3-2-2021

Color and morphological features extraction and nuclei classification in tissue samples of colorectal cancer

Sameer Akhtar Syed
University of Windsor

COLOR AND MORPHOLOGICAL FEATURES EXTRACTION AND NUCLEI CLASSIFICATION IN TISSUE SAMPLES OF COLORECTAL CANCER

By

Sameer Akhtar Syed

A Thesis
Submitted to the Faculty of Graduate Studies through the School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science
at the University of Windsor

Windsor, Ontario, Canada

2020

COLOR AND MORPHOLOGICAL FEATURES EXTRACTION AND NUCLEI CLASSIFICATION IN TISSUE SAMPLES OF COLORECTAL CANCER

by

Sameer Akhtar Syed

APPROVED BY:

M. Belalia
Department of Mathematics and Statistics

I. Ahmad
School of Computer Science

B. Boufama, Advisor
School of Computer Science

21 December 2020

Declaration of Originality

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

Cancer is an important public health problem and the third most leading cause of death in North America. Among the highest impact types of cancer are colorectal, breast, lung, and prostate. This thesis addresses the features extraction by using different artificial intelligence algorithms that provide distinct solutions for the purpose of Computer-Aided Diagnosis (CAD). For example, classification algorithms are employed in identifying histological structures, such as lymphocytes, cancer-cells nuclei and glands, from features like existence, extension or shape. The morphological aspect of these structures indicates the degree of severity of the related disease. In this paper, we use a large dataset of 5000 images to classify eight different tissue types in the case of colorectal cancer. We compare results with another dataset. We perform image segmentation and extract statistical information about the area, perimeter, circularity, eccentricity and solidity of the interest points in the image. Finally, we use and compare four popular machine learning techniques, i.e., Naive Bayes, Random Forest, Support Vector Machine and Multilayer Perceptron to classify and to improve the precision of category assignation.The performance of each algorithm was measured using 3 types of metrics: Precision, recall and F1-Score representing a huge contribution to the existing literature complementing it in a quantitative way. The large number of images has helped us to circumvent the overfitting and reproducibility problems. The main contribution is the use of new characteristics different from those already studied, this work researches about the color and morphological characteristics in the images that may be useful for performing tissue classification in colorectal cancer histology.

Acknowledgements

I praise and thank Allah SWT for his greatness and for providing me with strength, patience and courage to complete this thesis.

I would like to take this opportunity to express my deepest gratitude and sincere appreciation to my supervisor, Prof. Dr. Boubakeur Boufama, for his continuous belief, patience and guidance throughout the thesis. Words fall short to express my gratitude.

I would like to thank my committee members Prof. Dr. Imran Ahmad and Dr. Mohamed Belalia for their time, constructive comments, suggestions during the thesis. I would like to extend my thanks to Prof. Dr. Dan Wu for his kind acceptance to be the chair of my thesis defense.

I would like to thank my parents, Mr. Syed Moulana and Mrs. Saira Banu, my sister and brother-in-law for their immense support and belief in me.

Lat but not least, I would like to thank the special person in my life, Khatia for all the encouragement.

Table of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Computer Vision

As humans, we perceive the three-dimensional structure of an image taken by the world around us. Otherwise, as we can perceive evident elements, we can disconcert for some common optical illusions that makes it difficult to detect useful information. As mentioned in [11], visual information is being attracted by our eyes at great speed. Much of this information is redundant and compressed by several layers in the visual cortex, so that the higher centers of the brain interpret only a small fraction of the information.

Computer Vision is the study of methods and techniques to extract more information than we can observe, this system can be efficiently used in practical applications. Therefore, it encompasses both the science and engineering of vision. Computer vision focuses on extracting relevant information from images at a high level to perform vision-based tasks [5].

Research topics under computer vision include motion detection, autonomous navigation, scene reconstruction and recognition, augmented reality (AR), object recognition, object tracking and many more vision-based tasks. Consequently, there are factors involved that make vision such a difficult task for machines to accomplish. Computer vision systems often face challenges such as scale change, variations in lighting conditions, point of view changes, partial occlusion, deformation of non-rigid objects, or intra-class variation of visual

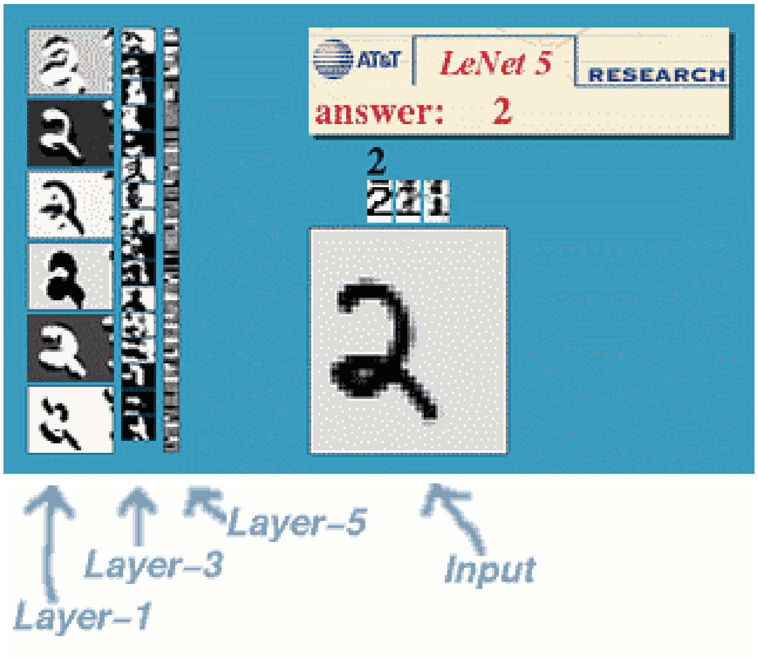


Figure 1.1: Optical character recognition (OCR). Image acquired from `http://yann.lecun.com/exdb/lenet/`

perspectives [10].

Some typical applications of computer vision can be found in optical character recognition (OCR) which extract printed or handwritten text from images (Figure. 1.1). In other field, with the help of computer vision it is possible to detect the movement of cars and to track and count the different vehicles by analyzing a camera picture (Figure 1.2). Another example of an useful application of computer vision is face detection, known as the first step towards many face-related technologies, with recognition or verification purpose (Figure 1.3).

Figure 1.4 is an example of a medical assistance, oriented to Computer-aided diagnosis (CAD). CAD is known as one of the major research subjects in medical imaging and histological diagnostic, that the performance by computers complement physicians view [4].

Computer vision methods for assisting medical diagnosis generally follow a similar structure: it relies on the extraction and combination of several features obtained from pre-processed images, and uses them to build models that can be generalized to invisible data. This image processing methodology can be used to extract out of sight information, thereby providing valuable outcomes to perform related tasks in CAD.

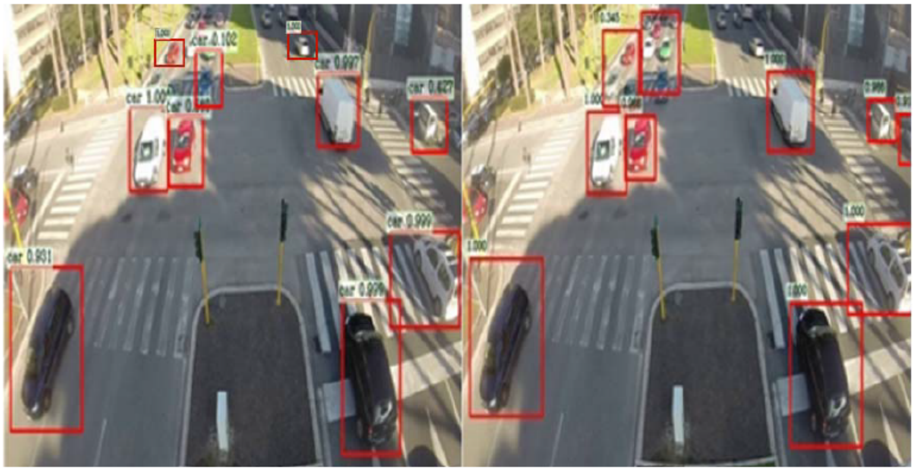


Figure 1.2: Surveillance and traffic monitoring (Image acquired from [7]).

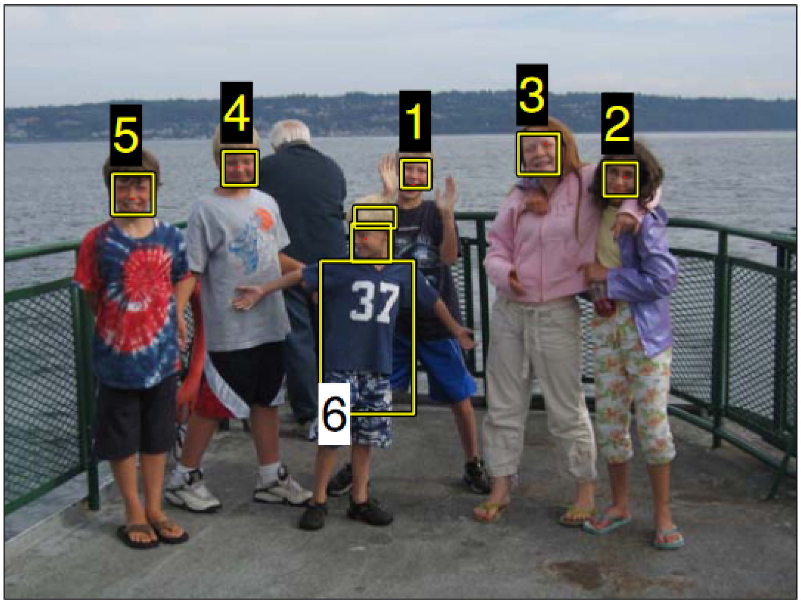


Figure 1.3: Face detection algorithms can locate and recognize the individuals in this image. (Image acquired from [9]).

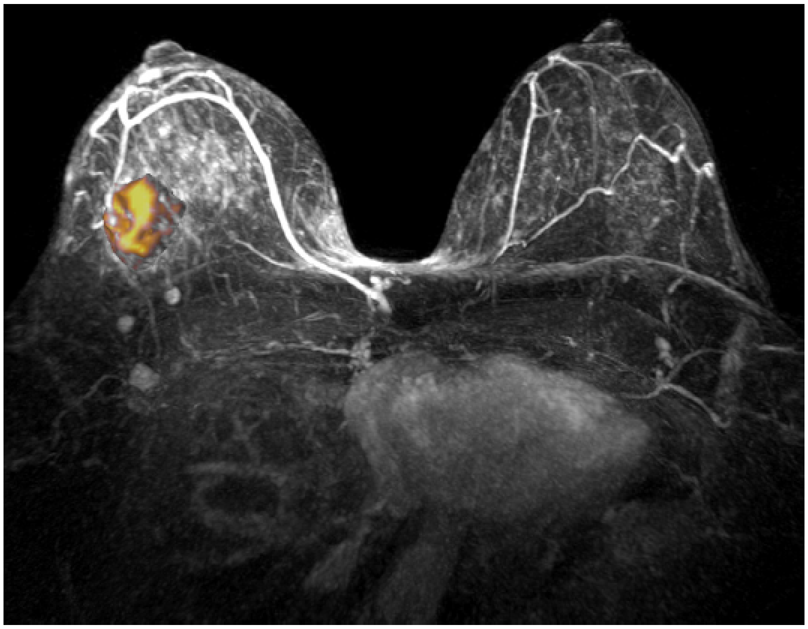


Figure 1.4: Medical imaging - radiological evaluation of dynamic contrast-enhanced from a magnetic- resonance imaging (MRI) of the breast (Image acquired from [8])

We further explain machine learning and supervised techniques to features extraction in section 1.2 and deep learning in computer vision in section 1.3. We present an overview of Digital Pathology in section 1.4. Then, the problem statement in section 1.5, application in section 1.6, and finally the thesis organization in section 1.7.

1.2 Machine Learning

In order to solve a problem on the computer, we need an algorithm. An algorithm is a set of instructions that should be executed to convert input to output.There may be various algorithms, we intend to find the most efficient. *Machine learning* is the programming of computers to optimize performance, using sample data or past experience as a standard to solve a given problem [25].

Machine learning uses statistical theory to build mathematics model to make inferences based on samples. The core task is twofold: First, in training, we need to write efficient algorithms to solve optimization problems and storage and processing data. Two, a model is learning, its representation and reasoning needs algorithm solutions. We build a model depending on some parameters, learning is the execution of a computer program using

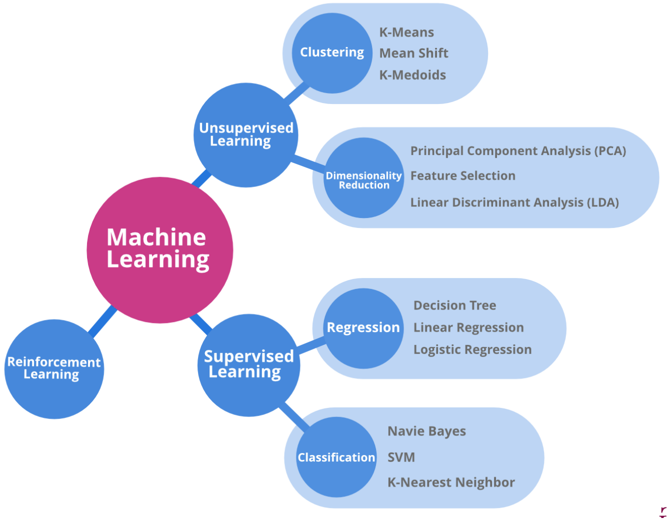


FIGURE 1.5: Types of Machine Learning algorithms (Image acquired from `https://www.geeksforgeeks.org/`)

training data to optimize model features to get a model that may be predictive and acquire knowledge from data.

There are 3 different types of Machine Learning (ML) techniques named:

- **Unsupervised ML Algorithms**: refer to drawing inferences from unlabelled data.
- **Supervised ML Algorithms**: refer to training and testing the algorithms over labelled data.
- **Reinforcement ML Algorithms**: refer to learning a lesson by optimal actions through trial and error. This means that the algorithm decides the next action by learning behaviors that are based on its current state and that will maximize the reward in the future [25]. We summarize ML types on Figure 1.5

Furthermore, we discuss along this thesis, a classification algorithm trained by *supervised learning*. In supervised learning, classification algorithms are used when the output variable is categorical, which means there are two or more classes such as Yes-No, True-false, etc. Figure 1.6 shows a model trained using labeled data sets, where the model learns for each data type. After the training process is completed, the model will be tested on the test data (a subset of the training set), and then the output will be predicted. As example, Figure 1.6 describes supervised leaning when we know that there are three classes, three

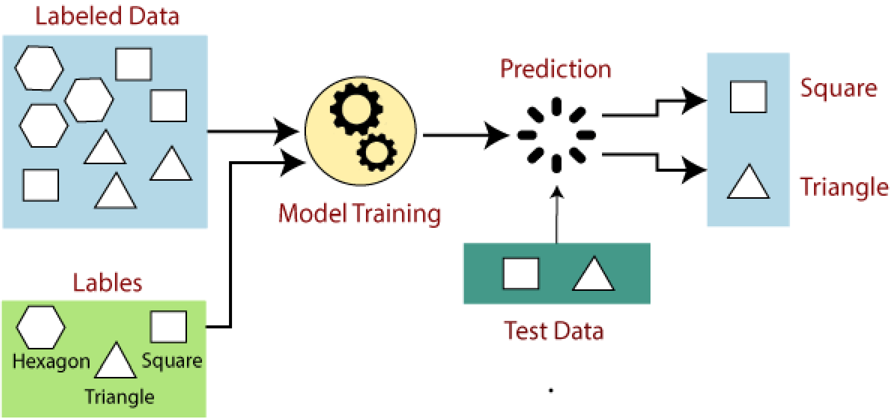


Figure 1.6: Supervised and Unsupervised learning. (Image acquired from https://www.javatpoint.com/supervised-machine-learning)

hypotheses induced, and, each one is covering the instances of one class and leaving outside the instances of the other two classes

With the help of classification tasks in Machine learning field, we find useful applications in eHealth, such as early diagnosis and prognosis of cancer types. Cancer research has been popularized due to the importance of classifying cancer patients into high-risk or low-risk categories has led many research teams.

Therefore, these technologies have been used for the purpose of simulating the progress and treatment of cancer conditions. In addition, the ability of machine learning tools to detect key features from complex data sets reveals their importance. A variety of these techniques, including Bayesian Networks (BNs), Support Vector Machines (SVMs), Decision Trees (DTs), and Multilayer Perceptrons (MLP) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making [26].

Some machine learning techniques are majorly related with colorectal cancer (CRC) diagnosis and prognosis. This thesis focuses mainly on detecting the primary site of cancer. Then, there is need of popular ML techniques required for the early detection of this type of cancer as we present:

- **Naive Bayes**

  Naive Bayes classifiers are statistical classifiers that can predict class membership probabilities, such as the probability that a given sample will belong to a particular case. The naive Bayes classifier depends on Bayes' theorem. This classifier assumes that the influence of attribute values on a given class is independent of the values of other attributes. [29].

- **Random forest**

  Random forest is a collection (ensemble) of decision trees. It is a popular ensemble technique in pattern recognition. Random forest for cancer classification is based on gene expression and generated with the different number of features per node split [29].

- **Support Vector Machines (SVMs)**

  Support vector machines are supervised learning models. In the SVM algorithm, each data item is drawn as coordinates in an n-dimensional space, where n is the total number of elements used for classification, and the value of each element is represented by the coordinates of the data point. SVM contains a decision hyperplane, which is used to divide different types of data points using maximum margin. The data points located near the hyperplane are called support vectors. This classification process generates a nonlinear decision boundary and classifies data points that are not represented in a vector space [27].

- **Multilayer Perceptron (MLP)**

  The **multilayer perceptron** is a feedforward artificial neural network. MLP is a supervised learning algorithm used to learn functions by training on the basis of a given data type. MLP can learn a nonlinear function approximator for any classification. [29]

These techniques allow to classify important characteristics of pathological diagnosis, but we need to compare each other. With this in mind, we use 4 performance measures for classification algorithm [30] which will be discussed further.

1.3 Deep Learning

Deep learning is a subset of machine learning, consisting of algorithms inspired by the structure of a human brain. It is particularly useful for working with unstructured data which nevertheless follows some internal rules: images, videos, text, etc. Through the use of deep learning, computer vision has improved rapidly. Using deep learning models, such as Convolutional Neural Networks (CNN), image processing has become a very broad field, covering a variety of methods to reproduce human visual capabilities, such as recognizing human faces, objects or patterns [15].

