FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

AI-based methods for cancer cells quantification using whole slide imaging

Sara Carvalho Alves

Mestrado em Engenharia Eletrotécnica e de Computadores

Supervisors: Tânia Pereira, Hélder Oliveira and Fernando Schmitt

November 16, 2022

Resumo

O principal procedimento para o diagnóstico do cancro é a análise das lâminas de tecido histopatológico por um patologista, incluindo a estimativa do conteúdo de células tumorais. Este procedimento, para além de demorado, apresenta uma elevada variabilidade entre os patologistas. Com o aparecimento e a comercialização de *scanners* de lâminas digitais, foi possível obter a digitalização das lâminas inteiras. Consequentemente, a utilização de métodos de análise de imagem permitem obter resultados robustos, num curto período de tempo, na estimação da percentagem de células tumorais. Combinando a vantagem destes métodos computarizados com a experiência dos patologistas, pode-se ainda obter resultados mais precisos.

Neste contexto, e com o intuito de ajudar os patologistas, esta dissertação visa desenvolver uma metodologia de aprendizagem computacional para prever a percentagem de células tumorais presentes em imagens de cancro da mama obtidas a partir da digitalização das lâminas. Para tal, foi investigada a influência de três factores diferentes na previsão da percentagem de células tumorais realizada por uma ResNet-18. Um destes factores diz respeito à rede (camada classificadora) e os outros dois ao conjunto de dados (remoção das imagens com celularidade zero dos dados de treino e normalização da cor). Destes, verificou-se que os dois factores associados ao conjunto de dados são mais críticos para o desempenho do modelo. Em particular, os melhores resultados foram obtidos quando as imagens com zero celularidade foram removidas do conjunto de dados de treino e não foram utilizadas técnicas de normalização de cor.

Adicionalmente, a utilização do pré-treino foi também investigada, visto que a utilização de conjuntos de dados com um carácter geral para o pré-treino pode melhorar a eficácia e o desempenho do modelo. Para esse efeito, foram consideradas três abordagens: (i) ResNet-18 pré-treinada com ImageNet (conjunto de dados gerais), (ii) ResNet-18 pré-treinada com PCam (dados específicos) e (iii) ResNet-18 treinada a partir do zero. Os resultados obtidos demonstram que a rede pré-treinada com a ImageNet tem um desempenho superior às outras duas abordagens, sendo a pré-treinada no conjunto de dados PCam a pior.

Finalmente, foi também realizado um estudo preliminar relativo a uma técnica de refinamento. Esta consiste em descongelar progressivamente os pesos das redes que foram pré-treinadas com os conjuntos de dados da ImageNet e da PCam. Os resultados obtidos são promissores e demonstram uma melhoria do modelo pré-treinado com a PCam. Relativamente à rede pré-treinada com a ImageNet, não foi observada uma influência significativa, no entanto, é necessária uma investigação mais aprofundada.

Abstract

The main procedure for cancer diagnosis is the analysis of histopathology tissue slides by a pathologist, comprising the estimation of the tumor cell content. This line of action is time-consuming and presents high interobserver variability. With the appearance and commercialization of digital slide scanners, it was possible to obtain the digitalization of entire histology slides. Digital image analysis methods may be able to produce robust and reproducible results in the estimation. Moreover, by combining the strengths and experience of pathologists with the advantages of computerized methods, it may be possible to obtain more accurate results.

In this context, this dissertation aims to develop a machine learning methodology to predict the tumor cells percentage in Whole-Slide Imaging data of breast cancer in order to assist pathologists. The influence of three different factors on the prediction of the tumor cells done by a ResNet-18 was investigated. One of these factors is directly related to the network (classifier layer), and the other two are associated to the dataset (removal of the images with zero cellularity from the training data and color normalization). Of these, it was found that the two factors associated with the dataset are more critical to the performance of the model. In particular, the best results were obtained when the images with zero cellularity were removed from the training dataset and no color normalization techniques were used.

Furthermore, the use of pre-training was also investigated, as it has been questioned if pre-training with a very general dataset is an effective way of improving the performance of the networks. To that end, three approaches were considered: (i) ResNet-18 pre-trained with ImageNet (general dataset), (ii) ResNet-18 pre-trained with PCam (context specific) and (iii) ResNet-18 trained from Scratch. The results obtained show that the network pre-trained with the ImageNet outperforms the other two approaches, with the one pre-trained on the PCam dataset being the worst one.

Finally, a preliminary study concerning a fine-tuning technique was also conducted. This consisted on progressively unfreezing the weights of the networks that were pre-trained with the ImageNet and with the PCam datasets. The results were promising, demonstrating an improvement of the model pre-trained with the PCam. No significant influence was observed for the network pre-trained with ImageNet, nevertheless, further investigation is required.

Acknowledgments

I would like to express my gratitude to my supervisor Tânia Pereira for always helping and guiding me towards the right direction. Also, I would like to thank Professor Hélder Oliveira and Professor Fernando Schmitt for the opportunity of developing this dissertation and inspiring me to work on such a relevant area. A special thanks to Francisco Silva, who was always available to reply to my infinite questions.

Moreover, to my family, Mãe, Pai, Kika, Aníbal, Zé, Lu, Tuco e Riga a very special thank you for listening to my emotional crisis and supporting me throughout this time. To my friends, both the FEUP ones and the Zinde ones, thank you for being a part of this journey. Last but not least, I would like to thank my boyfriend Gonçalo, who was always my rock and capable of putting me in a good mood.

