelligence and evolutionary computation. In: ____. [S.l.]: Springer International Publishing, 2015. cap. Binary Flower Pollination Algorithm and Its Application to Feature Selection, p. 85–100. ISBN 978-3-319-13826-8.

Rodriguez-Diaz, E.; Singh, S. K. Computer-assisted interpretation of the nice criteria for colorectal polyps using near-focus narrow-band imaging. *Gastroenterology*, v. 150, p. S–434, 2016.

Rosa, G. H. et al. Learning parameters in deep belief networks through firefly algorithm. In: ____. *Artificial Neural Networks in Pattern Recognition: 7th IAPR TC3 Workshop, ANNPR*. [S.l.]: Springer International Publishing, 2016. p. 138–149. ISBN 978-3-319-46182-3.

Rosa, G. H. et al. Fine-tuning convolutional neural networks using harmony search. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. [S.l.]: Springer International Publishing, 2015. v. 9423, p. 683–690. ISBN 978-3-319-25750-1.

Rosenfeld, A. et al. Using data mining to help detect dysplasia: Extended abstract. *2014 IEEE International Conference on Software Science, Technology and Engineering*, p. 65–66, 2014.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, v. 1, n. 5, p. 206–215, 2019. ISSN 2522-5839.

Russakovsky, O. et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, v. 115, n. 3, p. 211–252, 2015.

Sabol, P. et al. Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. *Journal of Biomedical Informatics*, v. 109, p. 103523, 2020. ISSN 1532-0464.

Salakhutdinov, R.; Hinton, G. E. An efficient learning procedure for deep boltzmann machines. *Neural Computation*, v. 24, n. 8, p. 1967–2006, 2012. ISSN 0899-7667.

Salakhutdinov, R.; Mnih, A.; Hinton, G. Restricted boltzmann machines for collaborative filtering. In: *Proceedings of the 24th international conference on Machine learning*. [s.n.], 2007. p. 791–798. ISBN 9781595937933. Disponível em: <https://doi.org/10.1145/1273496.1273596>.

Sandfort, V. et al. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific Reports*, v. 9, p. 16884, 12 2019. ISSN 2045-2322.

Santana, M. et al. A novel siamese-based approach for scene change detection with applications to obstructed routes in hazardous environments. *IEEE Intelligent Systems*, v. 35, n. 1, p. 44–53, 2019. ISSN 1941-1294.

Sattlecker, M.; Stone, N.; Bessant, C. Current trends in machine-learning methods applied to spectroscopic cancer diagnosis. *TrAC Trends in Analytical Chemistry*, v. 59, p. 17–25, 2014. ISSN 0165-9936.

Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks*, v. 61, p. 85–117, 2015. ISSN 0893-6080.

Seguí, S. et al. Generic feature learning for wireless capsule endoscopy analysis. *Computers in Biology and Medicine*, v. 79, p. 163–172, 2016. ISSN 0010-4825.

Seibel, E. J. et al. Tethered capsule endoscopy, a low-cost and high-performance alternative technology for the screening of esophageal cancer and barrett's esophagus. *IEEE Transactions on Biomedical Engineering*, v. 55, n. 3, p. 1032–1042, 2008. ISSN 1558-2531.

Serpa-Andrade, L. et al. An approach based on fourier descriptors and decision trees to perform presumptive diagnosis of esophagitis for educational purposes. In: *2015 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*. [S.l.: s.n.], 2015. p. 1–5. ISBN 978-1-4673-7121-6.

Shaheen, N. J. et al. Radiofrequency ablation in barrett's esophagus with dysplasia. *The New England Journal of Medicine*, v. 360, p. 2277–2288, 2009. ISSN 1533-4406.

Sharma, P. et al. Development and validation of a classification system to identify high-grade dysplasia and esophageal adenocarcinoma in barrett's esophagus using narrow-band imaging. *Gastroenterology*, v. 150, n. 3, p. 591–598, 2016. ISSN 0016-5085.