CNNs are developed based on Artificial neural networks (ANNs). ANNs are one of the most famous machine learning models. It was introduced as early as the 1950s and has been actively researched since then [15]. Roughly speaking, a neural network consists of many connected computing units (called neurons), which are arranged in layers. There is an input layer from which the data enters the network, followed by one or more hidden layers, which transform the data as it flows through the data, and then terminate in the output layer that produces the neural network predictions.

Describing briefly the diagram shown in Figure 1.7, Neural Networks consists of the following components: An input layer x, an output layer y, an arbitrary amount of hidden layers, a set of *weights* (W) and *biases* (b) between each layer and a choice of activation function for each hidden layer (in this case σ).

In computer vision, this methods begins from a random value, training repeatedly until the network learns the most appropriate weights for each neuron and the output layer provides an accurate prediction. The network is adjusted accordingly to update each weight of each neuron until the pattern recognized by the network produces a good prediction on the training data. However, we need to evaluate the "goodness" of these predictions. Hence, we use a *loss function* that allows to calculate an error rate which show us the difference between each predicted value and the actual value. There are many available loss functions, we choose the most convenient according to the nature of our problem, once we determine one, the goal in training is to find the best set of weights and biases that minimizes the loss function. Furthermore, we need to update our weights and biases based on *Backpropagation*. Backward propagation of errors (*Backpropagation*), is an algorithm

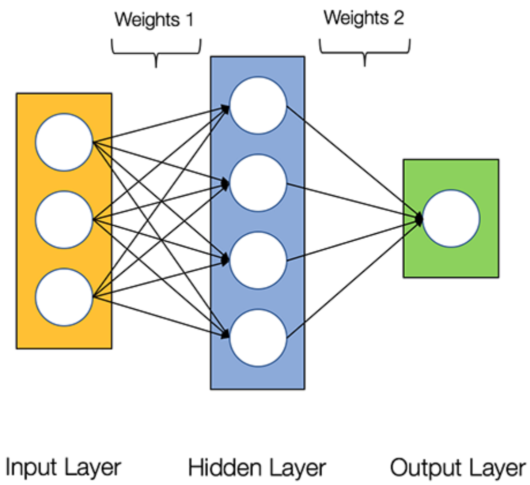


FIGURE 1.7: Architecture of a 2-layer Neural Network (input layer excluded). (Image acquired from `https://towardsdatascience.com`)

that calculates the gradient of the error function with respect to the neural network's weights [18].

At this part, a standard neural network can be constructed and trained. However, deep learning is not possible because the training process involves a different activation function in contrast to classical bounded activation like $sign(x)$, $\sigma(x)$, and $tanh(x)$. These functions are associated to each node, and are applied to node inputs to produce node outputs. In this case we use the **Rectified Linear Activation Function** ($RELU$). This is a non-linear function that will output the input directly if it is positive, otherwise, it will output zero. It has become the default activation function for many types of CNNs because a model that its derivative with respect to its input, is always 1 for positive input value, hence it solves the problem of vanishing gradient and makes easier to train and often achieves better performance [17].

In Figure 1.8, the Rectified Linear Unit (RELU) function, known as the most widely-used activation function in neural networks today is shown. One of the greatest advantages RELU has over other activation functions is that it does not activate all neurons at the same time it converts all negative inputs to zero. This makes it very computationally efficient as few neurons are activated at any given time. In practice, RELU converges six times faster than other functions like *tanh* and *sigmoid* [18].

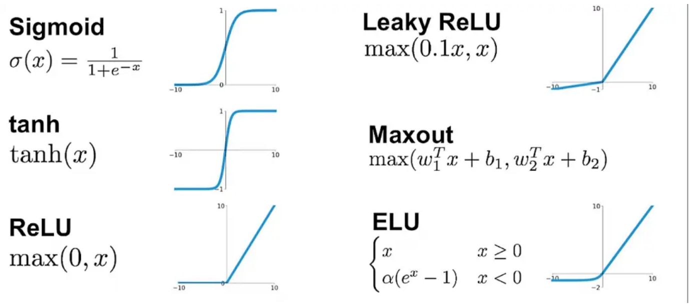


Figure 1.8: Common activation Functions. (Image acquired from `https://7-hiddenlayers.com/deep-learning-2/`)

Returning to computer vision, CNNs are the most researched machine learning algorithms in medical image analysis for preserving spatial relationships when filtering input images. A CNN takes an input image of raw pixels, and transforms it via *Convolutional Layers*, *Rectified Linear Unit (RELU) Layers* and *Pooling Layers*. Then, this output feeds into a final Fully Connected Layer which assigns class scores or probabilities, thus classifying the input into the class with the highest probability [19].

The *convolutional layer* is the core component of CNN. It bears the main part of the network computing load. The main purpose of convolution is to extract features such as edges, colors, and corners from the input. As we go deeper into the network, the network also begins to recognize more complex features, such as shapes, numbers, and faces.

Convolution is defined as the operation of two functions. In image analysis, one function composed of the input value (such as pixel value) at a certain position in the image, and the second function is a filter (or kernel); each can be represented as an array of numbers. Calculating the dot product between two functions to get the output. This process is repeated until cover the entire image, generating a *feature map*, a map where the filter is strongly activated and highlight a feature such as a straight line, a dot, or a curved edge [16].

Moreover, the *RELU layer* uses an activation function that turns into negative input values to zero. Thus, simplifies and accelerates computing and training, and helps to evade vanishing gradient problem. The *pooling layer* is inserted between the convolutional layer and the RELU layer to reduce the number of parameters to be calculated and the size of the image (width and height, but not depth). Finally we find in a CNN, the *Fully Connected*

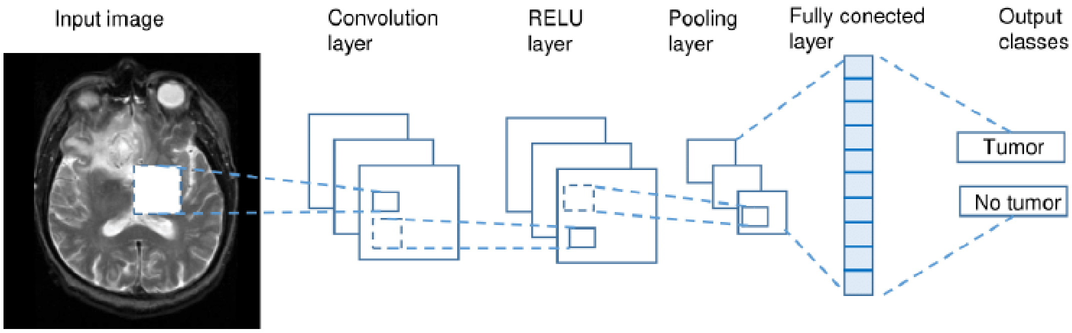


Figure 1.9: Disease classification task. (Image aquired from [16]).

Layer. This layer takes the output of the previous layer (convolution, RELU or pooling) as input and calculates the probability score of the classification from different classes available. As shown in Figure 1.9 input image is an abnormal axial slice MRI brain going through a schematic of a Convolution, and RELU and pooling layers, before classification by the fully connected layers. We notice that final connected layer looks at the combination of the most strongly activated functions that will determine that the image belongs to a particular category. Other example where we can apply this CNN, is on histological slides. Cancer cells have a higher ratio of DNA to cytoplasm than normal cells, ff the DNA signature is strongly detected from convolution, RELU or pooling layer, CNN is more likely to predict its existence cancer cell.

1.4 Digital Pathology

In the last century, the basic process of pathologists for diagnosis has remained relatively unchanged, but advances in information technology have provided huge opportunities for image-based diagnosis and research applications. Pathology has lagged behind other health care practices, such as histology, where this technology is widely adopted in digital medicine. As the equipment that generates full slide images becomes more practical and affordable, practice will increasingly adopt the technology and eventually generate new data that will quickly make the present large amount of histological imaging data overshadowed [20].

This work is framed into what is called *Digital Pathology*, which is a relatively recent area of research which is dedicated to providing accurate and efficient computational methods to support quantitative detection, diagnosis, and prognosis in pathology. It presents various computational learning methods and frameworks for the automatic representation of histopathology images by learning from databases of different digital pathology tasks including the detection, localization and quantification of tumours and tissues in various types of cancer.

The role of machine learning in computer vision is very significant. There are many supervised learning techniques which are capable of segmenting, measuring and classifying images to be applied to routine pathological tasks, including the quantification of antibody staining, the identification and classification of cells, and the characterization of essentially multi cellular or regional microstructures [20]. The development and evolution of computational techniques to classify categories such as K-means algorithms, Support Vector Machines (SVM), Multilayer Perceptron (MLP), Random Forest Algorithm among others have allowed the appearance of some solutions to apply on diagnosis assistance of malignant tumours and tumour vs non-tumour identification.

1.4.1 Histopathology

Delving deeper into the field of pathology and histopathology, we find Histopathology slides provide a more comprehensive view of the disease and its impact on the tissue, because the preparation process preserves the underlying tissue structure. In this way, certain disease characteristics, such as lymphocytic infiltration of cancer, can be inferred only from histopathological images [21].

In the last decade, technological advances in software and hardware have made possible the digitization of histological slides and the creation of histopathological images, through robotic microscopes, known as operating tissue scanners. Digitized slides are generated from image data of the entire sheet in high resolution. These are visualized through computer platforms that emulate the functionalities of a microscope, increasing with information technologies through the internet and accessed through the use and exchange of data and images online through computers or mobile devices. These technological developments offer great advantages over the traditional treatment of using microscopes, such as ease of access, interconsultation, second opinion, as well as multiple applications and useful tools in clinical diagnosis, research and education [22].

With the arrival and mass use of digitized slides, great interest has arisen in the development and use of image analysis algorithms that allow the quantification of various biological markers. The benefits provided by such algorithms include improvements in accuracy and precision in the detection, classification and measurement of morphometric patterns. In this last aspect, these analyzes are currently being applied mainly in the quantification of the expression of tumour markers in cancer.

In recent years, to analyze these characteristics and discover abnormalities of some type of cancer, computerized systems have applied segmentation algorithms (unsupervised machine learning). Other studies have focused on computer classification of the type of tissue present, in order to make a diagnosis based on it. The fundamental problem in most of the algorithms develop is the large number of parameters that need to be tuned, which is why work has been carried out to adjust them automatically. The process of segmentation allows to distinguish between different regions in the image and thus finding regions of interest.

1.4.2 Features Extraction

The systems used for digital pathology are designed to process images obtained from different microscopes. Multiple visual characteristics (colors, shapes, etc.) are obtained from the regions of interest (ROI) to identify and make a better diagnosis from tissue samples, through image segmentation and feature extraction [23].

Additionally, morphological characteristics provide information about the size and shape of the described region, object or image. For example, the glandular area can be used as a discriminatory criterion to classify between cancer and non-cancer, the perimeter can also be used as a descriptor of the traits to characterize the size of the cells in the segmentation process. Taking advantage of the pathologist's experience in the diagnosis of Oral Submucosal Fibrosis (FSO), Muthu and the other researchers represented the image using morphological characteristics such as the perimeter of eccentricity, thus being the diameter of the equivalent area to describe the nucleus of the cell. Not only can classification tasks be performed using these types of characteristics, there is also an automated system for differential White Blood Cell (WBC) count based on 19 characteristics such as area, perimeter, convex area, solidity, orientation and eccentricity. In addition, these features have been used to construct Content Based Image Retrieval (CBIR) to find histopathology images of the prostate based on morphological similarity. Finally, most of the estimation of morphological measurements is made based on a previous segmentation, so its performance depends on the precision of said segmentation.

The intensity characteristics provide information on gray level or color of the pixels located in the regions of interest. This feature extraction approach uses different color spaces where the hue channel of the HSV (Hue-Saturation-Value) color space conversion of the original image is used. It can also work in the white/pink/purple dimension with the objective of evaluating color models. It allows to compare the incidence on the performance in a classification task and makes sure that there is no single model that works better than others in all cases. Figure 1.10, presents an example of the perception-based feature values of an epithelium and stroma tumour from a histopathological slide along with the feature values from a colorectal cancer database consisting of 1332 image of epithelium and stroma tissue sample extracted from the patients [24].

1.5 Problem Statement

All these studies are motivated due to the fact that Cancer is a global public health problem in the 21st century. According to WHO (World Health Organization), cancer is ranked as the third cause of death in the world and the second in developed countries, only surpassed by cardiovascular diseases. The WHO GLOBOCAN developments aim to provide updated estimates on the incidence, mortality and prevalence of the main types of cancer, at the national level, for 184 different countries. These estimates are based on the most recent data available from the International Agency for Research on Cancer (IARC) and on information available to the public on the Internet.

According to estimates of cancer treatment and survival rates, on January 2019, more than 16.9 million Americans had a history of cancer, only to population growth and aging, this number is expected to exceed 22.1 million statistics by 2030. The report is produced every three years by the American Cancer Society and the National Cancer Institute to

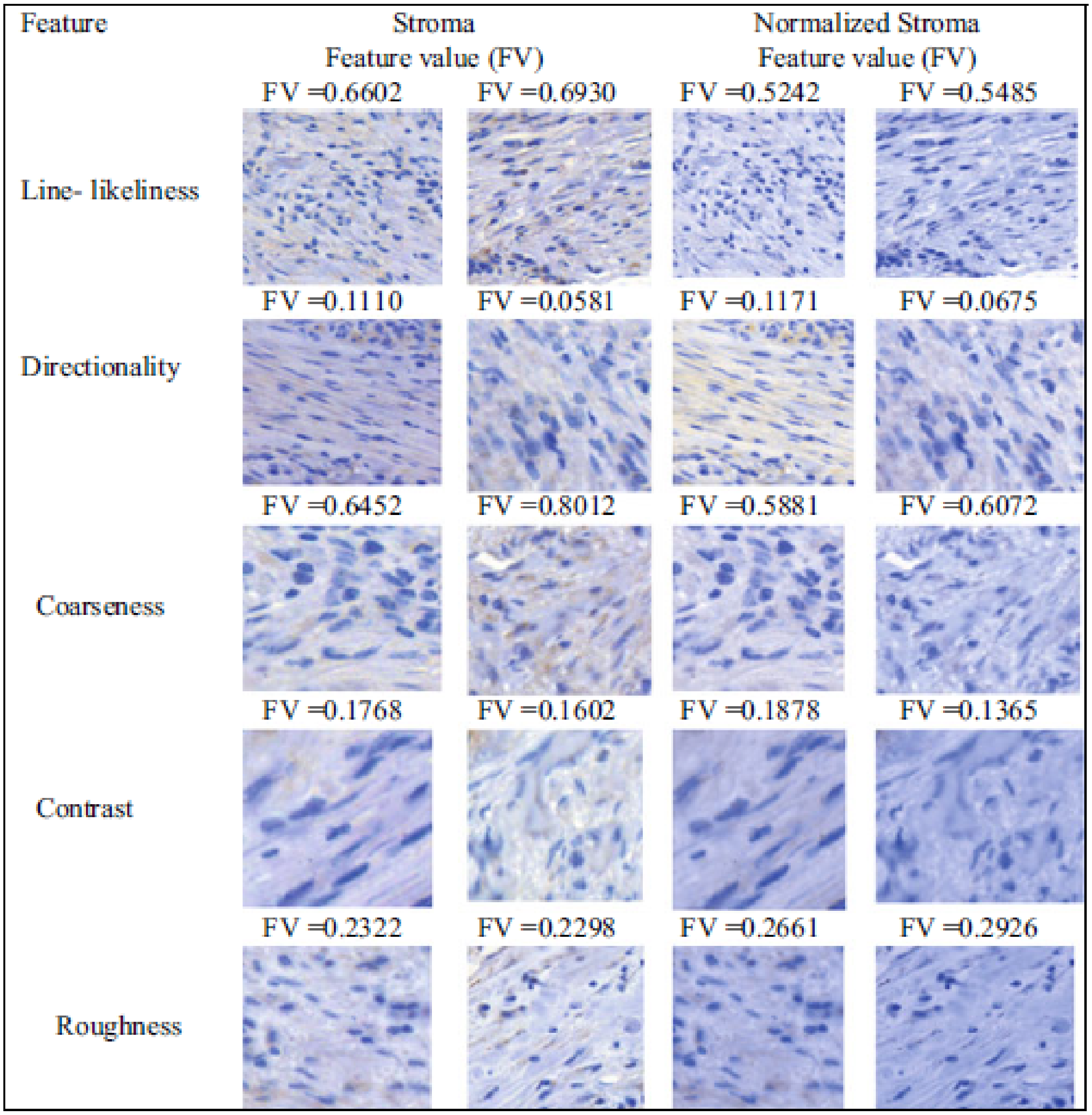


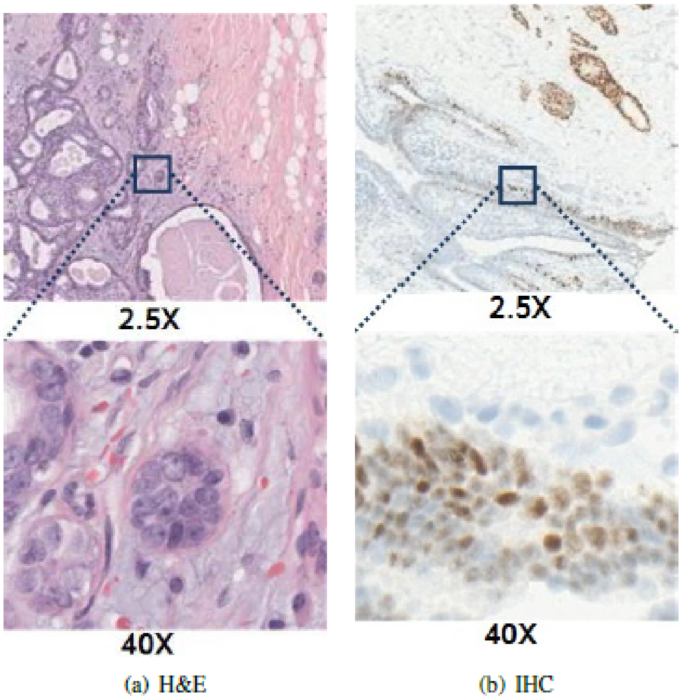
Figure 1.10: Sample image of epithelium tumour and stroma tumour (Image acquired from `http://fimm.webmicroscope.net/supplements/epistroma`) (Image acquired from [24])

help the public health community better serve this growing population [41]. The current numbers mean that one in five men and one in six women will eventually get cancer by the time they reach 75. One in eight and one in 12 women will pass away from the disease [66]. One of the most concerning factors was that 57% of new cancer cases and 65% percent of deaths came from the developed world.