Sara Alves

"Invisible things are the only realities."

Edgar Allan Poe

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
BreastPathQ	SPIE-AAPM-NCI BreastPathQ
CI	Confidence Interval
CNN	Convolutional Neural Network
DC	Dice Coefficient
DL	Deep Learning
DNN	Deep Neural Network
FCN	Fully Convolutional Network
FFPE	Formalin-Fixed Paraffin-Embedded
GA	Genetic Algorithm
GBDT	Gradient Boosting Decision Tree
H&E	Hematoxylin & Eosin
ICC	Intraclass Correlation Coefficient
LR	Learning Rate
LTU	Linear Threshold Unit
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
mRMR	Minimum Redundancy Maximum Relevance
MSE	Mean Squared Error
NAT	Neoadjuvant Therapy
PCA	Principal Component Analysis
PCam	PatchCamelyon
P_k	Prediction Probability
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
RS	Rough Set
SVM	Support Vector Machine
SVR	Support Vector Regression
WHO	World Health Organization
WSI	Whole-Slide Imaging
τ_b	Kendall's Tau-b

Chapter 1

Introduction

1.1 Context and Motivation

At the moment, cancer corresponds to the main cause of death worldwide [1]. As stated by the World Health Organization (WHO) in 2019 (Figure 1.1), cancer was the main cause of death in 57 countries and the second in 55 countries, making a total of 112 countries out of 183 that have to deal with the mortality of this disease on a daily basis. It is important to notice that the numbers provided do not possess information on the impact of the virus SARS-CoV-2 responsible for the coronavirus disease.

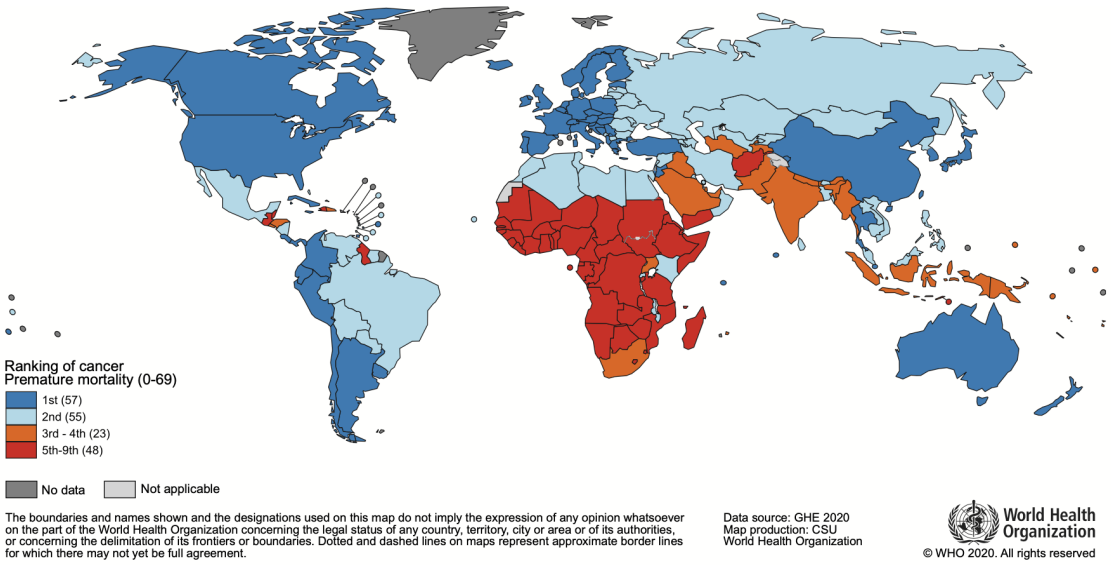


Figure 1.1: Cancer as a cause of death for people with less than 70 years in 2019. Source: World Health Organization (as it appears in [1]).

With the emergence of several new personalized and targeted cancer therapies, there has been an increase in the development of new molecular tests, making molecular pathology essential for defining mutation status [2]. Molecular tests sensitivity depends largely on the tumor cell

percentage present in a sample compared to the percentage of normal cells. Since the presence of non-neoplastic cells may dilute the percentage of tumor DNA and lead to false negative results, a previous step of estimating if the neoplastic cell content is enough to meet the threshold criteria of the test is required [3]. Estimation of the tumor cell percentage is usually made in hematoxylin & eosin (H&E) stained Formalin-Fixed Paraffin-Embedded (FFPE) tissue slides by a pathologist. Macrodissection may be used to enrich the tumor cell content in the sample for posterior molecular profiling, by eliminating non-neoplastic cells [4].

Studies have shown that estimations of the tumor cell content made by pathologists are subjective, may not be accurate and present high interobserver variability [3]. Pathologists are also prone to overestimating the neoplastic cell percentage, which in the case of a molecular test, may result in a false negative result. This may have severe consequences in the treatment of the patient. Moreover, pathologists are not oblivious to visual and cognitive traps as sources of bias [5]. In an attempt to minimize the variation of the estimation between pathologists, studies were made to reach consensus-based recommendations that could lead to more accurate results [2][6].