Sharma, P. et al. White paper aga: Advanced imaging in barrett's esophagus. *Clinical Gastroenterology and Hepatology*, v. 13, n. 13, p. 2209 – 2218, 2015. ISSN 1542-3565.

Shi, J.; Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 8, p. 888–905, 2000. ISSN 0162-8828.

Shi, Y. Brain storm optimization algorithm. In: *Proceedings of the Second International Conference on Advances in Swarm Intelligence (ICSI'11)*. [S.l.: s.n.], 2011. v. 1, p. 303–309. ISBN 978-3-642-21515-5.

Shin, H. et al. Medical Image Synthesis for Data Augmentation and Anonymization using Generative Adversarial Networks. *CoRR*, abs/1807.10225, 2018.

Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features through Propagating Activation Differences. In: *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*. [S.l.: s.n.], 2017. v. 70, p. 3145–3153.

Silva, L. A. et al. Learning spam features using restricted boltzmann machines. *IADIS International Journal on Computer Science and Information Systems*, v. 11, n. 1, p. 99–114, 2016. ISSN 2047-4962.

Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR*, abs/1312.6034, 2014.

Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Song, K.; Kittler, J.; Petrou, M. Defect detection in random colour textures. *Image and Vision Computing*, v. 14, n. 9, p. 667 – 683, 1996. ISSN 0262-8856.

Souza Jr., L. A. et al. Learning visual representations with optimum-path forest and its applications to barrett's esophagus and adenocarcinoma diagnosis. *Neural Computing and Applications*, v. 32, 2019.

Souza Jr., L. A. et al. Barrett's esophagus Identification using Optimum-Path Forest. In: *Proceedings of the 30th Conference on Graphics, Patterns and Images (SIBGRAPI'17)*. [S.l.: s.n.], 2017. p. 308–314. ISSN 2377-5416.

Souza Jr., L. A. et al. Barrett's esophagus identification using color co-occurrence matrices. In: *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. [S.l.: s.n.], 2018. p. 166–173. ISSN 2377-5416.

Souza Jr., L. A. et al. Barrett's esophagus analysis using SURF features. In: *Bildverarbeitung für die Medizin 2017 (BVM'17)*. [S.l.: s.n.], 2017. p. 141–146. ISBN 978-3-662-54345-0.

Souza Jr., L. A. et al. Convolutional neural networks for the evaluation of cancer in barrett's esophagus: Explainable ai to lighten up the black-box. *Computers in Biology and Medicine*, v. 135, p. 104578, 2021. ISSN 0010-4825.

Souza Jr., L. A. et al. A survey on Barrett's esophagus analysis using machine learning. *Computers in Biology and Medicine.*, v. 96, p. 203–213, 2018. ISSN 0010-4825.

Souza Jr., L. A. et al. Assisting barrett's esophagus identification using endoscopic data augmentation based on generative adversarial networks. *Computers in Biology and Medicine*, p. 104029, 2020. ISSN 0010-4825.

Souza Jr., L. A. et al. Fine-tuning generative adversarial networks using metaheuristics. In: *Bildverarbeitung für die Medizin 2021 (BVM'21)*. [S.l.: s.n.], 2021. p. 205–210. ISBN 978-3-658-33198-6.

Spearman Rank Correlation Coefficient. In: THE Concise Encyclopedia of Statistics. New York, NY: Springer New York, 2008. p. 502–505. ISBN 978-0-387-32833-1.

Springenberg, J. et al. Striving for Simplicity: The All Convolutional Net. In: *ICLR (workshop track)*. [S.l.: s.n.], 2015.

Struyvenberg, M. R. et al. A computer-assisted algorithm for narrow-band-imaging-based tissue characterization in Barrett's esophagus. *Gastrointest Endosc.*, 2020. ISSN 0016-5107.

Suen, P.-H.; Healey, G. Modeling and classifying color textures using random fields in a random environment. *Pattern Recognition*, v. 32, n. 6, p. 1009 – 1017, 1999. ISSN 0031-3203.

Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017.

Suzuki, C. T. N. et al. Automatic segmentation and classification of human intestinal parasites from microscopy images. *IEEE Transactions on Biomedical Engineering*, v. 60, n. 3, p. 803–812, 2013. ISSN 1558-2531.