The three most common cancers among men in 2019 are prostate cancer (3,650,030), colon and rectal cancer (776,120) and skin melanoma (684,470). Among women, the three most common cancers are breast cancer (3,861,520), endometrial (uterine body) (807,860), and colon and rectal cancer (768,650). The author's estimate of the number of cancer survivors in 2030 (22.1 million) is based on population projections [41]. Therefore, the less developed regions and that have few resources will occur 57% (8 million) of new cancer cases, 65% (5.3 million) of cancer deaths and 48% (15.6 million) of people with the disease (5 years after diagnosis) [66].

1.6 Applications

The histopathology acquisition process follows a well-defined methodology in its first steps, according to Figure 1.12 [48]: First, the biological sample is taken from an organ. Then, a fixation process is done over the biopsy to ensure chemical stability on the tissue and to avoid post-mortem changes. After this, it must be cut into sections that can be placed onto glass slides. The sections are stained to reveal cellular components by chemical reactions. The most common dyes used are Hematoxylin - Eosin (H&E), which stain cell nuclei in a dark blue or purple and cytoplasm and connective tissue in bright pink (i.e. as shown in Figure 1.11). Finally, the section covers slipped into being viewed and digitized with a microscope.



(a) H&E

(b) IHC

FIGURE 1.11: Examples of H&E and IHC stained images (Image acquired from [23])

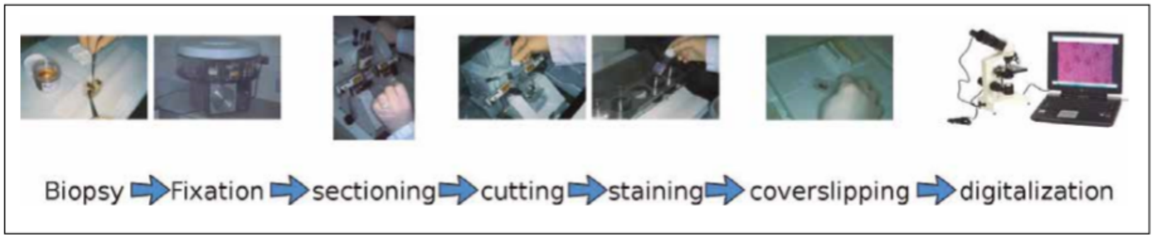


FIGURE 1.12: Acquisition workflow diagram in histopathology images [48].

Human factors in this process constitute sources of variability. This visual heterogeneity is due to three factors at least [48]: (i) **Magnification**, referring to increasing the proportion of biological structures which are visible under the microscope according to the set of lenses; (ii) **Staining**, that is, improve the contrast in a biological sample seen under a microscope; (iii) **Slice Orientation**: how the tissue appearance changes when the cut is done in a longitudinal or cross-sectional orientation in smooth muscle tissue using the same staining (H&E).

The usual histopathology image workflow usually goes as follows [48]:

- **Image preprocessing:** Raw image data is transformed to reduce visual variability and noise, as well as making it more suitable for the subsequent steps.
- **Feature extraction:** Its goal is to produce a more descriptive representation of the image-making important explicit information which is not directly manifest from the raw pixels.
- **Pattern recognition:** In which interesting visual patterns are detected and identified through supervised and unsupervised algorithms.

As described above, the development of techniques to solve the public health problem caused by cancer has evolved significantly in recent years. Even if this has been caused by the emergence of new algorithms due to the boom in computer science, the growth and development of the applications of these techniques in cancer detection has been significantly high than others areas due to the high investment that countries have invested in the study for the treatment and cure of cancer. For example, the European Union in recent times has been increasing its budget for cancer research and treatment, measures that have been taken by each and every one of the member countries in proportion to their economic capacities, which shows the importance given to the subject in this region and the understanding that prevention is cheaper than treatment.

To summarize, research in the field of computer vision for assistive medical diagnosis is increasing because it provides objective goals about the patient's condition. Therefore, a non-invasive way to detect colorectal cancer, is essential for medical diagnosis, because developed detection algorithms are still rarely used in clinical practice, their reliability is still worthy of attention due to lack of research on this type of cancer.

As a result, this thesis analyzes histopathological images to decide whether a tissue contains cancer or not. Using computational learning tools which automatically find patterns for healthy and abnormal tissues, we contribute a fundamental support for the diagnosis of colorectal cancer with this research. It is also embedded into histology which studies the microscopic anatomy of biological tissues. In this study histopathological cancer images are the cornerstone to understand the state of biological structures, provide the diagnosis and analyze the state of diseases [48].

1.7 Organization

This thesis is organized as follows: Chapter 2 reviews the literature in image processing and computational algorithms used in cancer detection. Chapter 3 will go over the methodology used to address the problem of deciding whether a tissue is cancer positive or not as stated in section 1.6. Chapter 4 will go over the results showing the performance of the algorithms used and tested to classify the histopathological images. Finally Chapter 5 Resumes the mains aspects and conclusion of this work.

Chapter 2

LITERATURE REVIEW

This chapter reviews the literature related to classification of histopathological images for diagnosis/procedures. We reviewed work focused on an increasing amount of literature with different datasets and objectives, having in common a high degree of accuracy. The common feature in CADx works is that specific image analysis problems require specific image representation schemes. Therefore, we make an overview of the most common features extraction scheme for histopathological image representation listed on the visual features which describe and provide the most relevant information for specific machine learning algorithms, depending on the type of image, some features may be more significant than others.

Thus, we review features extraction in section 2.1. This section reviews visual features extraction and the properties of tissues images which will be considered as inputs of machine learning and deep learning techniques to construct a model. Subsequently, we notice some challenges from high-level semantic concepts as connected areas and objects which could determine the visual pattern of disease if it exists, or, a false positive case. Thus, we focus on processing histopathological challenges for CADx in section 2.2.

2.1 Features Extraction

Numerous methods have been used to obtain a host of features. The latter represents clinically meaningful information after the preprocessing of images. Indeed, according to [48] *Feature extraction methods* are usually performed on the following feature types:

2.1.1 Texture Features

Texture features provide data about variance of intensities inside of a specified region. A texture is a set of connected pixels that repeatedly appear in an image. It provides information about changes that determine the surface strength by quantifying properties such as smoothness, roughness, and regularity. To extract information from these features one can use statistical properties like correlations, means, etc, to relate between each others to tissue identification. Similarly, after such statistical processing, these features can be used to classify sub-images. For example, the image is divided into squares of sub-images and texture features are extracted from these squares, and then the squares containing malignant cells are distinguished from the squares composed [62].

2.1.2 Color Features

Color features or visual features based on intensity provide information on the gray level or color of pixels located in the ROI. This type of function does not provide any information about the spatial distribution of pixels. Intensity histogram units are used to define characteristics. For example, using the gray value of the pixel to define the optical density of the pixel and using the pixel values in a single color channel may establish a relationship between color values in different channels.

Figure 2.1 shows an example of color differences between a normal and a cancerous tissue from a histopathological sample. Also, at the same figure we can distinguish patterns like texture that can be extracted.

In Welsh approach, the transfer result depends on the luminance information of reference image. Since the process of Gupta approach is on the purpose of propagating color

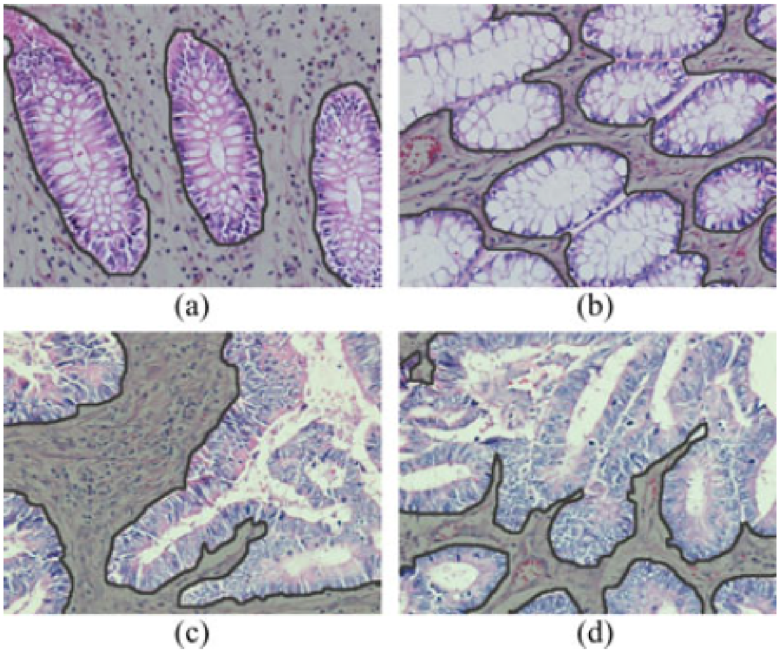


Figure 2.1: Example of color and texture differences between histopathological images of colon tissues: (a), (b) Normal and (c), (d) cancerous. Nonglandular regions in images are shaded with gray. Image aquired from [62]

information using the least-squares optimization method, the result shows limited relevance to the reference image luminance.

2.1.3 Topological Features

Topological features provides information about some structure within the image. The structure of a tissue by quantifying the spatial distribution of its cells. Thus, graph construction, local and global graph features is a usual method to extract patterns that consist of a bunch of cells. Therefore, the local metrics correspond to the properties of segmented images or sub-images. Moreover, the global ones are the giant connected component ratio, spectral radius. As example, Figure 2.2, shows a sample image and its respective graph model.

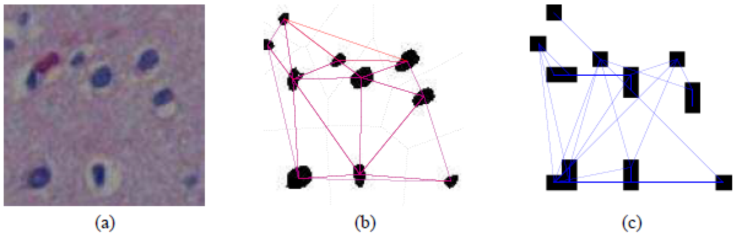


Figure 2.2: a) A sample image, (b) the Voronoi diagram of the image (black dotted lines) and its Delaunay triangulation (red solid lines), and (c) a cell-graph of the image. Image acquired from [32]

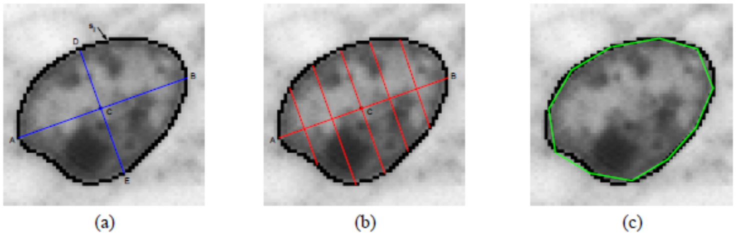


Figure 2.3: (a) A sample of a nucleus with its boundary points and centroid, (b) line segments used in symmetry computation, (c) chords used in concavity computation. Image acquired from [32]

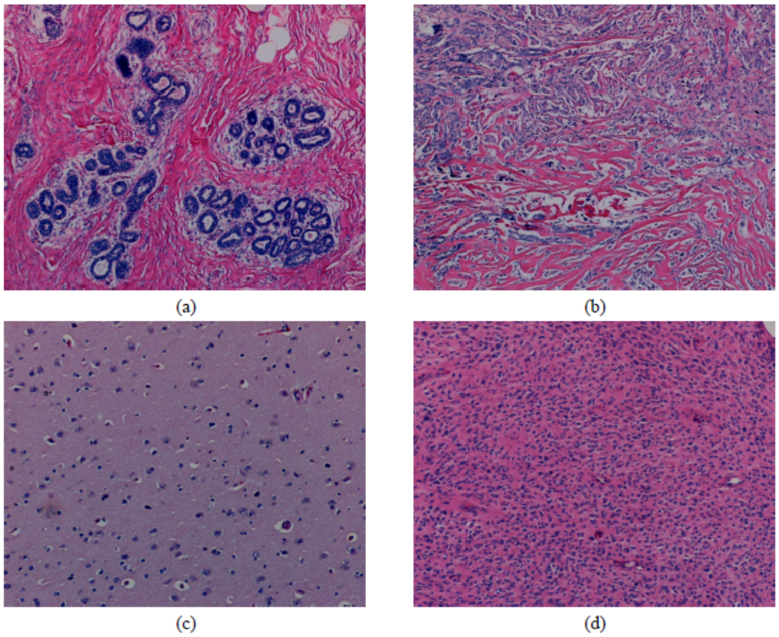


Figure 2.4: Similarity relationship from a histopathological images of biopsy samples: (a) a healthy breast tissue, (b) a cancerous breast tissue, (c) a healthy brain tissue, and (d) a cancerous brain tissue. Image acquired from [32]

2.1.4 Morphological Features

Morphological features provide information about the size and shape of the described region, object in an image. This information is particularly useful on the segmentation task, where graph theory is commonly the most used in this group of features. Area, perimeter, angle and other geometrical features are calculated. On the other hand, the shape consists of compactness, roundness, smoothness, length of major and minor axes, symmetry, concavity and circumference [32]. These characteristics can be found in Figure 2.3 and a sample image to describe which segments and patterns may be calculated.

2.1.5 Similarity Features

Similarity measures provide information about similar image sequences. Commonly used to compare characteristics from different samples. Proof of this, Figure 2.4, shows a couple of different sample images that can be related. Furthermore, this review can summarize the above features in Table 2.1.

Feature Type	Feature Extraction Method or Metric	Authors
Texture	First-Order Statistics	[53]
	Lower Order - High Order Histograms	[62]; [54]
	Local Binary Patterns (LBP)	[56], [54], [58]
	Run Length Matrix (RLM)	[53]
	Gray-level co-matrix (GLCM)	[46], [54], [53], [58]
	Haralick Descriptor	[56], [46], [52], [58]
	Gabor filters	[56], [54], [58]
	Perception-like features: coarseness, contrast, directionality, line-likeness, and roughness.	[54]
Color	Sample Entropy	[49]
	First Order Statistics	[52], [53]
	Contrast Measure	[56]
	Quantile Normalization	[55]
Morphological	Perimeter (Size)	[50]
	Ratio of major to minor axis (Shape)	[50]
	Mean of the gray-level intensity (Nuclear Appearance)	[50]
	Axis of Least Inertia	[61]
Topological	Graph Analysis: Diameter, Degree, Clustering Coefficient	[47]
	Delaunay Triangulation	[62], [47]
Similarity	Grid-Based Approach	[62]
	Voting Approach	[62]
	Bag of Words Approach	[62]

Table 2.1: Feature extraction method for different feature types

2.1.6 Extracting Methods Review

According to Table 2.1, most used computational learning models are less heterogeneous due to low sample sizes. While supervised learning techniques are intended to classify new observations in the defined categories, Support Vector Machine (SVM) is the most common method and provides the most accurate estimations in general. Unsupervised learning is used mostly in preprocessing. To detect Usual Ductal hyperplasia (UDH), [50] segmented cell regions by clustering the pixel data with Gaussian Mixture Models (GMMs) in four cytological regions (i.e., cellular: nuclear and cytoplasmic, extracellular, regions with hues and Illumina). The pixels classified as cellular components are further clustered by using dynamic thresholding to eliminate pixels with less luminance. In the end, individual cells are segmented by converting them to gray-level and using watershed. The idea is an amount of waterflows along with a topographic relief following a certain descending path to eventually reach a catchment basin. Blobs in the image can be separated by identifying the limits of adjacent catchment basins and then separating them. Even so, [53] classifies regions of interest containing benign/malignant cancerous colon tissues. Overall, RGB intensities of the samples are converted to optical density values and, therefore, the saturation of each stain. Besides, a K-Means clustering is usually used to extract nuclei structures by identifying pixels with high hematoxylin concentration, generally with only exception presented in [61], which used Self Organizing Maps to split normal and tumorous epithelium tissues.

Computational Learning Class	Method	Authors
Supervised Learning	Support Vector Machine	[47], [50], [56], [62], [54], [46], [52] & [49]
	Logistic Regression	[49]
	Neural Networks	[46], [54], [49]
	Linear Discriminant Analysis	[55]
	Ensemble Trees	[54]
	Decision Trees	[49]
	Random Forests	[53], [49]
	RBF	[49]
	Multiple Instance Learning	[50]
	Resampling Based-Markovian Models	[62]
	Bilinear Convolutional Neural Networks	[59]
	Convolutional Neural Networks	[60]
Unsupervised Learning	k-Means	[53], [46], [62]
	C-Fuzzy Means	[61]
	Self-Organizing Maps (SOMs)	[61]
	Gaussian Mixture Models	[50]

Table 2.2: Different types of methods for computational learning classes

The most popular state-of-the-art study of the reviewed literature is [60]. Using 86 H&E slides to obtain 100.000 image patches and data augmentation, they identify nine tissue types with several CNNs (Convolutional Neural Networks): VGG19, AlexNet, SqueezeNet, GoogLeNet, and Resnet50, with a 70/15/15 division of the dataset into train/validation/test sets, where the first one obtained a 98% accuracy. Visual representations of tissue classes are obtained through t-SNE on deep layer activations, with an almost perfect separation of the classes in the testing set; visualization of morphological features are derived from a Deep-Dream approach. Their excellent performance let them establish four categories of consensus molecular subtypes (CMSs) and calculate a deep stroma score.

The performance of the algorithms depends on the ways the data are incorporated into them. As shown in [53] the performance of eight algorithms is tested for different ways of incorporating the raw data. see Table 2.3:

Classifiers	Vector with 24 Attributes	Vector with 13 Attributes	Vector with 37 Attributes (24+13)
Random Forest	0.742	0.769	0.79
J48	0.72	0.651	0.742
SMO	0.603	0.687	0.7
Rotation Forest	0.923	0.896	0.913
Multi-Class Classifier	0.889	0.94	0.886
Multilayer Perceptron	0.891	0.917	0.907
Logistic	0.889	0.94	0.866
RBFNetwork	0.641	0.682	0.675
Average $\pm SD$	0.787±0.111	0.810±0.113	0.807±0.081

Table 2.3: Performance of eight algorithms according to [53].