In the beginning of this century, digital slide scanners started to become widely available [7]. Whole-slide Imaging (WSI) is the most recent and important imaging technique in pathology nowadays. It allowed the digitization of the entire H&E slides, maintaining a high resolution. WSI made collaborations between pathologists in different corners of the world possible, without having to transport slides between places [8]. Furthermore, WSI facilitated digital image analysis and the implementation of computer-based methods that would be capable to detect important features.

The development of different image analysis techniques may become an unbiased option that would help to obtain robust and reproducible results and a viable tool to aid pathologists in estimating the tumor cell content [5].

1.2 Problem Statement and Objectives

Nowadays, the estimation of the tumor cell content in a histopathology tissue slide is made by a pathologist. Several studies have already demonstrated that this procedure is time-consuming and presents high interobserver variability [3]. On the other hand, it has been recently shown that computerized methods are advantageous for aiding pathologists in the clinical diagnosis and in further treatment planning [9]. Moreover, these methods also allow experts to save time for tasks where their knowledge is imperative. In addition, these methods have shown good results in estimating the neoplastic content in the presence of different types of cancer, such as in lung cancer [4], breast cancer [10] and colorectal cancer [11]. By combining the strengths of digital image analysis with the strengths of the human pathologist, it is possible to overcome the weaknesses of each approach, therefore obtaining better results (Figure 1.2).

The main objective for this work is to contribute to the development of an automated approach based on artificial intelligence (AI) methods to quantify breast cancer that will assist pathologists in performing an accurate diagnosis and treatment planning. In order to achieve this goal, it

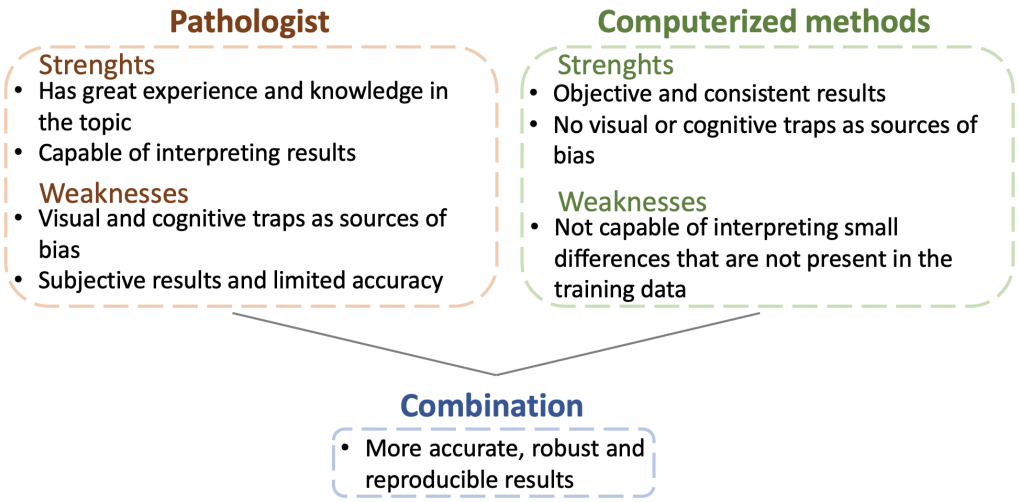


Figure 1.2: Combination of the capacities of pathologists and computerized methods (adapted from [5]).

is necessary to optimize a number of parameters that influence the performance of the model and, therefore, in this work, an investigation regarding the effect of different factors in the tumor quantification will be conducted.

1.3 Contributions

The following contributions were achieved in this work:

- an assessment of the impact of using color normalization techniques;
- a study on the effect of using transfer learning with a context specific dataset;
- a study on the impact of removing the images with zero cellularity from the training dataset;
- the development of an AI-based method that will be capable of accurately and efficiently estimating the tumor cell content in a tissue sample of breast cancer;
- a scientific publication with the main results of this work: Sara Alves, Francisco Silva, Fernando Schmitt, Tânia Pereira and Hélder P. Oliveira, AI-based Methods for Cancer Cells Quantification using Whole Slide Imaging. *npj Breast Cancer* (In Preparation).

1.4 Document Structure

The remaining part of this document is organized as follows. Chapter 2 presents an overview of theoretical aspects required for the understanding of the problem, namely histopathology and WSI, machine learning and deep learning methods. Then, chapter 3 contains an extensive literature review addressing automated methods for breast cancer and lung cancer. Chapter 4 presents the

description of the data that will be used throughout this work, some necessary data preprocessing steps and the methodology selected in this dissertation. Furthermore, Chapter 5 presents the results for the cancer cells quantification by using a network that was trained from scratch, one that was pre-trained on a general dataset, and one that was pre-trained with a context specific dataset. Finally, Chapter 6 contains the conclusions and future work.

Chapter 2

Background

This chapter presents some fundamental knowledge necessary for this work. It starts with information regarding the pathology field and the main steps required to obtain a tissue slide. Then, it addresses the technique of WSI and its main advantages and disadvantages. Furthermore, a few images that are representative of histopathology datasets are presented. In addition, the incidence and mortality of breast cancer are mentioned, as well as other relevant aspects of this disease. The chapter continues with a review of machine learning and some evaluation metrics. Finally, deep learning is also introduced, with emphasis given to Convolutional Neural Networks.