Swager, A.-F. et al. Feasibility of a computer algorithm for detection of early barrett's neoplasia using volumetric laser endomicroscopy. *Gastroenterology*, v. 150, n. 4, Supplement 1, p. S56, 2016.

Swager, A.-F. et al. Identification of volumetric laser endomicroscopy features predictive for early neoplasia in barrett's esophagus using high-quality histological correlation. *Gastrointestinal Endoscopy*, v. 85, n. 5, p. 918–926.e7, 2017. ISSN 0016–5107.

Swager, A.-F. et al. Computer-aided detection of early Barrett's neoplasia using volumetric laser endomicroscopy. *Gastrointestinal Endoscopy.*, v. 86, n. 5, p. 839–846, 2017. ISSN 0016-5107.

Szegedy, C. et al. Inception-v4, inception-resnet and the impact of residual connections on learning. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17)*. [S.l.: s.n.], 2017. p. 4278–4284.

Szegedy, C. et al. Going deeper with convolutions. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2015. p. 1–9. ISSN 1063-6919.

Tanaka, F. H. K. dos S.; Aranha, C. Data Augmentation Using GANs. *CoRR*, abs/1904.09135, 2019.

Tang, Z. et al. Interpretable classification of alzheimer's disease pathologies with a convolutional neural network pipeline. *Nature communications*, v. 10, n. 2173, p. 1–14, 2019. ISSN 2041-1723.

Taylor, G. W.; Hinton, G. E.; Roweis, S. T. Modeling human motion using binary latent variables. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2006. p. 1345–1352.

Tieleman, T. Training restricted Boltzmann machines using approximations to the likelihood gradient. In: *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*. [S.l.: s.n.], 2008. p. 1064–1071. ISBN 978-1-60558-205-4.

Tjoa, E.; GUAN, C. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. *CoRR*, abs/1907.07374, 2019.

Tonekaboni, S. et al. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In: *Proceedings of Machine Learning Research*. [S.l.: s.n.], 2019. v. 106, p. 359–380.

Toth, E.; BRATH, A. Multistep ahead streamflow forecasting: Role of calibration data in conceptual and neural network modeling. *Water Resources Research*, v. 43, W11405, p. 1–11, 11 2007.

Tsochatzidis, L. et al. Integrating segmentation information into cnn for breast cancer diagnosis of mammographic masses. *Computer Methods and Programs in Biomedicine*, v. 200, p. 105913, 2021. ISSN 0169-2607.

van de Wouwer, G. et al. Wavelet correlation signatures for color texture characterization. *Pattern Recognition*, v. 32, n. 3, p. 443–451, 1999. ISSN 0031-3203.

van der Putten, J. et al. Multi-stage domain-specific pretraining for improved detection and localization of Barrett's neoplasia: A comprehensive clinically validated study. *Artificial Intelligence in Medicine*, v. 107, p. 101914, 2020. ISSN 0933-3657.

van der Putten, J. et al. Deep principal dimension encoding for the classification of early neoplasia in Barrett's esophagus with volumetric laser endomicroscopy. *Comput Med Imag Grap.*, v. 80, p. 101701, 2020.

van der Sommen, F. et al. Computer-aided detection of early neoplastic lesions in Barret's esophagus. *Endoscopy*, v. 48, n. 7, p. 617–624, 2016.

van der Sommen, F. et al. Computer-aided detection of early cancer in the esophagus using hd endoscopy images. In: *Proceedings of The International Society for Optical Engineering (SPIE)*. [S.l.: s.n.], 2013. v. 86700.

van der Sommen, F. et al. Supportive automatic annotation of early esophageal cancer using local gabor and color features. *Neurocomputing*, v. 144, p. 92–106, 2014. ISSN 0925-2312.

van Riel, S. et al. Automatic detection of early esophageal cancer with CNNS using transfer learning. In: *Proceedings of 2018 IEEE International Conference on Image Processing (ICIP'2018)*. [S.l.: s.n.], 2018. p. 1383–1387. ISBN 978-1-4799-7062-9.