According to Table 2.3, the best algorithm is Rotation Forest allowing to have the best result with 24 and 37 attributes, while better results are obtained using 13 vectors with Multi-Class Classifier and logistic regression algorithms.

2.2 Processing Challenges

A significant methodology related to computer vision systems in medical imaging involves a stage of preprocessing, for example a segmentation algorithm to easily capture the variability induced by texture or color inside histopathological samples, among others.

Most preprocessing steps aims to reduce noise when they are properly applied. Nonetheless there are many challenges in high precision algorithms of tissue classification and high-efficiency algorithms in computation time. Specifically, the number of samples in different studies is too small to draw a general conclusion [54]. The natural high-dimensionality of images increases the complexity of analysis and the over-division into patches could increase the number of false-positive observations, even though the overall accuracy degree is high. In most practical applications the number of labeled images is too low and the staining process depends on laboratory procedures.

Hence, in [59] aiming at internalizing the transformation processes, the authors tried to solve these problems by proposing a Bilinear Convolutional Neural Network (BCNN) for the classification of histopathological images of tissues with colorectal cancer. Also, enhancing the feature representation, we find that the categories specified by [54], as shown on Figure 2.5, were learned from hematoxylin-eosin (*H&E*) components jointly and the classification of the images. They applied a decomposition algorithm to obtain H and E representations through a stain decomposition algorithm. Moreover, with the feature functions associated, they use two parallel CNNs, which are a hierarchy of neural units, including a convolutional, pooling, and non-linear layer. The Feature outputs are combined at each location using the outer matrix product before the classification is done. Similarly in [54], the authors used CNNs for category classification, combining a DeepNet architecture with a focal loss function. Still, they carried out magnification and color normalization to images.

The authors in [62] proposed to generate perturbed images from training data and modeling them by a Markov process. The classification of the images is done using their perturbed samples and thereby, reducing the negative outcomes due to the large variance in tissue images (Figure 2.6). They present their solution as follows. They randomly select points of the stained images and locate a window at the center of each point to extract four-color and texture features. Following, a k-Means algorithm is run to quantify the pixels into dominant cluster-colours of H&E staining (white, pink and purple). The features are then discretized into observation symbols. Each new data point is labeled into the cluster of the nearest neighbour, so each observation has a set of observation symbols that can be ordered by the shortest distance. A sequence can be therefore defined, and the noise is reduced due to the existence of large sequences, modeling images in a Markov framework.

In another related work [49], the authors start dividing stained images into equal-size cells and a rolling squared-window of shape m is defined by moving from upper left to lower right (see Figure 2.7). Then, they define a similarity measure based on comparing maximal individual distances to average intensity in each window and channel. A probability of similarity of each window and the other ones is calculated, which lets them calculate a sample entropy measure from the average of these probabilities as the window moves. The sample entropy curves are described as a function of a constant tolerance and the window size, which let them calculate the area under the curve, obliquity, area ratio, maximum point value and maximum point scale metrics under different configurations. These calculated entities are used as 13 features for classification between benign and malignant images of colorectal adenocarcinoma.

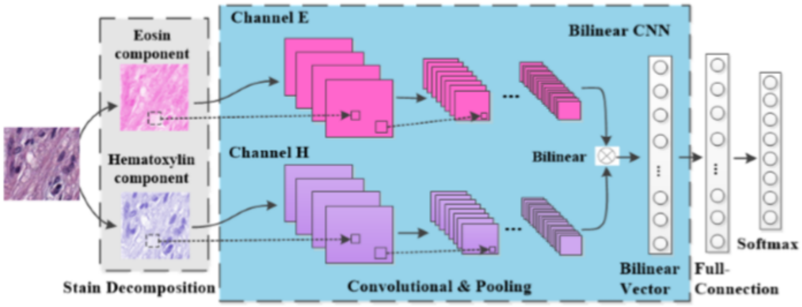


FIGURE 2.5: Flowchart of a proposed BCNN-based classification method for histopathological images [59].

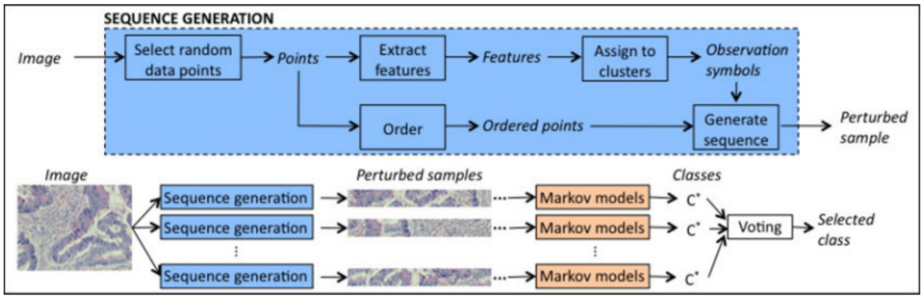


FIGURE 2.6: Schematic overview of the resampled-based Markovian model (RMM) to classify given images [62].

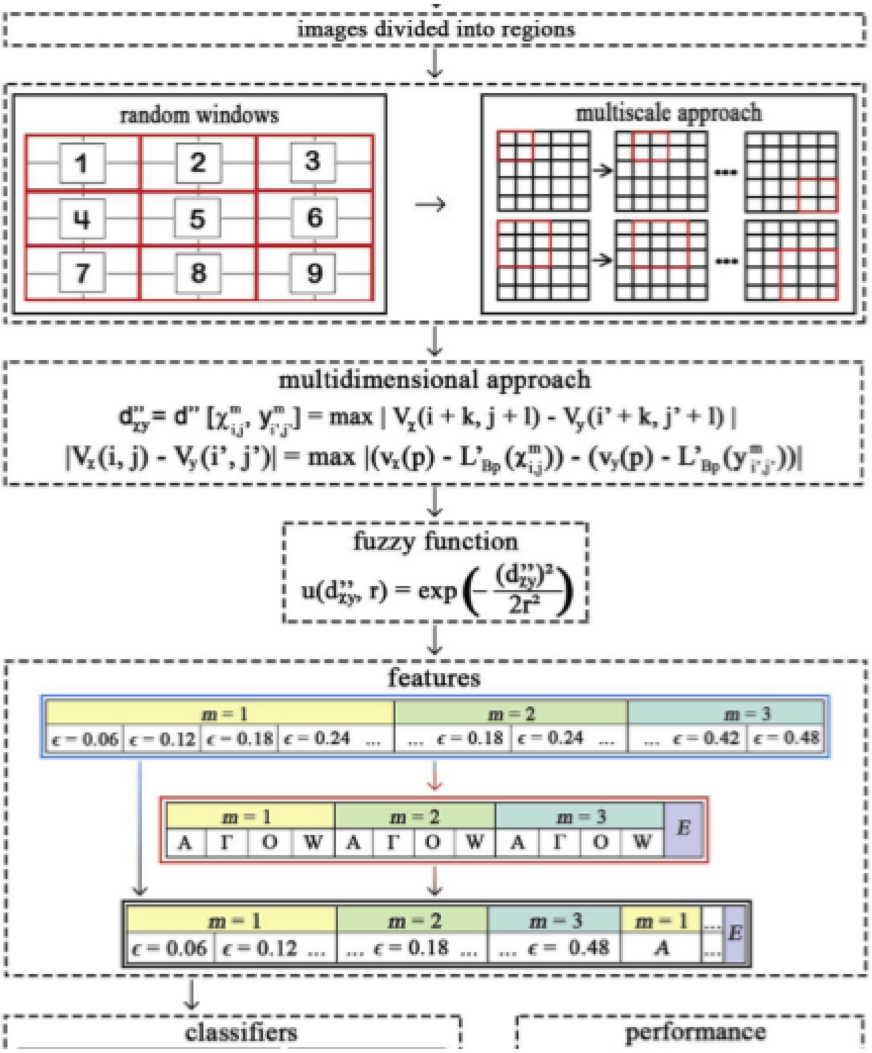

images divided into regions
random windows
1
2
3
4
5
6
7
8
9
multiscale approach
multidimensional approach
$d''_{zy} = d''[\chi^m_{i,j}, y^m_{i',j'}] = \max | V_z(i+k, j+l) - V_y(i'+k, j'+l) |$
$|V_z(i, j) - V_y(i', j')| = \max |(v_z(p) - L'_{Bp}(\chi^m_{i,j})) - (v_y(p) - L'_{Bp}(y^m_{i',j'}))|$
fuzzy function
$u(d''_{zy}, r) = \exp\left(-\frac{(d''_{zy})^2}{2r^2}\right)$
features
m = 1
m = 2
m = 3
ε = 0.06
ε = 0.12
ε = 0.18
ε = 0.24
ε = 0.42
ε = 0.48
A
Γ
O
W
E
classifiers
performance

Figure 2.7: Stages defined for multidimensional and fuzzy approaches [49].

Finally, In Table 2.4, we present a record of literature which exhibits components of relative works with a performance measure and a sample from a database of histopathological images to establish a complete medical diagnosis/prognosis context.

Reference	Classification Objective	Dataset Type	Sample Size	Accuracy
[54]	Tumor epithelium; Simple and complex stromas; Immune cell groups; debris and mucus; mucosal glands; adipose tissue; background.	Colorectal Cancer	5000 images (625 samples for each class)	0.85

[55]	Basophilic structures; Eosinophilic intra & extra cellular proteins; Lumen of glands; Red blood cells.	Renal Tumor;
		Glioblastoma;
		Ovarian.
			Two renal tumors (RCC1 and RCC2 with 55 and 47 images, respectively), one glioblastoma (Gbm, 52 images), and one ovarian (Ov, 50 images).	0.87

[50]	Binary classification of Usual Ductal Hyperplasia (UDH)	UDH	327 regions of interest are used for training; 149 for testing.	0.879
[56]	Epithelium Stroma Tissues	Epithelium Stroma	576 images of regions of interest were used for training and 720 images for testing.	0.995
[49]	Benign and malignant colorectal adenocarcinoma.	Colorectal Cancer.	50/50 images of benign/malignant cancer for training and 17/34 for testing.	0.983
[47]	Normal; Low-Grade Adenocarcinomatous; High-Grade Adenocarcinomatous Colon Tissues.	Colon Cancer Diagnosis	213 microphotographs 115 images of tissues are used for training and 98 images for testing.	0.8265
[53]	Benign and malignant colon tissues	Colon Cancer Diagnosis	44 benign and 43 malignant H&E stained tissue samples.	0.91

Reference	Classification Objective	Dataset Type	Sample Size	Accuracy
[52]	Well, intermediate, poor cancer-degree	Colon Cancer Diagnosis	92 malignant colon biopsy samples (23/44/25 samples associated with poor/mid/well-degrees of cancer)	0.9613
[59]	Tumor epithelium, stromas, immune cell groups, debris and mucus, mucosal glands, adipose tissue, background	Colon cancer	125 images for each class	0.985
[58]	Normal, Hyperplastic polyp (HP), Tubular Adenoma with low-grade dysplasia (TA_LG) and Carcinoma (CA)	Colorectal Cancer Diagnosis	200 images, containing 50 images from each of the four classes	0.9317
[62]	Normal; Low-grade cancerous; High-grade cancerous.	Colorectal Cancer Diagnosis	3226 images of colon tissues. 1644 images for training (510 normal, 859 low-grade cancerous, 275 high cancerous) and 1592 for testing (491 normal, 849 low-grade cancerous, 257 high-grade cancerous).	0.9522 (Normal); 0.8945 (Low); 0.8646 (High).
[46]	Cancerous; Normal	Colorectal Cancer Diagnosis	113 colon images: 64 cancerous & 49 normal.	0.833
[61]	Tumour and Normal Epithelium	Colorectal-Cancer	Diagnosis 134 images of normal and tumorous regions; 14 whole slide images stained for different biomarkers.	0.81-0.96

Reference	Classification Objective	Dataset Type	Sample Size	Accuracy
[63]	Background; Adipose; Mucus; Tumor Epithelium; Mucosal glands; Muscle; Stroma; Blood vessel; Immune cell; Necrosis	Colorectal Cancer Diagnosis	660 digitalized colorectal cancer specimens	0.72-0.96
[60]	Adipose tissue; Background; Debris; Lymphocytes; Mucus; Smooth muscle; Normal colon mucosa; Cancer-associated stroma; Colorectal adenocarcinoma epithelium.	Colorectal Cancer.	86 H&E slides of human cancer tissue to create 100.000 image patches for training; 25 H&E slides to create 7.180 patches for testing.	0.943

Table 2.4: Literature Review

2.2.1 Relation to our Work in Nuclei Classification

Our work in Nuclei Classification uses components from computer vision and classification algorithm. In particular we start at cellular-level feature extraction. We build a nuclei segmentation algorithm to determine the locations of the nuclei/cells in a tissue. The type of the feature extraction method to be deployed depends on the sample. In this case, computer vision allows us to segment a sample into an specified area of interest. On the other hand, the complex nature of histopathological tissue images means a challenge to characterize an individual nucleus as well as an entire tissue by aggregating the features of its nuclei. For this reason, we use color and morphological features to determine the exact locations of nuclei beforehand. Briefly, our objective is to establish a relationship between color characteristics and morphology for the classification of tissues with colorectal cancer. Furthermore, color and morphological methods allow us to analyze changes and quantify the size and shape characteristics of cell nuclei in order to construct a nuclei classification algorithm.

Finally, to determine the exact details of cells and to address this issue, in section 3 we detailed the next steps to show how we treat the difficulty of the complex nature of image scenes, i.e., stain related problems including lack of dark separation lines between a nucleus and its surroundings, inhomogeneity of the interior of a nucleus, and occurrence of non-nuclei stain artifacts in a tissue [34].

Chapter 3

METHODOLOGY

This chapter presents our proposed methodology for the extraction of color and morphological features and its implementation of nuclei classification in Colorectal Cancer (CRC) digital histopathological images using machine learning. In section 3.1, we explain the motivation behind the feature extraction techniques with a brief overview. Then, we explain the details of the methodology in section 3.2. Section 3.2 is divided into four main phases. Subsection 3.2.1 reviews the *dataset collection*. Then subsection 3.2.2 details the extraction of *image classification features*. In subsection 3.2.3, we emphasize on *modelling of cancer classification algorithms* and finally subsection 3.2.4, we explain the *assessment metrics* used to measure the performance of classification algorithms, to compare it later.

3.1 Motivation

In this approach, we apply artificial intelligence in the biomedical area to support colorectal cancer diagnosis using digitized histopathological images. The analysis of digital histopathological images often uses general image recognition technology but this has some problems, such as texture analysis since this approach is used in pixels, it is sensitive to noise in the values of pixels. To address this issue, this research focuses on the relationship between *morphological* and *color* characteristics. We take advantage of this type of information to classify images of histopathology.

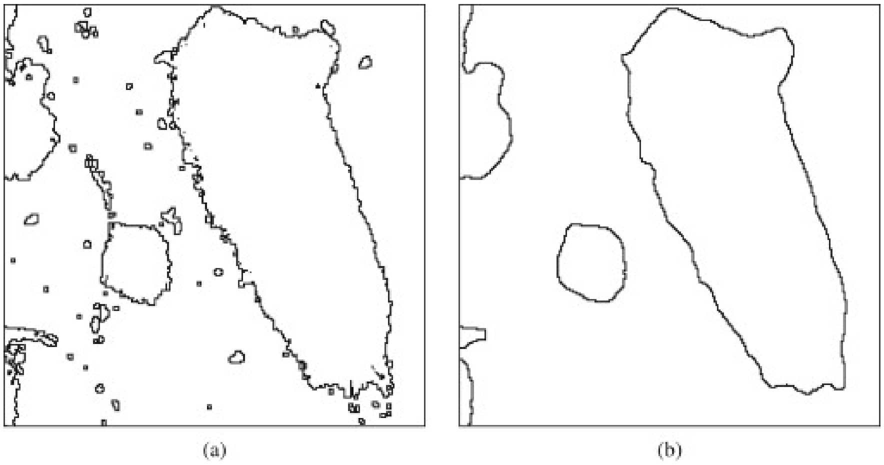


Figure 3.1: Object boundary detection using morphological filter. a) Image without filter b) Image with Filter [42].

Overall, *morphological* features are commonly used in boundary detection and to highlight its importance to preserve, uncover, and detect the geometric structure of image objects. Also, morphological features are quite relevant at shape analysis offering efficient solutions to apply denoising-filters like median-type and stack filters. In short, boundary detection becomes quite sensitive to small noise artifacts, for example, we present on Figure 3.1 an example of boundary detection and its properties for noise reduction, detail preservation, and artifact-free images [42].

On the other hand, *color* features are important for expressing saliency, and color uniqueness can be clearly incorporated into the design of image feature detectors. We analyze different color channels to uncover characteristics such as specific edges that only appear on determined color channel, according to the state of the art. Hence, for the most part detection methods are based on the statistical analysis of color derivatives [43]. Figure 3.2 presents results of a corner detector algorithm based on color feature trained accordingly where more pixels are located. Briefly, color is determined by concentration of pixels, it means intensity of each color channel (Red,Green,Blue for RGB images).

To achieve this, we work with a method split into four main steps. First, we obtain two digitized histopathological images dataset in order to select the types of tissue for analysis based on previous works [48] [36]. The second step is to construct a classification model able to determine if a nuclei is cancerous or not, from color and morphological

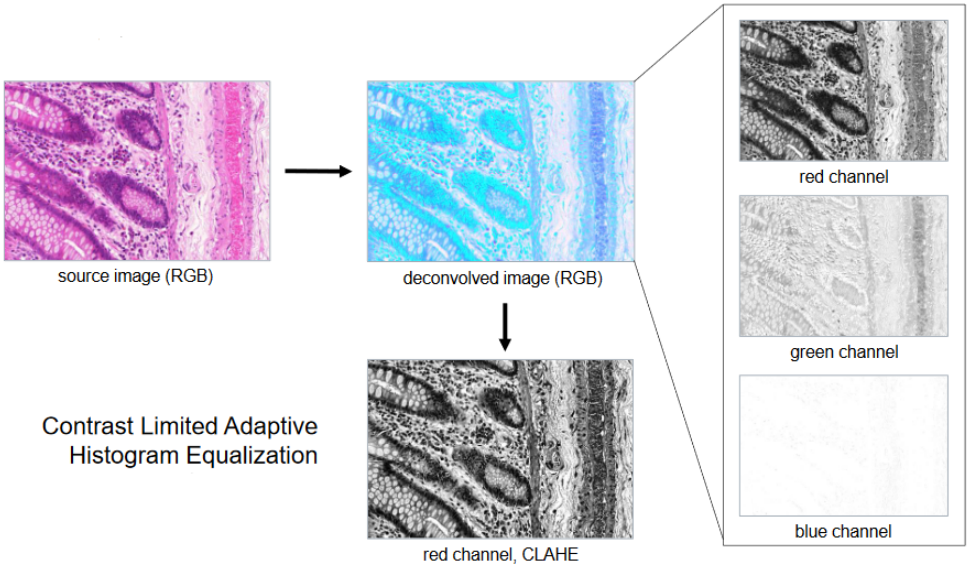


FIGURE 3.2: Color Channels - Color Deconvolution [45].