2.1 Histopathology

Pathology corresponds to a field of medical science that is dedicated to evaluating the causes of the disease of the tissues, cells and organs [12]. In addition, histology can be described as the study of the microscopic anatomy of cells and tissues. Taking these two definitions into account, it is possible to describe histopathology. Histopathology corresponds to the study of the tissues in order to localize and classify the disease [13]. Examples of histopathology images of different types of lung cancer is represented in Figure 2.1.

In terms of cancer, the analysis of the histopathology tissue slides under a microscope is considered to be the major procedure for diagnosis [5]. As already mentioned, the estimation of the tumor cells in the histopathology images by pathologists may not be accurate [3]. In Figure 2.2, annotations were made by two different pathologists, where it is possible to observe the variability between the experts. Furthermore, a more reproducible and robust approach may be of interest in order to reduce this variability.

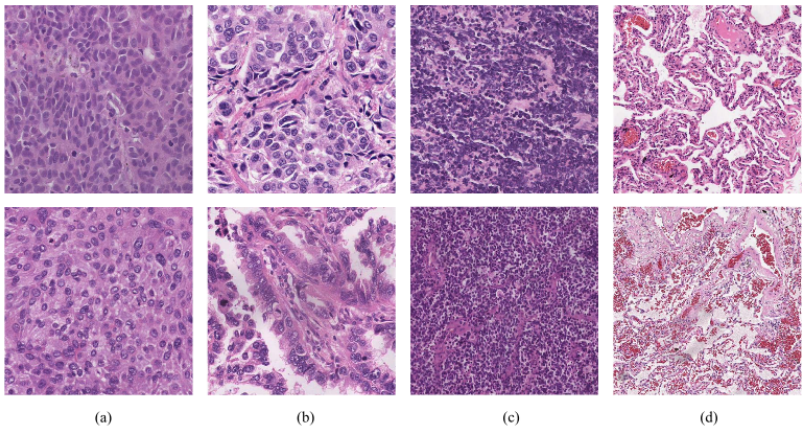


Figure 2.1: Examples of histopathology images of different types of lung cancer. (a) corresponds to squamous cell carcinoma, (b) to adenocarcinoma, (c) to small cell lung cancer and (d) to normal tissue (extracted from [14]).

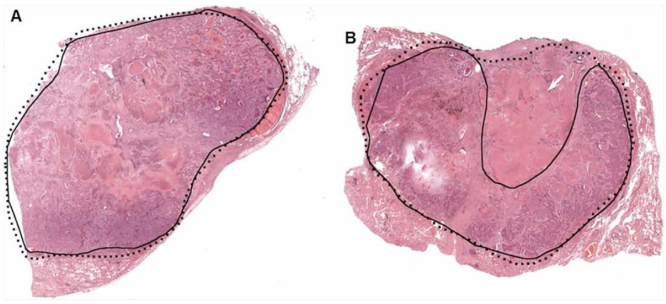


Figure 2.2: Annotations made by 2 different pathologists, where the dotted line corresponds to one of the pathologists and the solid line to the other. A shows good similarity between the annotations of both pathologists. B already shows a significant variance between the annotations made (extracted from [4]).

2.1.1 Tissue slides and WSI

In order to obtain the histopathology slide to be analyzed by the pathologist, it is necessary to prepare the tissue sample. The steps for the preparation of the slide are described in Figure 2.3. After obtaining a tissue sample, posterior processing steps, such as fixation and embedding, are of extreme importance, since it facilitates the cutting of sections with a small thickness (relevant for microscopy) [13]. The tissue slides are usually FFPE and stained with H&E [15]. Hematoxylin is responsible for staining the nuclear structures of the cells as dark blue or purple and eosin stains the cytoplasm of the cells as different shades of pink [13].

With the surfacing of whole-slide scanners [7] and these devices starting to become commercialy available all over the world, digital image analysis methods can be implemented in a more straightforward way. The main advantage of these scanners is that they provide all characteristics that normal microscopy already does, such as alternating between different magnifications and orientations. However, due to the high resolution of the images, the size of a WSI image is very

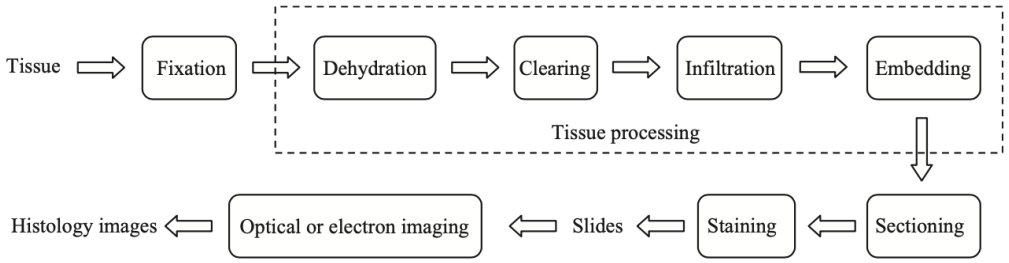


Figure 2.3: Preparation of the tissue slides (extracted from [13]).

large, which it can go from about 3.5 GB to 14.5 GB if the image is scanned at 20× magnification or at 40× magnification, respectively [8]. With this, the need for a great amount of storage space is necessary, which constitutes a disadvantage of this digitization technique. In fact, the cost of the storage space for the images can be much higher compared to the actual price of the scanner. In order to visualize the WSI data at different resolutions, the images are stored in a pyramidal architecture. In Figure 2.4, each plane in the pyramid represents the image at different resolutions, where the peak of the pyramid corresponds to the lowest resolution possible. With this strategy, it is possible to load the image with the intended resolution and even select a particular area without downloading the entire image.