Veronese, E. et al. Hybrid patch-based and image-wide classification of confocal laser endomicroscopy images in barrett's esophagus surveillance. In: *2013 IEEE 10th International Symposium on Biomedical Imaging*. [S.l.: s.n.], 2013. p. 362–365. ISSN 1945-8452.

Wang, C. et al. SaliencyGAN: Deep Learning Semisupervised Salient Object Detection in the Fog of IoT. *IEEE Transactions on Industrial Informormatics*, v. 16, n. 4, p. 2667–2676, 2020. ISSN 1941-0050.

Wang, C. et al. TPSDicyc: Improved Deformation Invariant Cross-domain Medical image Synthesis. In: *Machine Learning for Medical Image Reconstruction*. [S.l.: s.n.], 2019. p. 245–254. ISBN 978-3-030-33843-5.

Wang, Z. et al. Novel optical coherence tomography image analysis reveals subsquamous glandular structures as strong predictors of poorer response to radiofrequency ablation in barrett's esophagus. *Gastroenterology*, v. 150, n. 4, p. S–434, 2016. ISSN https://doi.org/10.1016/S0016-5085(16)31507-4.

Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, v. 1, p. 80–83, 1945. ISSN 00994987.

Xie, Y.; GAO, G.; CHEN, X. A. Outlining the Design Space of Explainable Intelligent Systems for Medical Diagnosis. *CoRR*, abs/1902.06019, 2019.

Yang, G. et al. DAGAN: Deep De-Aliasing Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction. *IEEE Transactions on Medical Imaging*, v. 37, n. 6, p. 1310–1321, 2018. ISSN 1558-254X.

Yang, S.-S.; Karamanoglu, M.; He, X. Flower pollination algorithm: A novel approach for multiobjective optimization. *Engineering Optimization*, v. 46, n. 9, p. 1222–1237, 2014.

Yang, X.-S. Firefly algorithm, stochastic test functions and design optimisation. *International Journal Bio-Inspired Computing*, v. 2, n. 2, p. 78–84, 2010. ISSN 1758-0366.

Yang, X.-S.; Deb, S. Engineering optimisation by cuckoo search. *International Journal of Mathematical Modelling and Numerical Optimisation*, v. 1, n. 4, p. 330–343, 2010.

Yang, X.-S.; Gandomi, A. H. Bat algorithm: a novel approach for global engineering optimization. *Engineering Computations*, v. 29, n. 5, p. 464–483, 2012. ISSN 0264-4401.

Yoshida, H. et al. Automated histological classification of whole-slide images of gastric biopsy specimens. *Gastric Cancer*, p. 249–257, 2017.

Yosinski, J. et al. How transferable are features in deep neural networks? In: *Adv Neural Inf Process Syst*. [S.l.: s.n.], 2014. p. 3320–3328.

Yu, C. et al. Otsu's thresholding method based on gray level-gradient two-dimensional histogram. In: *Proceedings of 2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR 2010)*. [S.l.: s.n.], 2010. v. 3, p. 282–285. ISSN 1948-3422.

Yu, J.-s. et al. A hybrid convolutional neural networks with extreme learning machine for wce image classification. In: *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. [S.l.: s.n.], 2015. p. 1822–1827. ISBN 978-1-4673-9675-2.

Yu, S. et al. Deep De-aliasing for Fast Compressive Sensing MRI. *CoRR*, abs/1705.07137, 2017.

Zanini, R. A.; Colombini, E. L. Parkinson's Disease EMG Data Augmentation and Simulation with DCGANs and Style Transfer. *Sensors-Basel*, Multidisciplinary Digital Publishing Institute, v. 20, n. 9, p. 2605, 2020. ISSN 1424-8220.

Zeiler, M. D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In: *Proceedings of European Conference in Computer Vision (ECCV'14)*. [S.l.: s.n.], 2013. v. 8689. ISBN 978-3-319-10590-1.