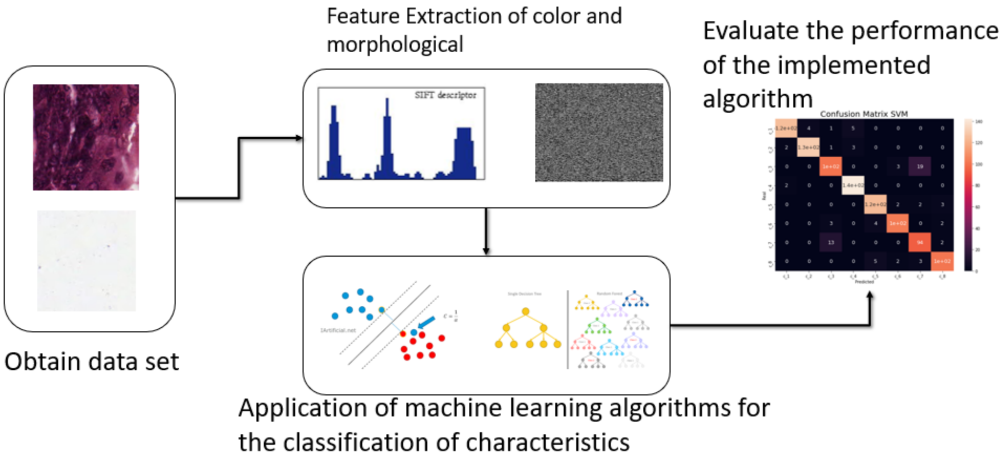

Feature Extraction of color and morphological
SIFT descriptor
Evaluate the performance of the implemented algorithm
Confusion Matrix SVM
Obtain data set
Application of machine learning algorithms for the classification of characteristics

FIGURE 3.3: Proposed Methodology divided by four steps

characteristics. Then, on third step, we apply the constructed model using machine learning algorithms to our dataset. Finally, at fourth step, we evaluate performance of algorithms through a comparison of four popular assessment metrics to detect changes and test the best performing algorithms. See Figure 3.3 where we provide the scheme of proposed methodology.

3.2 Proposed Methodology

3.2.1 Dataset Collection

First phase consists of obtaining two different datasets:

First Dataset - 8 Classes

The Medical Center of Mannheim (University of Heidelberg, Germany) provides us a database which contains 8 types of labels separated in folders on Figure 3.4. These classes were selected from the most representative images from our dataset. These ones shows the wide variation of illumination, stain intensity and tissue textures present in routine histopathological images [36]. Images were extracted from 10 independent samples of colorectal cancer (CRC) primary tumours.

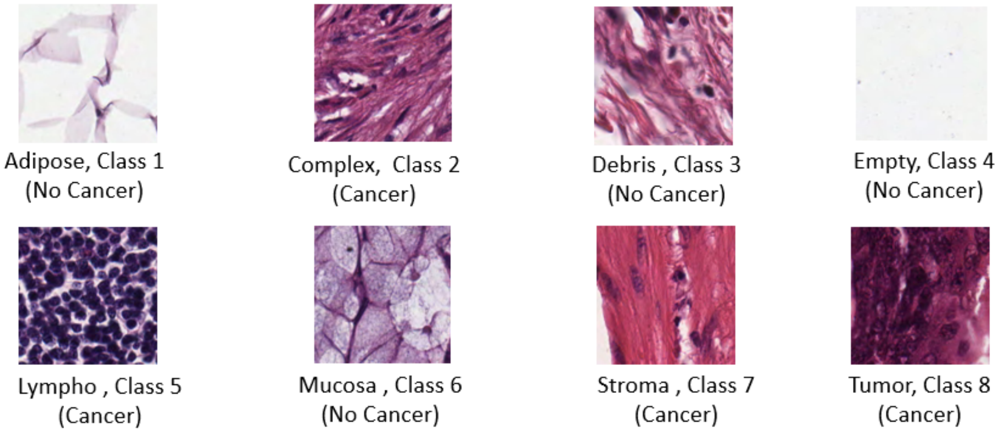

Adipose, Class 1 (No Cancer)
Complex, Class 2 (Cancer)
Debris , Class 3 (No Cancer)
Empty, Class 4 (No Cancer)
Lympho , Class 5 (Cancer)
Mucosa , Class 6 (No Cancer)
Stroma , Class 7 (Cancer)
Tumor, Class 8 (Cancer)

FIGURE 3.4: Selected types of cancer and non cancer cells for the study

Labeling histopathological images as cancerous or non-cancer regions is a key task in cancer diagnosis. It is also important to divide the cancer tissue into different categories to give a medical context. In this thesis, we analyze a dataset compound by anonymized H&E stained Colorectal Cancer (CRC) tissue slides. Each tissue image in Figure 3.4 is a label that explains if an image is cancer diagnosed or not, if it is the only tissue, or a part of the sheet, among others. However, supervised approaches require intensive work and time to obtain this labels. In this case, dataset used is already labelled and exhibited on Table 3.1. The slides were first digitized [37]. Moreover, contiguous tissue areas were manually annotated and tessellated [36], we use 625 non-overlapping tissue tiles of dimension $150px$ x $150px$ (74μm x74μm). In summary, the specifically required images correspond to previously labeled histopathology images of hematoxylin-eosin (H&E) from colorectal cancer. That data is covered by an MIT license [35].

Class	**Name**	**Diagnostic**	**Description**
1	Adipose	No Cancer	Adipose tissue
2	Complex	**Cancer**	Containing single tumour cells and/or few immune cells
3	Debris	No Cancer	Including necrosis, hemorrhage and mucus
4	Empty	No Cancer	No tissue
5	Lympho	**Cancer**	Immune-cell conglomerates and sub-mucosal lymphoid follicles
6	Mucosa	No Cancer	Normal mucosal glands
7	Stroma	**Cancer**	Homogeneous composition, includes tumour stroma, extra-tumoural stroma and smooth muscle
8	Tumor	**Cancer**	Tumour epithelium

Table 3.1: Description of selected types of cancer for the study.

Second Dataset - 2 classes

Second dataset was acquired from ”GlaS MICCAI’2015: Gland Segmentation Challenge Contest” databases. This dataset consists of 165 images derived from 16 *H&E* stained histological sections of stage *T3* or *T4* colorectal adenocarcinoma on images *.bmp*. Colorectal adenocarcinoma originating in intestinal glandular structures is the most common form of colon cancer. In clinical practice, the morphology of intestinal glands, including architectural appearance and glandular formation, is used by pathologists to inform prognosis and plan the treatment of individual patients [44].

Each segment belongs to a different patient. They were processed in the laboratory on different occasions. Digitization of these histological sections into whole slide images (WSI) was performed using a Zeiss MIRAX MIDI slide scanner with $0.465 \mu m$ pixel resolution. The WSIs were subsequently rescaled to a pixel resolution of $0.620 \mu m$ (equivalent to a 20X objective magnification).

A total of 52 visual fields of benign and malignant areas across the entire set of WSIs were selected to cover the widest possible tissue variety. An expert pathologist (DRJS) then rated each visual field as "benign" or "malignant", according to the general glandular architecture. The pathologist also delineated the boundary of each individual glandular object in that visual field. Second dataset labels are exhibited on Table 3.2.

Class	**Name**	**Diagnostic**	**Description**
1	Adenocarcinoma	Cancer	Malignant tumours arising from glandular epithelium
2	Benign Tissue	No Cancer	Healthy or benign sample

Table 3.2: Description of selected types of cancer for the study.

3.2.2 Image Features

The second phase involves the extraction of both morphological and color features. In Table 3.3, we present the corresponding features to analyze and build the model.

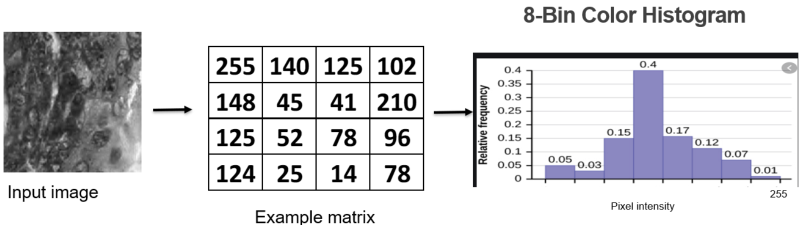
Color Features	**Morphological Features***
Mean	Area
Median	
Standard deviation	Perimeter
Skewness	
Kurtosis	Circularity
Energy	
Entropy	Eccentricity
Color Histogram (8 bins)	

Table 3.3: Features to extract. * For each of the morphological characteristics, a mean and standard deviation is obtained

- **Color Features**

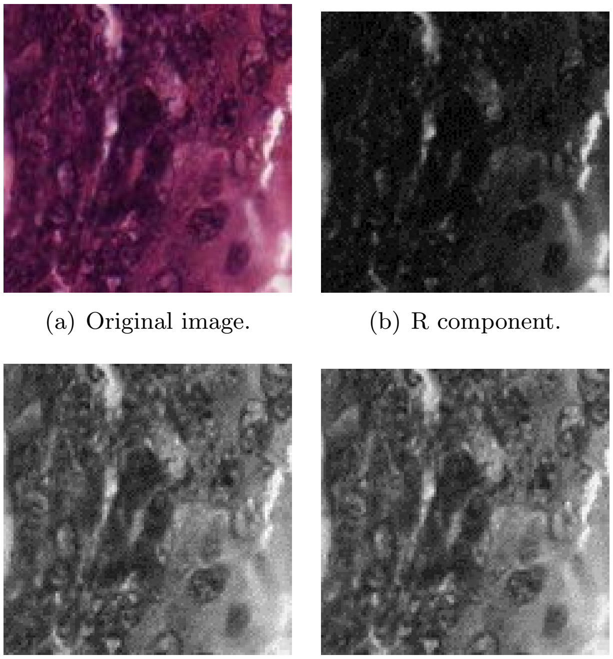
  The obtention of the color characteristics will be done with first-level statistics, which are the mean, median, standard deviation, kurtosis, skewness, and the color histogram (which obtains an even quantity of distribution and normalizes). The previous process is done for different color spaces such as HSV, RGB, YUV, and grayscale. Figure 3.6 displays the process for the RGB space, Figure 3.7 shows the process for the HSV components and Figure 3.8 exhibits the process for the YUV components.

  Also, we extract a color histogram in Figure 3.5. This is fundamental to extract valuable information before starting a process of core segmentation. After having all channels divided, we proceed to segment image and obtain the statistical measurements.



255	140	125	102
148	45	41	210
125	52	78	96
124	25	14	78

Figure 3.5: Sample color histogram



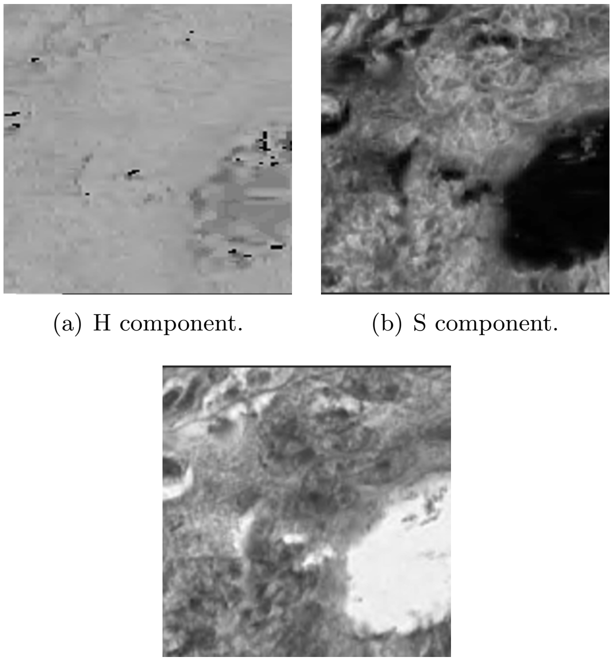
(a) Original image.

(b) R component.

(c) G component.

(d) B component.

Figure 3.6: Decomposition of an image in the RGB components.

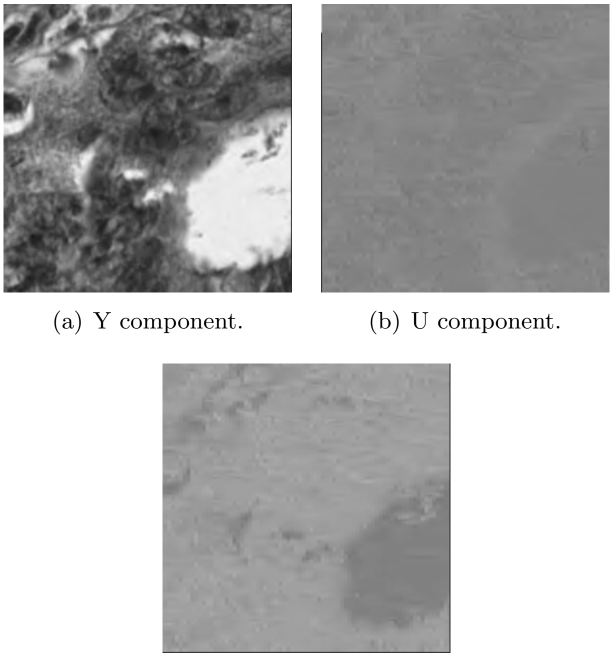


(a) H component.

(b) S component.

(c) V component.

Figure 3.7: Decomposition of a image in the HSV components.



(a) Y component.

(b) U component.

(c) V component.

Figure 3.8: Decomposition of a image in the YUV components.

- **Morphological Features**

  On the other hand, morphological features are extracted through a nucleation segmentation. This nucleation segmentation process is carried out by creating a hematoxylin and eosin color space; then, the color component in hematoxylin is taken out applying a watershed algorithm to detect nuclei as shown in Figure 3.9. Besides, morphological characteristics have to undergo a process of core segmentation. After having the segmented image we proceed to obtain the statistical measurements mentioned above. The extraction of the main features, such as circularity, eccentricity, solidity, area, and perimeter, is performed — features from which the mean and standard deviation are also extracted.

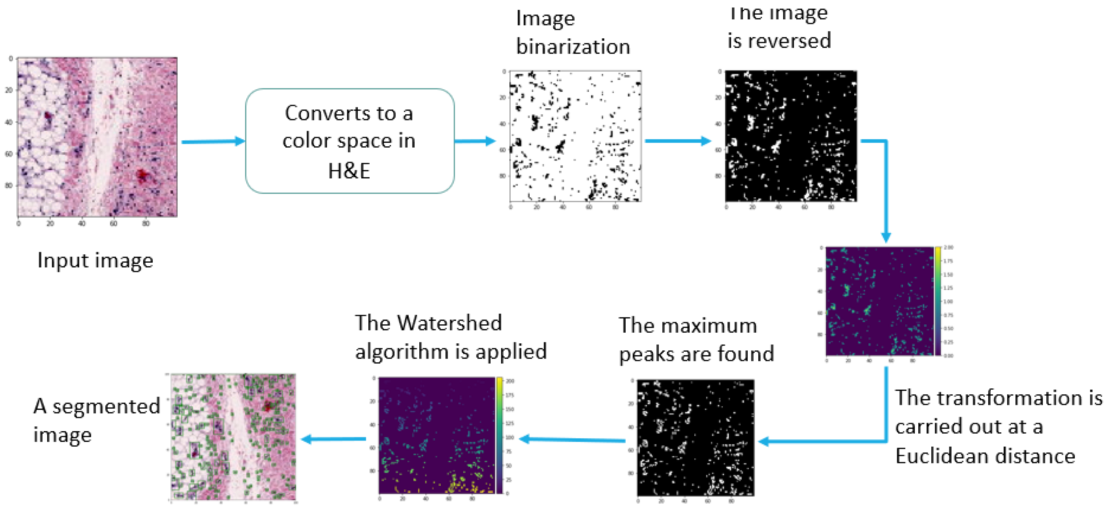

Input image
Converts to a color space in H&E
Image binarization
The image is reversed
The transformation is carried out at a Euclidean distance
The maximum peaks are found
The Watershed algorithm is applied
A segmented image

Figure 3.9: Nuclei segmentation.

After obtaining the characteristics of histopathological images, we need to mine the useful data. Thus, we create a clean dataframe on python, using *pandas* library to prepare features to being spare for next classification algorithms. In order to perform machine learning, the categorical variables in our datasets are usually regarded as discrete entities and coded as feature vectors. "Dirty" unorganized data will produce categorical variables with redundancy: multiple categories reflect the same entity [38]. This problem is solved by converting categorical data into numbers. We apply **one hot encoding**. It is a representation of categorical variables as binary vectors, where each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1. A one hot encoding allows the representation of categorical data to be more expressive, and causes that redundancy to learning algorithms will be deleted, bringing significant benefits to our model performance.

Additionally, we reduce size of both datasets by dropping variables which do not represent relevant data to build our model (i.e. "ID_of_sample"). Along with this, we evaluate if obtained datasets are balanced, it means that each class has equal number of samples and allow to build a most reliable model.

Finally, we separate each clean dataset into 80% for *training set*, and 20% for *test set*. Both are filled with random values of corresponding clean dataset. This methodology allows to find the appropriate performance of the next classification algorithms, where, in the end, we seek to have a generalization of the detection problem and obtain a classifier model as output of each algorithm.

3.2.3 Cancer Classification Algorithms

As a third stage, we build a classification of colorectal cancer diagnosis model based on machine-learning algorithms. To begin with, it is important to define an input and a labels set to feed our classification algorithms. Input set is conformed by the features dataset (already prepared as stated in 3.2.2), and labels set will be determined for number of classes of each dataset, eight classes for first and two classes for second dataset, as described in section 3.2.1.