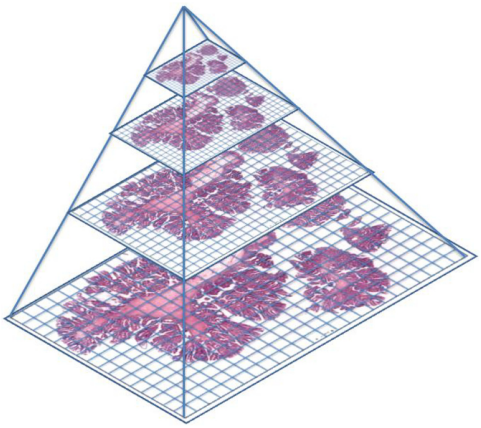


Figure 2.4: Pyramidal architecture for viewing the images in different resolutions, avoiding the transfer of the entire image into the computer. Each plane of the image is represented by a discrete number of tiles (extracted from [8]).

One of the main problems with storing a tissue slide for a long time is that the sample starts to lose the initial characteristics it presented. An example of this is represented in Figure 2.5, where it is possible to observe small bubbles in the sample and the colors of the stain in a darker tone [16]. This phenomenon can be avoided if the tissue slide was digitized after obtaining it. With this, even if the sample is stored and is no longer appropriate for diagnosis, there exists a digital format of the slide which can still be analyzed. Another advantage of using WSI is in the process of sharing the digitized slides with different pathologists. Sharing images between the experts is much easier

than having to transport slides among places.

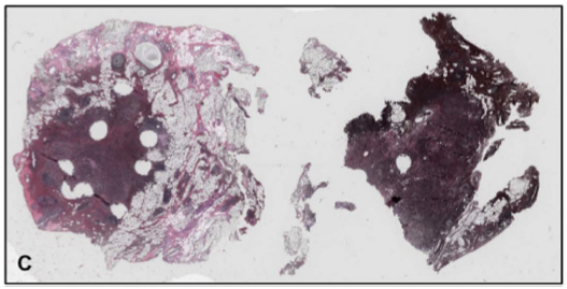


Figure 2.5: Bubbles in the tissue sample and darker stain color due to long storing time (extracted from [16]).

The datasets used by machine learning methods for cancer detection consist of digitized histopathology slides together with the annotations made by a pathologist. There are several datasets that were already used with this configuration, such as in breast cancer [16], lung cancer [17] and colorectal cancer [11]. Examples of such annotations are shown in Figure 2.6 and Figure 2.7 for the case of breast cancer and lung cancer, respectively. Another dataset configuration that may be used for quantifying cancer cells is having the WSI image only associated with the tumor cell percentage present in that region [9]. An example of this type of dataset regarding breast cancer is presented in Figure 2.8. It is easier to find histopathology datasets with this last configuration, since it only requires pathologists to supply the tumor percentage and not carefully annotating all cells in the image.

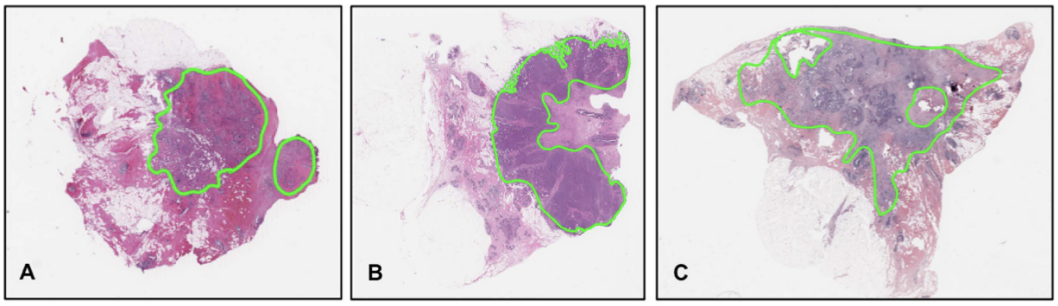


Figure 2.6: Examples of whole-slide images of breast cancer with annotations made by a pathologist (extracted from [16]).

One common analysis for the histopathology images is the classification of the tissue in benign, invasive carcinoma, *in situ* carcinoma and normal tissue (Figure 2.9). Invasive cancer happens when the tumor cells have proliferated and grown into another location beyond where it developed. Moreover, *in situ* cancer corresponds to tumor cells that did not spread from their original place, although it can still happen. Benign tumors have a slow growth rate and do not spread to other places than their original place. Another common approach is the division only in tumor or normal tissue.

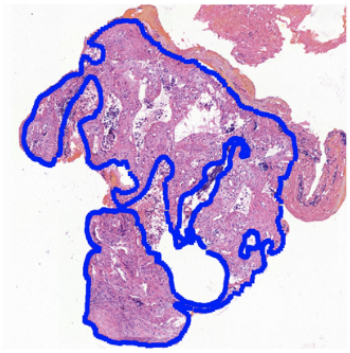


Figure 2.7: Example of a histopathology image of lung cancer with annotations made by a pathologist (extracted from [18]).