Zhang, J.; Sanderson, A. C. Jade: adaptive differential evolution with optional external archive. *IEEE Transactions on evolutionary computation*, v. 13, n. 5, p. 945–958, 2009. ISSN 1941-0026.

Zhang, Z. et al. High-order graph matching kernel for early carcinoma eus image classification. *Multimedia Tools and Applications*, v. 75, n. 7, p. 3993–4012, 2016. ISSN 1573-7721.

Zhao, H. et al. Synthesizing retinal and neuronal images with Generative Adversarial Nets. *Med Image Anal.*, v. 49, p. 14 – 26, 2018. ISSN 1361-8415.

Zhou, T. et al. Multi-modal latent space inducing ensemble svm classifier for early dementia diagnosis with neuroimaging data. *Medical Image Analysis*, v. 60, p. 101630, 2020. ISSN 1361-8415.

Zhou, W. et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, v. 13, n. 4, p. 600–612, 2004. ISSN 1941-0042.

Zhou, Z.; Sun, Y.; Li, Y. Multi-instance learning by treating instances as non-i.i.d. samples. In: *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*. [S.l.: s.n.], 2009. p. 1249–1256. ISBN 978-1-60558-516-1.

Zopf, S. et al. Narrow-band imaging for the computer assisted diagnosis in patients with barrett's esophagus. *Gastrointestinal Endoscopy*, v. 69, n. 5, p. AB376, 2009. ISSN 0016-5107.

Zucco, C. et al. Explainable Sentiment Analysis with Applications in Medicine. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. [S.l.: s.n.], 2018. p. 1740–1747. ISBN 978-1-5386-5488-0.

GLOSSARY

***k*-NN** – *k-nearest neighbors*

A-KAZE – *Accelerated-KAZE features*

ACM – *Association for Computing Machinery*

ANN – *artificial neural network*

AUC – *area under curve*

BE – *Barrett's esophagus*

BING – *BE's International NBI Group*

BoVW – *bag-of-visual-words*

CAD – *computer-assisted diagnosis*

CE – *capsule endoscopy*

CLE – *confocal laser endomicroscopy*

CM – *co-occurrence matrix*

CNN – *Convolutional Neural Network*

CT – *computed tomography*

Citation-*k*-nn – *Citation-k-nearest neighborhood*

DL – *Deep Learning*

EAC – *esophageal adenocarcinoma*

EAC – *gray-scale co-occurrence matrix feature*

ELM – *extreme learning*

ESCC – *esophageal squamous cell carcinoma*

EUS – *endoscopic ultrasonography*

GAN – *Generative Adversarial Networks*

GM – *gastric metaplasia*

HD-WLE – *high-definition white light endoscopy*

HRME – *high-resolution microendoscopic*

IM – *intestinal metaplasia*

LBP – *Local Binary Pattern*

LDA – *linear discriminant analysis*

LOO-CV – *leave-one-out cross-validation*

LOPO-CV – *leave-one-patient-out cross-validation*

LSO-CV – *leave-some-out cross-validation*

MDPI – *Multidisciplinary Publishing Institute*

MIL – *multiple instance leraning*

MLP – *Multilayer Perceptron*

ML – *machine learning*

NBI – *narrow band imaging*

NF-NBI – *near-focus narrow-band imaging*

NGLCM – *Normalized Gray Level Co-occurrence Matrix*

OPF – *Optimum-Path Forest*

OTH – *Ostbayerische Technische Hochschule*

PCA – *Principal Component Analysis*

RFA – *radiofrequency ablation*

ROC – *Receiver Operating Characteristic*

ROIs – *regions of interest*

SCM – *Single-Channel co-occurrence matrix*

SGS – *subsquamous glandular structures*

SIFT – *Scale-Invariant Feature Transform*

SURF – *Speed-Up Robust Features*

SVM – *Support Vector Machines*

VLE – *volumetric laser endomicroscopy*

VOC – *volatile organic compounds*

WCE – *wireless capsule endoscopy*

WLE – *white light endoscopy*

XAI – *Explainable Artificial Intelligence*

iRBM – *infinte Restricted Boltzmann Machines*