Four machine learning algorithms of *supervised type* are considered to be tested with this image dataset. The algorithms are the Naive Bayes, Random Forest, Support Vector Machine (SVM) and Multilayer Perceptron (MLP).

- **Naive Bayes Algorithm**

  It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. The reason why it is called 'Naive' because it requires rigid independence assumption between input variables. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes Classifier can be trained easily and fast and can be used as benchmark model. Also, Naive Bayes requires a strong assumption of independent predictors, so when the model has a bad performance, the reason leading to that may be the dependence between predictors.

  We deploy this algorithm on python using **Sklearn** library, constructing a classifier model with the command *"GaussianNB"* adjusted with default values (Ignoring prior probabilities of the classes and a portion of the largest variance of all features equal to $1e-9$).

- **Random Forest Algorithm**

  Random forest is a supervised classification algorithm, which consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

In this case, we also consider **Sklearn** library to set up the classifier model. The command *"RandomForestClassifier"* is adjusted with default values, except the maximum depth of the tree determined 50, and a value of zero (0) to the randomness of the bootstrapping [1].

- **Support Vector Machine Algorithm (SVM)**

  Support Vector Machine (SVM) analyzes data for regression and classification analysis. The objective of SVM is to find a surface in the n-space that separates the space in regions. The shape of the surface depends on the type used to make the separation. The surface is usually known as kernel. In the simplest form, the kernel is lineal and represents a hyperplane the space of dimension n.

  At this part, we continue using **Sklearn** library. In order to ensemble a SVM model we make use of command *"make_pipeline"* adjusted to default values and *"gamma SVC"* value set on *auto*.

- **Multilayer Perceptron Algorithm (MLP)**

  It is an artificial neural network as explained in section 1.3, consisting of many layers such that it allows to solve problems that are not linearly separables. It consists of an input layer, intermediate layers and an output layer. The set of the layers represent non linear functions.

To set up this model, we call **TensorFlow** library and use a series of commands to define the architecture of our MLP. We establish a 3-layer-sequential model configured with parameters to training our data exposed on Table 3.4 where each "Dense" means "Number of neurons of present layer".

[1] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Parameter	**Value**
Dense 1	100
Dense 2	50
Dense 3	8
Layer 1,2 -Activation Function	ReLu
Layer 3 - Activation Function	Softmax
Epochs	30
Validation Split	10%

Table 3.4: Parameters of MLP model proposed (Training set).

3.2.4 Assessment Metrics

To measure the performance of the machine-learning algorithms, we use the concepts of accuracy, Precision, Recall and F1-Score. The definition of these is as follows:

- **Accuracy** In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated. Its values are determined by Equation 3.1.

$$Accuracy = \frac{CorrectPredicitions}{TotalPredictions} \tag{3.1}$$

- **Precision**

  Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class, defined by Equation 3.2.

$$Precision = \frac{TruePositive}{PredictedPositive} \tag{3.2}$$

- **Recall**

  Recall is used to measure the fraction of positive patterns that are correctly classified.

Its values can be obtained using Equation 3.3.

$$Recall = \frac{TruePositive}{ActualPositive} \tag{3.3}$$

- **F1-Score**

  F1-Score is the harmonic mean between *Precision* and *Recall*, where the average is calculated per label (class), then, averaged across all labels Q (classes). If p_j and r_j are the *Precision* and *Recall* for all $\lambda_j \in h(x_i)$ from $\lambda_j \in y_i$, the *F1-Score* is represented by Equation 3.4.

$$F1Score = \frac{1}{Q}\sum_{j=1}^{Q}\frac{2p_j r_j}{p_j + r_j} \tag{3.4}$$

Chapter 4

RESULTS AND DISCUSSION

In this chapter, we present the results of the applied models for cancer classification in Section 4.1. In addition, we also describe an experimental algorithm to reduce dimension of the dataset before the classification. We apply the same calsification model on two different datasets to discuss and compare its performance and measure robustness of the built model. This way, we can obtain a better approach after comparing its results. Finally, in section 4.2, we compare classification performances on these two different datasets based on the metrics that we have previously discussed.

Before presenting results, it is important to remember the two datasets used.

- **First Dataset - 8 Classes**: This dataset consists of 8 classes previously described in section 3.2.1, containing 5000 rows and 161 of visual features: 10 features correspond to morphological features, and around 150 to color.

- **Second Dataset - 2 Classes**: A dataset of 165 images derived from 16 *H&E* stained histological sections of stage *T3* or *T4* colorectal adenocarcinoma. The visual field is rated as "Cancer" or "No Cancer", according to the general glandular architecture.

4.1 Cancer Classification Results

4.1.1 Experimental dimensionality reduction algorithm

This **Experimental stage** consists of applying a dimensionality reduction algorithm to reduce the size of the data without significantly reducing the useful information based on the variance analysis and finding the direction in which the variance changes in a maximum way. In our case, we use the Principal Component Analysis (PCA), which is one of the most used algorithms in CAD field [40] [39]. We intend to decrease computational load searching an optimal dimension that contains most useful information. However, after dimentionality reduction, every classification model designed provide less than 0.6 values for every performance measurements. Considering this, we assume that reduced dataset does present abnormalities on used features, which affect classification and suffer a high level of information loss in the distribution of tissue components, especially on morphological features.

Because of low performance using the PCA - dimensionality reduction algorithm, we only use both entire datasets on next stage.

4.1.2 8-Classes Dataset - Clasiffication Results

The four machine learning algorithms are applied to the first dataset (8 classes). The aim is to compare performances through SVM, MLP, Random Forest and Naive Bayes algorithms.

The following Table 4.1 presents results of performance for the algorithms. It should be noted that SVM algorithm attains a better result than the other three algorithms when compared by Accuracy, Precision and Recall. When compared by F1-Score we observed an outstanding performance of Random Forest and SVM approaches. Nevertheless, every algorithm reach a performance above 0.8, leaving the Accuracy value of MLP with 0.89 as the lowest value.

Algorithm	Accuracy	Precision	Recall	F1-Score
Support Vector Machine (SVM)	0.98	0.98	0.98	0.98
Random Forest	0.95	0.96	0.95	0.95
Naive-Bayes	0.93	0.93	0.93	0.93
Multilayer Perceptron (MLP)	0.84	0.89	0.84	0.84

Table 4.1: Result of the performance for the 4 algorithms - 8-Classes Dataset

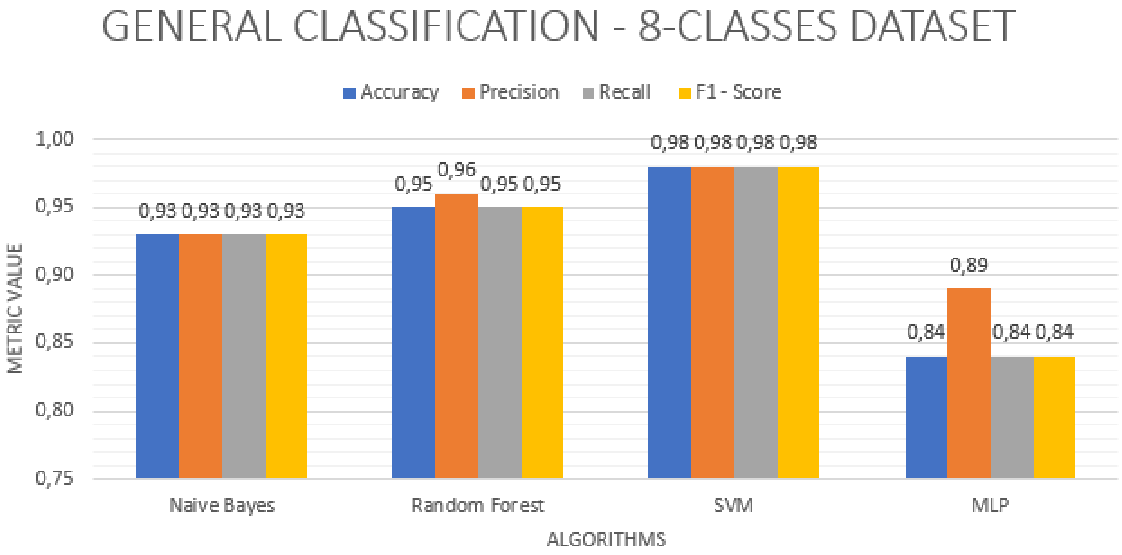


Figure 4.1: General Performance of the 4 algorithms - 8-Classes Dataset.

The performance of the four algorithms is also computed taking each class into account. The result of the four metrics of general classification of each algorithm is shown in Figure 4.1. From this figure we concluded that the best classification is carried out by the SVM and Random Forest above 0.95. In contrast, results Naive Bayes have a mean value of every measurement below 0.93 and the worse mean value (0.85) is achieved by MLP algorithm.

Subsequently, in Table 4.2, it is noticeable that the performance of F1-score for each class classified by each algorithm. It is important to include precision and recall, to appreciate proportions of F1-score and detail its relation between classes.

Furthermore, in Table 4.2 there are some details about which class represents a challenge to our models. Histopathological images have visual patterns with particularities that makes its analysis difficult. Some of these challenges may be found in sixth class for MLP algorithm, where precision is 0.64. This means MLP is not precise to predict a true positive value of ”Mucosa” a ”not cancer” class. However, there are four values in the Table 4.2 under 0.7 (Classes 2,3,5,6) for MLP .

Likewise, SVM represents a reliable computational diagnostic tool by evaluating its quantitative measures. In most cases, proposed metrics achieve a near value to 1 of F1-score. In short, SVM performs a model capable of classifying correctly each classs of dataset.

Following, it is possible to say that Random Forest algorithm is, a partially reliable computational diagnosis tool. Although this algorithm performs a value below 0.9 only in 3 cases, it is capable of classifying this large set of features and demonstrate that it requires less computational load to achieve a 1 value of precision or recall on classes apparently easy to distinguish.

Naive Bayes algorithm also has a reliable performance with metrics above of 0.9. Similar to MLP, Naive Bayes presents some weakness to classify fifth and sixth class. Finally, MLP presents an acceptable performance, but it is the only approach not sufficient to make it a reliable CAD tool.

In Figure 4.2, it can be seen the performance measured by precision for each class to clarify results between algorithms. Similar to this, Figure 4.3 shows performance measured by recall and finally Figure 4.4 presents F1-Score. At this time, we notice in F1-Score that there is a singular characteristic that makes the class 8 easy to classify with the algorithms of Random Forest, Naive-Bayes, SVM and MLP with very high values of Precision, Recall and F1-Score which are around 0.99.

Additionally, we seek to evaluate confusion matrix of all the algorithms to demonstrate classification challenges and similarities through classes.

Algorithm	Class	Precision	Recall	F1-Score
SVM	1	0.98	0.98	0.98
	2	0.95	0.99	0.97
	3	0.98	0.94	0.96
	4	0.97	0.99	0.98
	5	0.97	0.98	0.97
	6	0.98	0.97	0.98
	7	0.99	0.98	0.99
	8	0.98	0.98	0.98
Random Forest	1	0.76	0.98	0.86
	2	0.99	0.95	0.97
	3	0.98	0.92	0.95
	4	0.99	1	1
	5	0.99	0.94	0.96
	6	1	0.85	0.92
	7	0.99	0.96	0.97
	8	0.99	0.98	0.99
Naive-Bayes	1	0.98	0.95	0.97
	2	0.91	0.93	0.92
	3	0.9	0.92	0.91
	4	0.98	0.99	0.98
	5	0.89	0.88	0.88
	6	0.81	0.87	0.84
	7	0.98	0.92	0.95
	8	0.98	0.98	0.98
MLP	1	1	0.77	0.87
	2	1	0.52	0.69
	3	0.66	0.98	0.79
	4	0.85	0.98	0.91
	5	0.99	0.65	0.78
	6	0.64	0.99	0.78
	7	0.97	0.82	0.89
	8	1	0.98	0.99

Table 4.2: Result of the performance for the four algorithms at the level of Class.

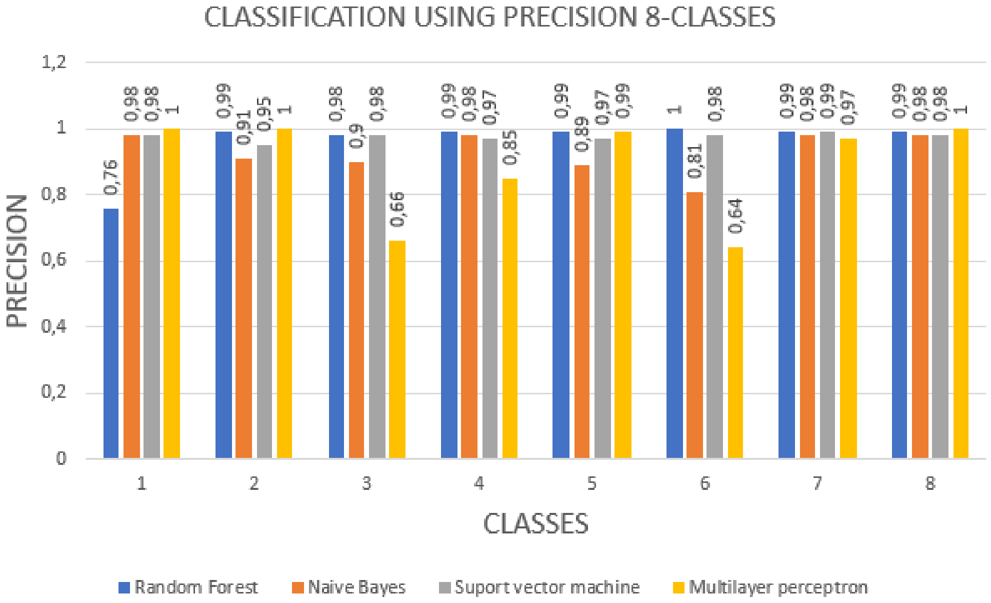


Figure 4.2: Performance of the 4 algorithms using Precision - 8-Classes Dataset.

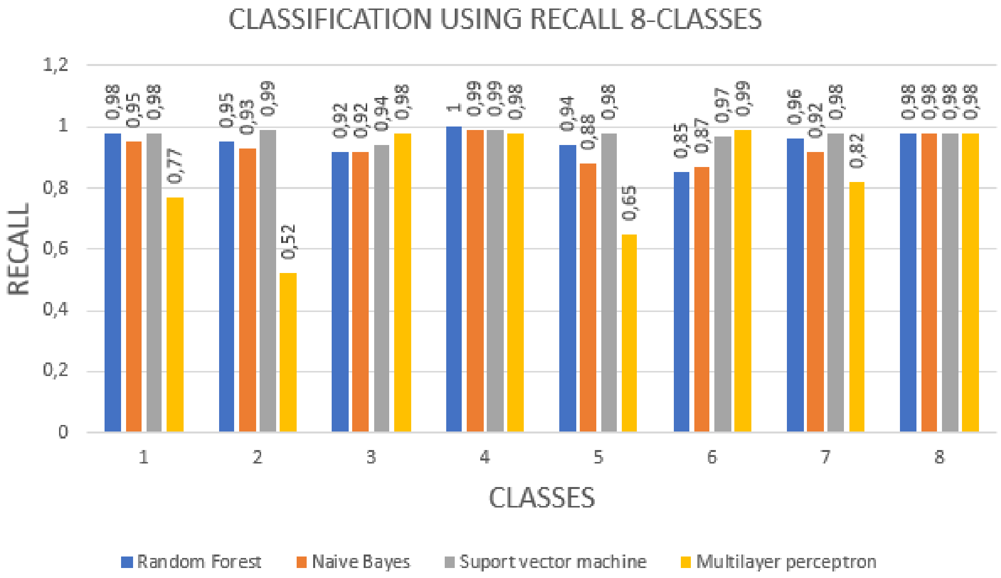


Figure 4.3: Performance of the 4 algorithms using Recall - 8-Classes Dataset.

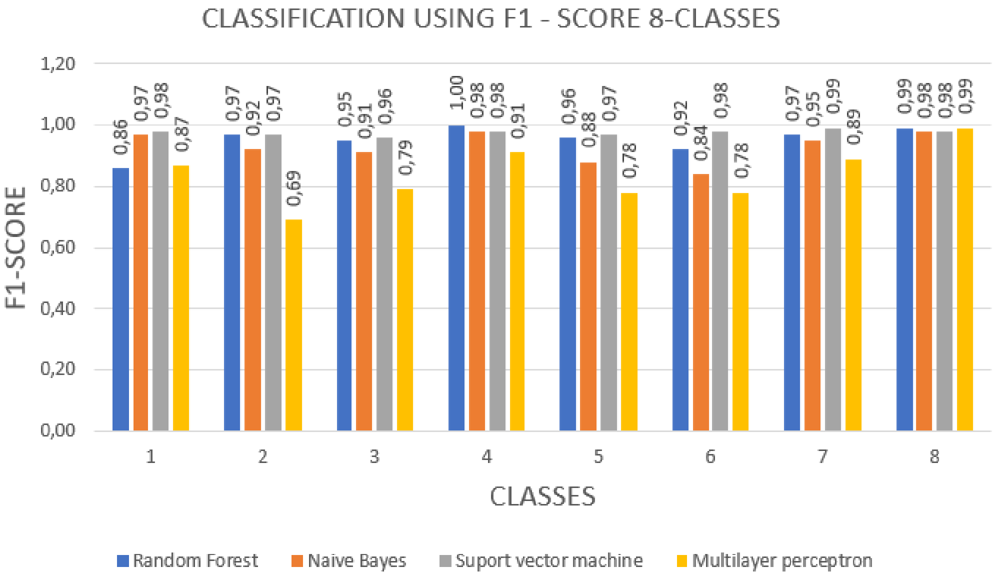
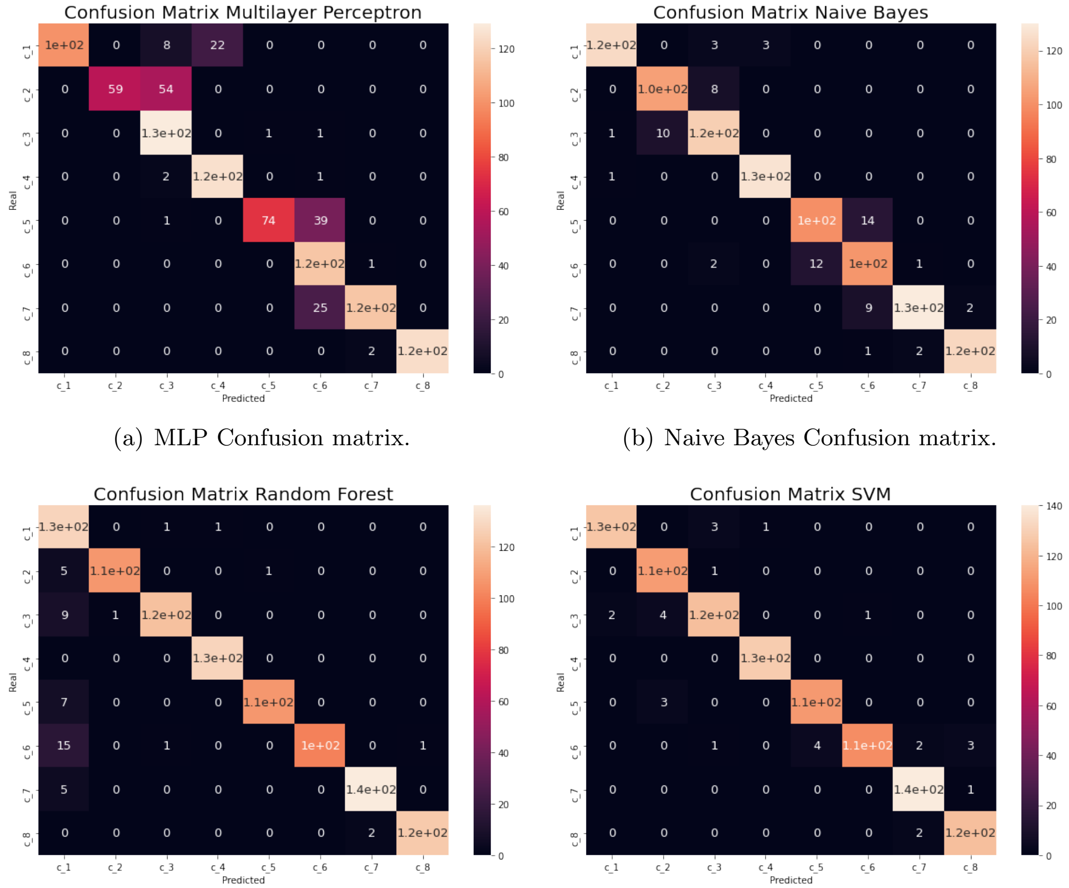


Figure 4.4: Performance of the 4 algorithms using F1-Score - 8-Classes Dataset.