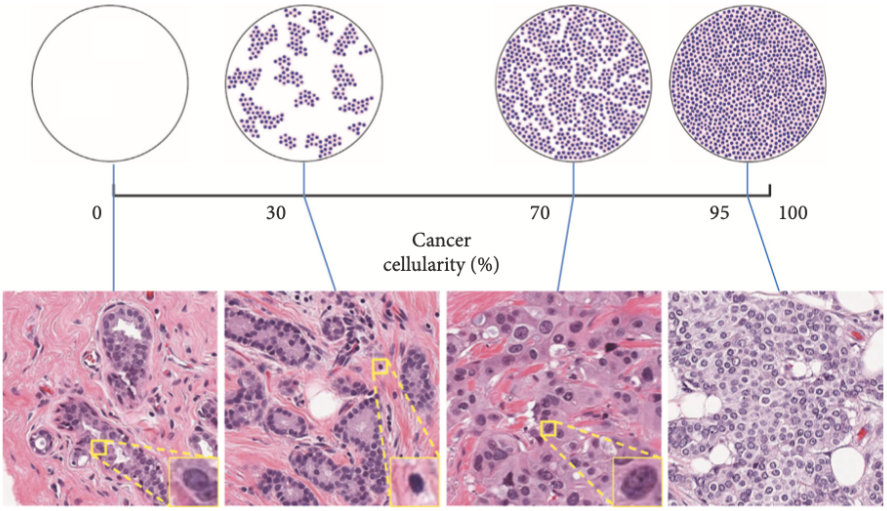


Figure 2.8: At the top, the circles represent diagrams that are used as a reference for manual estimation. At the bottom, the breast cancer WSI patches with their corresponding tumor percentage. Three yellow windows are also depicted, representing benign epithelial nuclei, lymphocyte, and malignant epithelial nuclei, respectively from left to right (extracted from [9]).

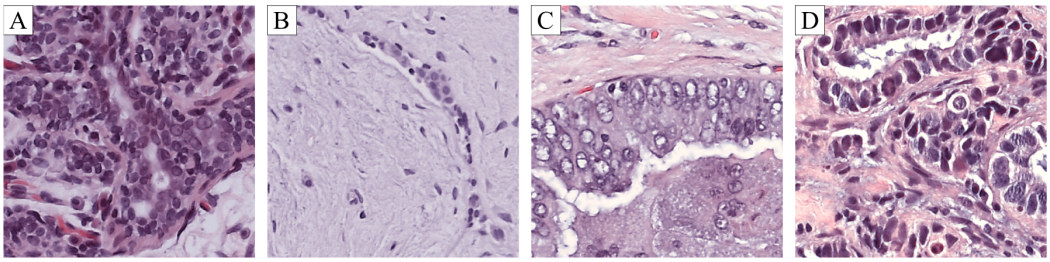


Figure 2.9: Examples of histopathology images. A corresponds to normal tissue, B to benign tumor, C to *in situ* carcinoma and D to invasive carcinoma (extracted from [10]).

2.2 Breast Cancer

Nowadays, cancer is one of the main causes of death worldwide [1]. In terms of cancer incidence in males, prostate cancer constitutes the most commonly diagnosed cancer followed by lung cancer. As for women, the most frequent diagnosis is breast cancer and cervical cancer. Regarding mortality, the results differ from cancer incidence. The cancer mortality in men is largely due to lung cancer and only then followed by prostate cancer, resulting in a switch on the two leading positions of cancer incidence. However, in women the main causes of cancer death are equivalent to the most frequently diagnosed types of cancer, with lung cancer only corresponding to the third cause. Taking into account both sexes, female breast cancer corresponds to the most commonly diagnosed type of cancer.

Regarding breast cancer, the two most frequently diagnosed types are invasive ductal carcinoma and invasive lobular carcinoma, where the former corresponds to around 75% of all diagnosed cases and the latter to 15% [19]. Examples of histopathology images of invasive ductal carcinoma are presented in Figure 2.1. The 5-year survival rate for breast cancer is around 85% [1], which is much higher than for other types of cancer, such as lung cancer. The survival rate of lung cancer patients 5 years after diagnosis ranges from 10% to 20%, however if detected early this number can increase considerably [20]. There are several factors that may contribute to developing breast cancer, such as age, hormone replacement and genetic factors, where the mutation of the BRCA1 and BRCA2 genes represent about 10% of all the breast cancer cases [21].

Furthermore, neoadjuvant (preoperative) systemic therapy (NAT) has been considered a valuable approach for almost all cancer patients that do not present any evidence for metastases. However, this therapy is mostly applied in cases where the cancer stage is considered to be locally advanced and in the presence of inoperable breast cancer, such as inflammatory breast cancer [22]. Effectively, several studies have shown that patients considered to be inoperable when submitted to therapy prior to surgery could actually become candidates for surgery [23]. Furthermore, patients who present a tumor size very large compared to the size of the breast are usually considered to be candidates for mastectomy. Nonetheless, neoadjuvant therapy made it possible that a significant part of these patients became candidates for breast conserving therapy. Following neoadjuvant therapy, pathologists are responsible for analysing the tissue and estimating the tumor cellularity in the tumor bed.