First, Figure 4.5(a) describes the confusion matrix of MLP, it can be seen that *class 2* is easily confused with *class 3* this confusion may be caused because of model fails to distinguish similar morphological characteristics (shapes) within "Debris"(3) and "Complex" (2) classes. Also, *class 6* has an closer value to *class 5*. In this confusion, we observed that there is apparently no common characteristics that can cause a classification issue, it might be interpreted as a model deficiency probably caused by a model validation error. Meanwhile, other classes achieve higher values and well-performed results, but, the confused classes are crucial to diagnose if sample has cancer or not, thus model gives acceptable results through 8-classes but cannot be used in real life, it has to keep improving results by researching over other datasets.

In the same way, the confusion matrix of Naive Bayes algorithm, on Figure 4.5(b), shows a strong classification where we can see clearly differences between classes. However, *class 5* and *class 6* are nearly confused, but it does not represent an issue classification between classes.

In Figure 4.5(c) Random Forest tends to get more confused when classifying the "Mucosa" (6) class, classifying it as "Adipose" (1). This can happen because there is little variability between shapes of "Adipose" and "Mucosa" classes. Nevertheless, most samples



Confusion Matrix Multilayer Perceptron

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
c_1	1e+02	0	8	22	0	0	0	0
c_2	0	59	54	0	0	0	0	0
c_3	0	0	1.3e+02	0	1	1	0	0
c_4	0	0	2	1.2e+02	0	1	0	0
c_5	0	0	1	0	74	39	0	0
c_6	0	0	0	0	0	1.2e+02	1	0
c_7	0	0	0	0	0	25	1.2e+02	0
c_8	0	0	0	0	0	0	2	1.2e+02

(a) MLP Confusion matrix.

Confusion Matrix Naive Bayes

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
c_1	1.2e+02	0	3	3	0	0	0	0
c_2	0	1.0e+02	8	0	0	0	0	0
c_3	1	10	1.2e+02	0	0	0	0	0
c_4	1	0	0	1.3e+02	0	0	0	0
c_5	0	0	0	0	1e+02	14	0	0
c_6	0	0	2	0	12	1e+02	1	0
c_7	0	0	0	0	0	9	1.3e+02	2
c_8	0	0	0	0	0	1	2	1.2e+02

(b) Naive Bayes Confusion matrix.

Confusion Matrix Random Forest

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
c_1	1.3e+02	0	1	1	0	0	0	0
c_2	5	1.1e+02	0	0	1	0	0	0
c_3	9	1	1.2e+02	0	0	0	0	0
c_4	0	0	0	1.3e+02	0	0	0	0
c_5	7	0	0	0	1.1e+02	0	0	0
c_6	15	0	1	0	0	1e+02	0	1
c_7	5	0	0	0	0	0	1.4e+02	0
c_8	0	0	0	0	0	0	2	1.2e+02

(c) Random Forest Confusion matrix.

Confusion Matrix SVM

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
c_1	1.3e+02	0	3	1	0	0	0	0
c_2	0	1.1e+02	1	0	0	0	0	0
c_3	2	4	1.2e+02	0	0	1	0	0
c_4	0	0	0	1.3e+02	0	0	0	0
c_5	0	3	0	0	1.1e+02	0	0	0
c_6	0	0	1	0	4	1.1e+02	2	3
c_7	0	0	0	0	0	0	1.4e+02	1
c_8	0	0	0	0	0	0	2	1.2e+02

(d) SVM Confusion matrix.

FIGURE 4.5: Confusion matrix of four algorithms - 8-Classes Dataset.

are correctly classified. Hence, this model does not represent an inefficient model, because although it confuses slightly between these classes, they are ”no cancer” classes and does not represent a severe CADx problem. Subsequently, Figure 4.5(d) shows SVM results. Unlike other models, SVM algorithm represents a robustness on classifying each class.

After comparison using general classification and confusion matrix, it is evident that the best algorithm to classify is the support vector machines having a F1-score of 0.98. In general, each model is a classical method that perform well. However, SVM has better support for working with large number of dimensions. In this case, both Random Forest algorithm and Support Vector Machine are better at classifying these types of features on 8-classes samples.

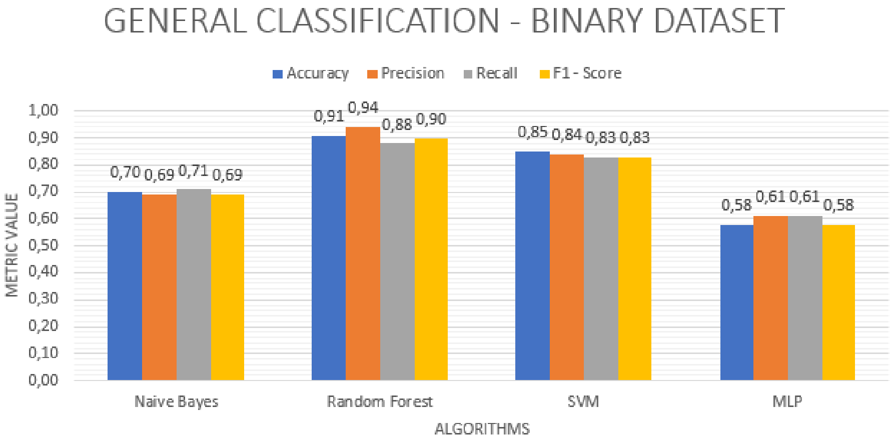


Figure 4.6: Performance of the 4 algorithms - Binary Dataset

4.1.3 Binary Dataset - Clasiffication Results

In this case, performance of SVM, MLP, Random Forest and Naive Bayes algorithms trained on **second dataset (2 classes)** is compared and evaluated.

The performance of the four algorithms is computed to classify into ”cancer” and ”non-cancer”. Overall, results show a great performance for SVM and Random Forest algorithm with metrics above 0.8. From Figure 4.6, we conclude that the best classification is carried out by the Random Forest with scores over 0.9. In contrast, results of MLP and Naive Bayes have a mean value of every measurement below 0.7.

In feature selection, a subset of features like color and morphology data can be selected to demonstrate and represent relevant characteristics from a dataset of small size to perform cancer diagnosis. However, small dataset carries weight training models and gives a proper output, avoiding the exhaustive process of searching a practical and big tissue dataset.

In case of the random forest algorithm, all its performances metrics are above 0.85, while the other algorithm have scores between 0.7 and 0.85. Next algorithms sorted by metrics are SVM, Naive Bayes and MLP respectively.

Besides, we add Figure 4.7 to see precision achieved for each algorithm to observe its details of classification on binary dataset. Figure 4.8 shows performance measured by recall and Figure 4.9 by F1-Score. As a result, we notice in precision and recall values that are variable among classes, and achieving F1-Score value of 0.9. From Figure 4.8, we can perceive that all models experience low scores on 0-class (”no cancer”) failing to reach scores higher than 0.75. Similarly, from Figure 4.7, we can notice the variance in scores in the algorithms. With the algorithms of Random Forest, Naive-Bayes, SVM and MLP with variable values of Precision, Recall and F1-Score we can say that there is a lack of data to feed to each model. Nevertheless, SVM and Random Forest achieve outstanding results and demonstrates that it is possible and reliable to perform cancer diagnosis, with only morphological and color features.

Also, we evaluate confusion matrix of algorithms to see its performance to distinguish between ’cancer’ or ’not cancer’.

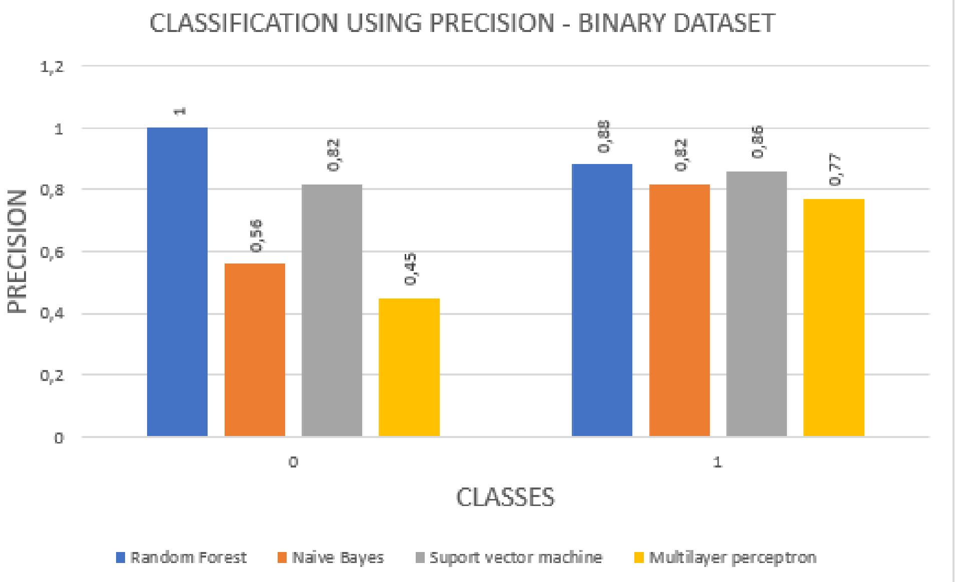


FIGURE 4.7: Performance of the 4 algorithms using Precision - Binary Dataset.

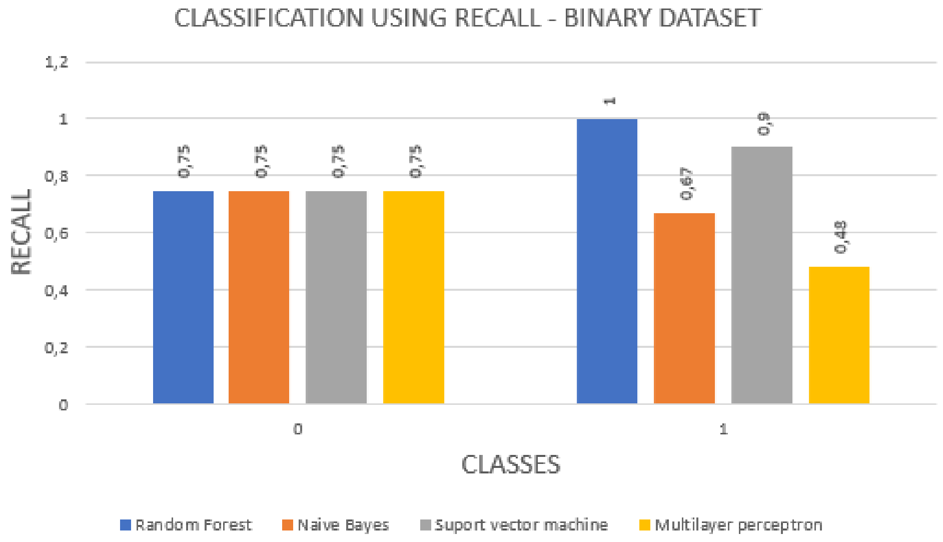


FIGURE 4.8: Performance of the 4 algorithms using Recall - Binary Dataset.

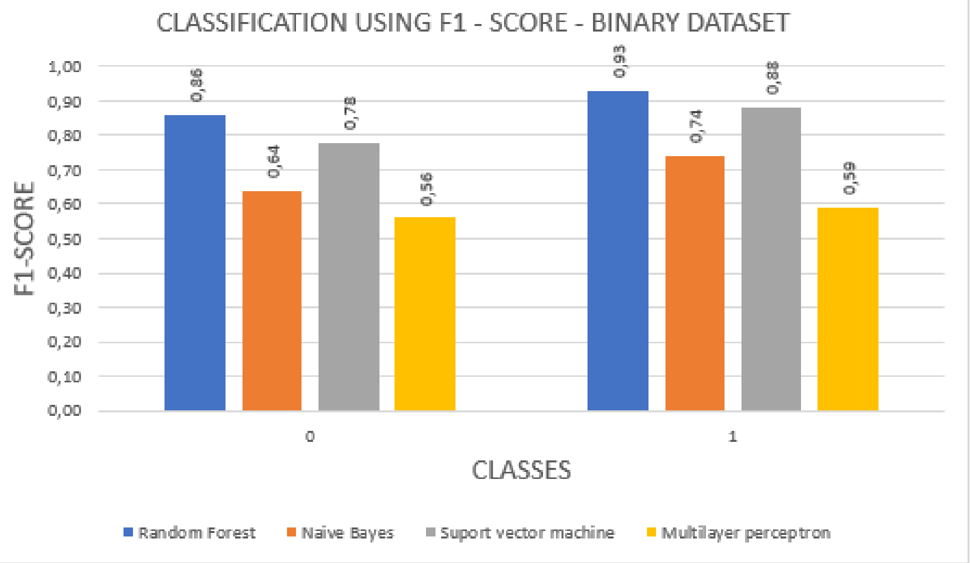
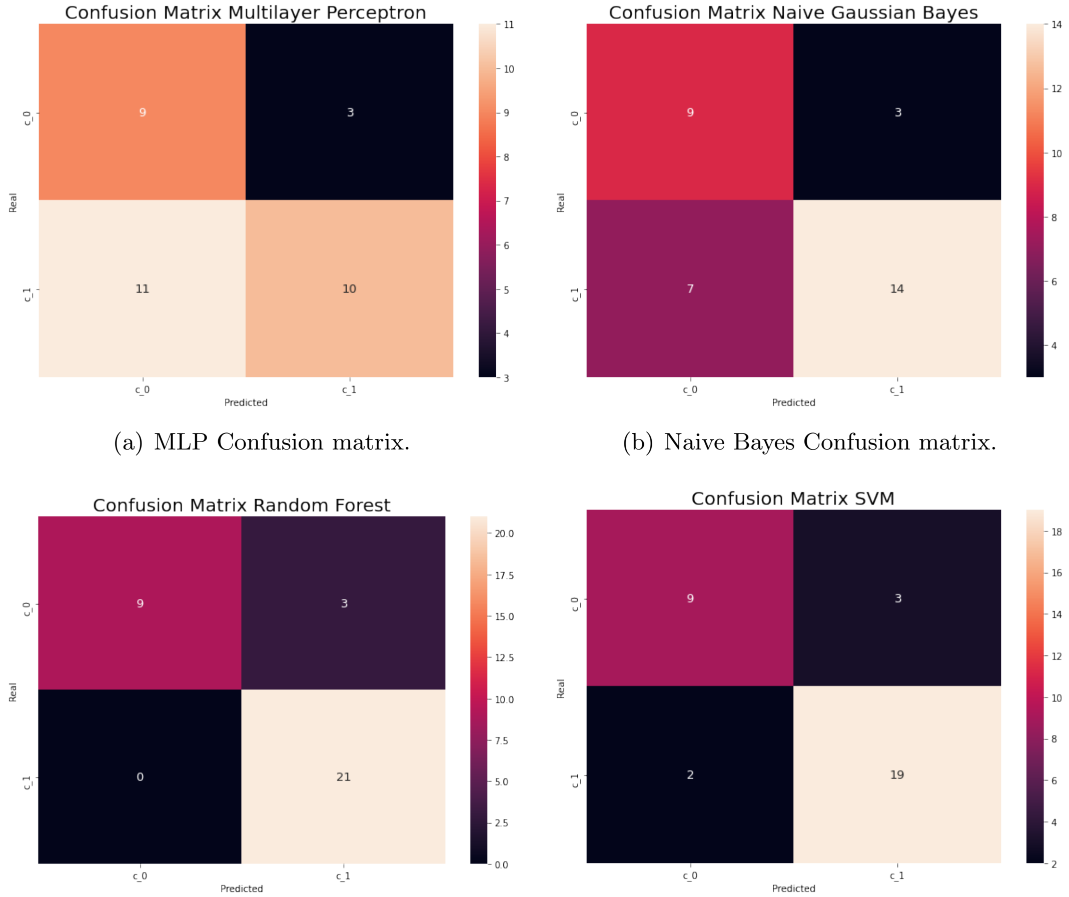


FIGURE 4.9: Performance of the 4 algorithms using F1-Score - Binary Dataset.

Differing from Subsection 4.1.2, Binary dataset represents 2-classes: "c_0 =cancer" or "c_1 =not cancer". First, MLP shows its confusion matrix in Figure 4.10(a) which describes that can it easily confuse "cancer" class with "not cancer", while "not cancer" is classified well by this algorithm.