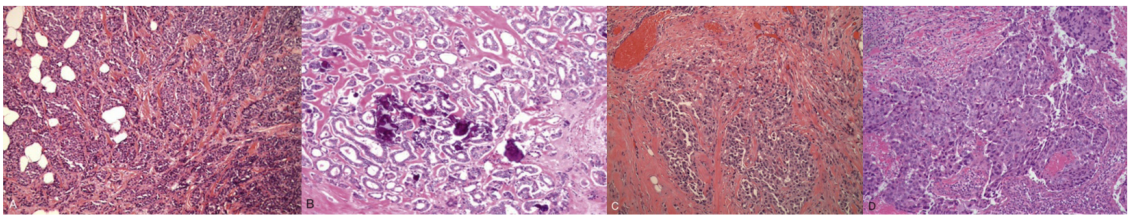


Figure 2.10: Four different examples of invasive ductal carcinoma. (extracted from [19]).

2.3 Machine learning

In 1956, there was a group of scientists who made the assumption that computers could be programmed to think, defining this principle as artificial intelligence [24]. The concept of machine learning was firstly defined in 1959 by Arthur Samuel. Machine learning (ML) is a branch of artificial intelligence that is capable of learning and gaining experience from training data [25].

Traditional programming differs from machine learning methods (Figure 2.11). Traditional programming consists in the processing of a previously implemented algorithm by a computer, when given a dataset, producing certain outputs. On the other hand, machine learning methods consist in the processing of a dataset and the corresponding outputs by a computer, obtaining an algorithm that delineates the correlation between the two inputs.

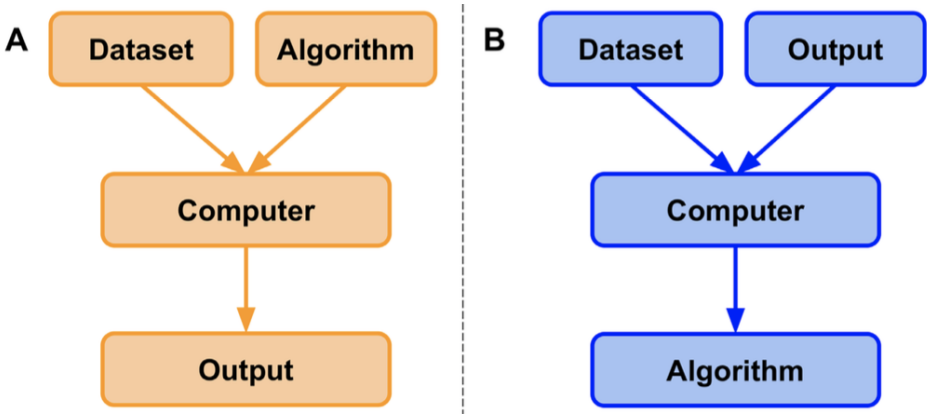


Figure 2.11: Traditional programming versus machine learning. (A) Traditional programming takes a dataset and an algorithm as an input, hands it to a computer and obtains the corresponding outputs. (B) Machine learning takes a dataset and the corresponding outputs as an input, hands it to a computer and obtains the algorithm that relates them (extracted from [24]).

There exist several different types of machine learning, but the most common are supervised learning, unsupervised learning and reinforcement learning [26].

In supervised learning, the training data comes in pairs, with an input and the correspondent expected output. This data is said to be labeled. If the prediction is based on labels that are discrete classes, it corresponds to a classification problem. However, if the label represents a continuous quantity, it corresponds to a regression problem. Examples of methods that are based on supervised learning are Support Vector Machines (SVMs) and Neural Networks [25].

Furthermore, in unsupervised learning, the data utilized for training does not present any labels. Some examples of algorithms that represent this type of learning are clustering methods and procedures for dimensionality reduction.

Finally, reinforcement learning tries to perform several actions, where each action grants a positive or negative reward, and find an optimal strategy that will concede the highest possible reward value.

There are some challenges that can pose as an obstacle for machine learning algorithms to perform suitably [25], such as:

- the amount of training data is not enough: most machine learning algorithms need great quantities of data in order to give accurate results.
- nonrepresentative training dataset: if the data used for training is not well representative of all cases we want the model to be able to generalize, the method will not be able to give good predictions.
- data with errors and non-relevant features: in the presence of data with errors or outliers (i.e. data that differs outstandingly from the rest) the model is prone to not give accurate results. Moreover, feature extraction and selection are important parts of machine learning algorithms and in the case of the presence of too many irrelevant features, the algorithm will not be able to perform well.
- overfitting: happens when the machine learning model is too complex and it is too adapted to the training dataset, giving bad results on a new test set. In addition, overfitting can be solved by increasing the size of the training dataset, making the model less complex or by reducing errors that are present in the data, such as outliers.
- underfitting: happens when the machine learning model is too simple and does not give good predictions in both training and test dataset. It can be solved by increasing the complexity of the model.

2.4 Deep Learning

Deep learning (DL), which in itself is a part of machine learning, consists of a group of methods that try to copy the functioning of the human brain by working with artificial neural networks (ANNs) [27].

Firstly, ANNs were introduced in 1943 by Warren McCulloch and Walter Pitts, where the model that tried to mimic the functioning of the biological neurons was based on artificial neurons that worked with propositional logic. In addition, the inputs and the output were always binary values [25].