Alternatively, on Figure 4.10(d) Random Forest and Support vector machines tend to get a better approach to classify classes, both can differentiate clearly if tissue sample has



(a) MLP Confusion matrix.

(b) Naive Bayes Confusion matrix.

(c) Random Forest Confusion matrix.

(d) SVM Confusion matrix.

Figure 4.10: Confusion matrix of four algorithms - Binary Dataset.

cancer or not. Among these algorithms, Random Forest is more precise than SVM.

Also, In Figure 4.10(b) Naive Bayes shows a similar result as the Random Forest algorithm on Figure 4.10(c), except Naive Bayes presents a higher level of confusion of ”no cancer” class. After comparison using classification metrics and confusion matrix, it is evident that the best algorithm to classify is Random Forest having an accuracy of 0.91 and a F1-score of 0.89.

4.2 Comparison

In this section, we compare algorithms and its results between the classification of two datasets.

Firstly, we compare Naive Bayes and MLP algorithm results from Figure 4.1 and Figure 4.6, their metric values are over 0.8 on first dataset and they are clearly reduced to 0.7-0.6 in second case. Also, it means that these models could get a better adjust, for example, we may adjust the MLP model adding a new layer or changing activation function of final layer. Similarly, on Naive Bayes algorithm, it could be useful replacing ”Gaussian” model and replacing with ”Multinomial” configuration. It is important to emphasize that second dataset is relatively smaller, which represents the reason of the difficulty to have better training (misclassification).

Subsequently, from the same Figures 4.1 and 4.6, when comparing Random Forest and SVM, it is easy to distinguish classes from each other on 8-class dataset using SVM. However, SVM algorithm faces a low level of classification on binary dataset in contrast to 8-class dataset, taking its mean metric values from 0.9 with its 8-classes classification performance to 0.83 in second dataset. On the other hand, Random Forest deploys a good performance with both datasets. In fact, in each case Random Forest's metric values stand above 0.9 and distinguishes without much confusion among classes on confusion matrix.

After analyzing metrics, we propose a way to prevent misclassification of color and morphological features, especially on second dataset. It is necessary to increase training samples. This will force the model to learn more number of samples and increase the number of models trained on the path.

To summarize, we see that yields of binary dataset have improved, because of the reduced number of classes regardless of less samples of dataset. In this case, there is a better classification, but also, it is possible to be doubtful due to a small dataset and its validation set.

Finally, the results show that the Random Forest Algorithm performed the best, when applied to both large and small dataset, considering the number of features detected properly. The MLP approach indicates its worst-case performance in 8-classes and binary dataset particularly due to less number of samples considering a neural network.

Chapter 5

CONCLUSIONS AND FUTURE WORK

The Random Forest algorithm has best performance with respect to Accuracy, Precision and Recall values above 0.95 for first dataset, and 0.9 for binary dataset.This feature extraction approach is tested on two different datasets to demonstrate the computational load and classification capacity is more effective to keep the accuracy high when less training data are used for learning. This is due to the ability of the random forest algorithm to increase the learner's generalization ability by increasing the size and variation of the training data.

Up to a point, SVM performs a similar result to Random Forest. As is it known, SVM is a traditional classification algorithm and its theory is based on the use of a hyperplane and dimension conversions where it is sought to draw a line to separate the classes. The possible reason is that the algorithm was able to separate the classes in the SVM hyperplane.

Different types of experiments can be carried out where an improvement of the configuration is done, thus making the algorithms more precise and accurate. Also, it is planned to deploy experiments with more images of adecarcinoma to increase samples an improve binary classification.

We also plan to compare other machine learning methods, making a complex supervised classification focusing on combining models Random Forest and SVM. We may explore the use of other adjustments that make a stronger model such as increasing depth in Random forest algorithm. Additionally, SVM can be improved using a sequence of more appropriate adjustment on its scaler capable of selecting points and such feature vectors more accurately, even with a reduced dataset.

Traditionally, the same methodology has been applied to classify images with cancer but with characteristics of texture, in this work it was possible to determine that the color and morphological characteristics are also useful for this purpose, opening another possibility for the study of this type of images. Including color and morphological features to colorectal cancer classification allows to have high algorithms performance for certain classes.

This thesis successfully addresses the issue of having limited labeled training data in the domain of histopathological tissue image classification. To this end, it presents a nuclei classification algorithm that performs a F1-score higher than 0.9 on Random Forest and SVM approach. We achieve our goal of contributing to medical CADx field, by using characteristics of color and shape to classify histopathology images. In addition, we described how characteristics of color and shape can be used percieve colors and shapes that the human eye cannot see to classify histopathology images.

To summarize, Random forest method performs a precise classification of cancer tissues. We can use its results to apply on a higher dataset with undesirable changes like morphological and color similarities, and we could improve model design to be capable of create an overall structure of potential results. This may benefit the CADx in automated tasks with the use of less features to extract and obtain a truly diagnosis type on posterior colorectal researches.

Finally, it is specifically designed for histopathological images of colon tissue, the proposed method may be used in different types of images and different types of organizations. This can also be considered as the future research direction of this article.

Bibliography

[1] Marr, D.: Vision: A computational investigation into the human representation and processing of visual information. San Francisco, CA, W.H. Freeman (1982).

[2] Thevenot, J., López, M. B. & Hadid, A.: A Survey on Computer Vision for Assistive Medical Diagnosis From Faces, IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1497-1511 (09 2018).

[3] Szeliski, R.: Computer Vision: Algorithms and Applications. 1st. ed. Springer-Verlag, Berlin, Heidelberg (2010).

[4] Doi, K.: Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. Computerized Medical Imaging and Graphics, vol 31, issues 4–5, pp. 198-211 (2007).

[5] Bluteau, R.: Obstacle and Change Detection Using Monocular Vision. Electronic Theses and Dissertations (2019).

[6] Kobatake, H.: Future CAD in multi-dimensional medical images: – Project on multi-organ, multi-disease CAD system. Computerized Medical Imaging and Graphics, vol. 31, issues 4–5, pp. 258-266 (2007).

[7] Mhalla, A., Chateau, T., Gazzah, S. & Amara, N. E. B.: An Embedded Computer-Vision System for Multi-Object Detection in Traffic Surveillance. IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 11, pp. 4006-4018 (11 2019).

[8] Behrens, S., Laue, H., Althaus, M., Boehler T., Kuemmerlen, B., Hahn, H. K., Peitgen, H.: Computer assistance for MR based diagnosis of breast cancer: Present and future

challenges - Computerized Medical Imaging and Graphics. vol 31, issues 4–5, pp. 236-247 (2007)

[9] Sivic, J., Zitnick C. L. & Szeliski R.: Finding People in Repeated Shots of the Same Scene. BMVC (2006).

[10] Yang, M.H., Kriegman, D. Ahuja, N. Detecting Faces in Images: A Survey. Pattern Analysis and Machine Intelligence. IEEE Transactions, vol. 24. pp. 34 - 58 (2002).

[11] Davies, E.: Computer Vision: Principles, Algorithms, Applications, Learning (2017).

[12] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G.: Recent Advances in Convolutional Neural Networks. Pattern Recognition (2015).

[13] Selvikvåg, A., Lundervold, A.: An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik. vol. 29, issue 2, pp. 102-127 (2019).

[14] Shen, D., Wu, G., Suk, HI.: Deep Learning in Medical Image Analysis. Annual Review of Biomedical Engineering (06 2017).

[15] Goodfellow, I., Bengio, Y. & Courvill, A.: Deep Learning. MIT Press. `http://www.deeplearningbook.org` (2016).

[16] Ker, J., Wang, L., Rao, J. & Lim, T.: Deep Learning Applications in Medical Image Analysis. IEEE Access, vol. 6, pp. 9375-9389 (2018).

[17] Maier, A., Syben, C., Lasser, T. Riess, C.: A gentle introduction to deep learning in medical image processing. Zeitschrift für Medizinische Physik (2019).

[18] Nielsen, M.: Neural Networks and Deep Learning, Determination Press. `http://neuralnetworksanddeeplearning.com/` (2015).

[19] Litjens G., et al.: A survey on deep learning in medical image analysis [Online]. `https://arxiv.org/abs/1702.05747` (2017).

[20] Cooper L., Carter A., Farris A., et al.: Digital Pathology: Data-Intensive Frontier in Medical Imaging: Health-information sharing. Proc IEEE Inst Electr Electron Eng. vol. 100(4), pp. 991-1003 (2012)

[21] Gurcan, M.N., Boucheron, L.E., Can A., et al.: Histopathological image analysis: a review. IEEE Rev Biomed Eng. vol. 2, pp. 147-171 (2009).

[22] Aeffner, F., Adissu, H.A., Boyle, M.C., et al.: Digital Microscopy, Image Analysis, and Virtual Slide Repository. ILAR J. vol. 59(1), pp. 66-79 (2018).

[23] Irshad, H.,Veillard, A., Roux, L., Racoceanu, D.: Methods for Nuclei Detection, Segmentation and Classification in Digital Histopathology: A Review Current Status and Future Potential. IEEE reviews in biomedical engineering, vol. 7, pp. 97-114 (2014).

[24] Jain A.K. & Lal, S.: Feature Extraction of Normalized Colorectal Cancer Histopathology Images. Ambient Communications and Computer Systems. Springer Singapore (2019).

[25] Alpaydin, E.: Introduction to Machine Learning, 2nd ed, The MIT Press (2010).

[26] Kouro, K., Exarchos, T., Exarchos, K., Karamouzis & M., Fotiadis, D.: Machine learning applications in cancer prognosis and prediction, Computational and Structural Biotechnology Journal, vol. 13, pp. 8-17 (2015).

[27] Gupta, M. & Gupta, B.: A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques. 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), pp. 997-1002 (2018).

[28] Okun O., Priisalu H.: Random Forest for Gene Expression Based Cancer Classification: Overlooked Issues. Pattern Recognition and Image Analysis. Lecture Notes in Computer Science, Springer, vol. 4478. Berlin, Heidelberg (2007).

[29] Kharya, S., Soni, S.: Weighted naive bayes classifier: A predictive model for breast cancer detection. International Journal of Computer Applications, vol. 133, no 9, pp. 32-37 (2016).

[30] Ferri, C., Hernández-Orallo, J. & Modroiu, R.: An experimental comparison of performance measures for classification. Pattern Recognition Letters, vol. 30, issue 1, pp. 27-38 (2009).

[31] Hosny, K.M., Kassem, M.A. & Foaud, M.M.: Skin melanoma classification using ROI

and data augmentation with deep convolutional neural networks. Multimed Tools Appl 79, pp. 24029–24055 (2020).

[32] Demir, C. & Yener, B.: Automated cancer diagnosis based on histopathological images : a systematic survey (2005).

[33] He, L., Long, L., Antani, S. Thoma, G.: Histology image analysis for carcinoma detection and grading. Computer methods and programs in biomedicine. vol. 107. pp. 538-56. (2012).

[34] Gil, J., Wu, H. & Wang, B.Y.: Image analysis and morphometry in the diagnosis of breast cancer, Microsc. Res. Techniq. vol. 59, pp. 109-118 (2002).

[35] Xu, Y., Zhu, J.Y., Chang, E., et al.: Weakly supervised histopathology cancer image segmentation and classification. Medical Image Analysis. vol. 18, issue 3. pp. 591-604 (2014).

[36] Kather, J. N., Weis, C. A., Bianconi, F., Melchers, et al.: Multi-class texture analysis in colorectal cancer histology. Scientific reports, vol. 6, pp. 27988 (2016).

[37] Kather, J.N., Marx, A., Reyes-Aldasoro, C.C., et al.: Continuous representation of tumor microvessel density and detection of angiogenic hotspots in histological whole-slide images. Oncotarget (2015).

[38] Cerda, P., Varoquaux, G. & Kégl, B.: Similarity encoding for learning with dirty categorical variables. Machine Learning, vol. 107, issue 8-10, pp.1477-1494 (2018).

[39] Dai, J.J., Lieu, L., Rocke, D.: Dimension reduction for classification with gene expression microarray data. Statistical Applications in Genetics and Molecular Biology, vol. 5(1), pp. 1–15 (2006).

[40] Truntzer, C., Mercier, C., Gautier, C., et al.: Importance of data structure in comparing two dimension reduction methods for classification of microarray gene expression data. BMC Bioinformatics, vol. 8(90) (2007).

[41] Miller, K.D., Nogueira, L., Mariotto, A.B., et al.: Cancer treatment and survivorship statistics.2019 CA A Cancer J Clin. vol. 69, pp. 363-385 (2019).

[42] Bovik, A.C..: Chapter 4 - Basic Binary Image Processing, The Essential Guide to Image Processing. Academic Press, pp. 69-96 (2009).

[43] Gevers, T., Gijsenij, A., Bagdanov, A.D., et al.: Color Feature Detection. In Color in Computer Vision. (2012).

[44] Sirinukunwattana, K., Snead, D.R.J. & Rajpoot, N.M.: A Stochastic Polygons Model for Glandular Structures in Colon Histology Images. IEEE Transactions on Medical Imaging (2015).

[45] Kainz, P., Pfeiffer, M. & Urschler, M.: Colon Gland Segmentation with Deep Convolutional Neural Networks and Total Variation Segmentation. Institute of Computer Graphics and Vision. MICCAI 2015 Challenge GLaS: Gland Segmentation in Colon Histology Images. Munich, Germany (2015).

[46] Akbar, B., Gopi, V. P., & Babu, V. S.: Colon cancer detection based on structural and statistical pattern recognition". 2nd International Conference on Electronics and Communication Systems (ICECS) (2015).

[47] Altunbay, D., Cigir, C., Sokmensuer, C. & Gunduz-Demir, C.: Color Graphs for Automated Cancer Diagnosis and Grading. IEEE Transactions on Biomedical Engineering, vol. 57(3), pp. 665-674 (2010).

[48] Arévalo, J., Cruz-Roa, A. & González, F.: Histopathology Image Representation for Automatic Analysis: A State of the Art Review. Revista Med, vol. 22 (2) (2014).

[49] Dos Santos, L., Alves, L., Botazzo, G., Gonçalves, M., do Nascimento, M. & Azevedo, T.: Multidimensional and Fuzzy Sample Entropy (SampEnMF) for Quantifying H&E Histological Images of Colorectal Cancer. Computers in Biology and Medicine, vol. 103, pp. 148-160 (2018).

[50] Dundar, M., Badve, S., Bilgin, G., Raykar, V., Jain, R., Sertel, O. & Gurcan, M.: Computerized Classification of Intraductal Breast Lesions Using Histopathological Images. IEEE Transactions on Biomedical Engineering, vol. 58(7), pp. 1977-1984 (2011).

[51] Hadid, A., Pietikainen, M. & Ahonen, T.: A Discriminative Feature Space for Detecting and Recognizing Faces. IEEE Conference on Computer Vision and Pattern Recognition, pp. 797-804 (2004).

[52] Iftikhar, M., Hassan, M. & Alquhayz, H.: A Colon Cancer Grade Prediction Model using Texture and Statistical Features, SMOTE and mRMR. 19th International Multi-Topic Conference (2016).

[53] Jørgensen, A., Rasmussen, A., Mäkinen, A., Andersen, S., Emborg, J., Røge, R. & Østergaard, L.: Using Cell Nuclei Features to Detect Colon Cancer Tissue in Hematoxylin and Eosin Stained Slides. Cytometry Part A, vol. 91, pp. 785-793 (2017).

[54] Kather, J., Weis, C., Bianconi, F., Melchers, S., Schad, L., Gaiser, T., Marx, A. & Zöllner, F.: Multi-class Texture Analysis in Colorectal Cancer Histology. Nature Scientific Reports, vol. 6, pp. 27988 (2016).

[55] Kothari, S., Phan, J., Moffitt, R., et al.: Automatic Batch-Invariant Color Segmentation of Histological Cancer Images. IEEE International Symposium on Biomedical Imaging: from Nano to Macro, pp. 657-660 (2011).

[56] Linder, N., Konsti, J., Turkki, R., et al.: Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. Diagnostic Pathology, vol. 7(22), pp. 1-11 (2012).

[57] Masood, K., & Rajpoot, N.: Texture based classification of hyperspectral colon biopsy samples using CLBP. IEEE International Symposium on Biomedical Imaging: From Nano to Macro (2009).

[58] Malik, J.:Colorectal Cancer Diagnosis from Histology Images: A Comparative Study. ArXiv, Computer Vision and Patter Recognition. `https://arxiv.org/abs/1903.11210` (2019).

[59] Wang, C., Shi, J., Zhang, Q., & Ying, S.: Histopathological image classification with bilinear convolutional neural networks. 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4050–4053 (2017).

[60] Kather, J. N., Krisam, J., Charoentong, P., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS medicine, vol. 16(1), (2019).

[61] Abdelsamea, M., Pitiot, A., Grineviciute, R. B., et al.: A cascade-learning approach for automated segmentation of tumour epithelium in colorectal cancer. Expert Systems with Applications, vol. 118, pp. 539-552 (2019).

[62] Ozdemir, E., Sokmensuer, C. & Gunduz-Demir, C. A resampling-based Markovian model for automated colon cancer diagnosis. IEEE transactions on biomedical engineering, vol. 59(1), pp. 281-289 (2012).

[63] Xu, J., Cai, C., Zhou, Y., Yao, B., et al.: Multi-tissue Partitioning for Whole Slide Images of Colorectal Cancer Histopathology Images with Deeptissue Net. European Congress on Digital Pathology, pp. 100-108 (2019).

[64] Paulin, F. & Santhakumaran, A.: Classification of breast cancer by comparing back propagation training algorithms. International Journal on Computer Science and Engineering, vol. 3(1), pp. 327-332 (2011).

[65] Eckhouse, S., Lewison, G. & Sullivan, R.: Trends in the global funding and activity of cancer research. US National Library of Medicine, National Institutes of Health (2008).

[66] Wild, C.: World cancer report 2014, World Health Organization, (Eds. C. P. Wild, B. W. Stewart), pp. 482-494. Geneva, Switzerland (2014).

Vita Auctoris

NAME:	Sameer Akhtar Syed
PLACE OF BIRTH:	Telangana, India
YEAR OF BIRTH:	1996
EDUCATION:	Bachelor of Technology, 2014-2018. GITAM University, Hyderabad, Telangana, India. Master of Science in Computer Science, 2019-2020. University of Windsor, Windsor, ON.