Secondly, in 1957 Frank Rosenblatt introduced the perceptron, which consisted in an ANN with an artificial neuron that differed from the ones presented in 1943. These new artificial neurons were called Linear Threshold Unit (LTU) and consisted in a set of inputs that are no longer binary values but numbers associated with weights. Moreover, a term of bias is also added (Figure 2.12). Several transfer or activation functions can be used, such as the sigmoid function, the hyperbolic tangent and the step function.

By using several LTUs is possible to obtain a multi-layer perceptron (MLP). This ANN consists of an input layer, a hidden layer of LTUs and an output layer with a single LTU. An ANN with more than one hidden layer is defined as a deep neural network (DNN).

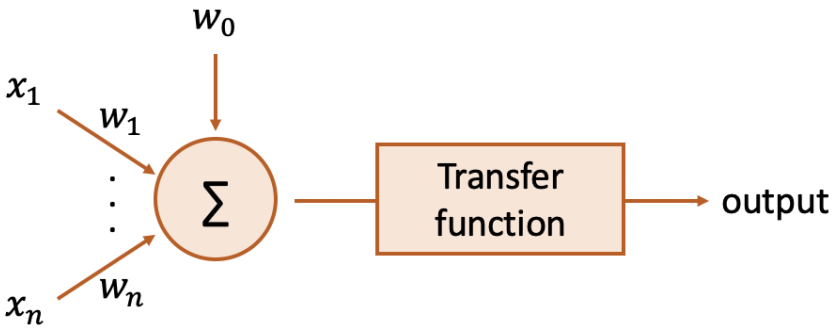


Figure 2.12: LTU artificial neuron (adapted from [25]).

Deep Neural Networks are trained using the backpropagation algorithm [25]. The main goal of this algorithm is to minimize the loss function by changing the values of the weights and the biases in the neural network. A loss function is described as a relation between the predicted output and the desired output. This function works as a good measure of the prediction capacity of the model. The backpropagation algorithm is based on a forward pass through each layer until reaching the output layer. Moreover, the difference between the obtained output and the desired output is calculated. It then follows to perform a reverse pass, to try to understand the contribution for the error of each neuron of the previous layer. In the final step, it tries to adjust the values of the weights in order to decrease the error.

The main advantage of deep learning methods is that feature extraction from the data is performed automatically, unlike other machine learning algorithms where to extract relevant features is necessary expertise and domain knowledge [28].

2.4.1 Convolutional Neural Networks

A Convolutional Neural Network (CNN) corresponds to a type of DNN where the input to the network is image data. The basic idea behind CNNs is that when the input is an image, pixels that are near each other usually share some common information [29]. With this in mind, it is possible to detect local features by using a sliding window through the image that works with the convolution operation [26].

There are three main types of layers in a CNN, more concretely:

- **Convolutional Layer** — In this layer, a filter or kernel slides through the image and performs the convolution operation. This process is done in order to automatically detect relevant features, like edges or gradients. In addition, to convolve the image and the filter, it is important to define stride and padding. Stride corresponds to the step for sliding the filter through the image. The bigger the stride, the smaller is the output of the convolution operation. Moreover, padding corresponds to the width of extra cells added to the image in order to obtain a bigger output matrix. Convolutional layers are usually followed by an

active layer, which is responsible for introducing nonlinearities to the network in order to be able to adapt to more problems [27].

- **Pooling Layer** — Pooling is responsible for reducing the dimensionality of the features, by also sliding a filter through the data. This layer usually follows the convolutional layer. The two most used methods for pooling are maximum pooling and average pooling. The former selects the pixel that presents the maximum value and the latter performs the average operation while the filter slides through the input. An example of maximum pooling is present in Figure 2.13.

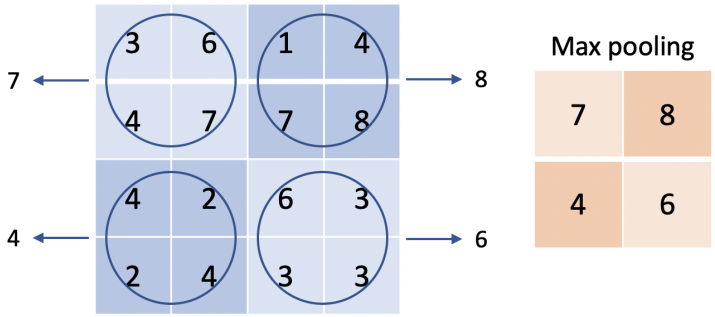


Figure 2.13: Maximum pooling using an input matrix with size 4x4, pooling filter of size 2x2 with a stride of 2 (adapted from [27]).

- **Fully Connected Layer** — This layer is responsible for the classification task based on the previously detected features. The fully connected layers can be added sequentially and usually use the softmax activation function [25] for multiclass classification problems.

CNNs have already been widely used in several different contexts, such as object detection [30], medical image segmentation [31] and face detection [32].

2.5 Evaluation Metrics

In order to evaluate if an algorithm provides good predictions on a given test set by comparing them with the expected results, it is necessary to define evaluation metrics. There are several indicators that are important for evaluating the performance of the developed model in a given test set, such as accuracy, sensitivity, specificity and precision [33]. These parameters can be calculated with the results that are usually represented in a confusion matrix (Figure 2.14).

Accuracy corresponds to the number of correct classifications by the model divided by all samples in the dataset and it can be defined as

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.1}$$

where TP (TN) corresponds to a true positive (negative) and FP (FN) to a false positive (negative).

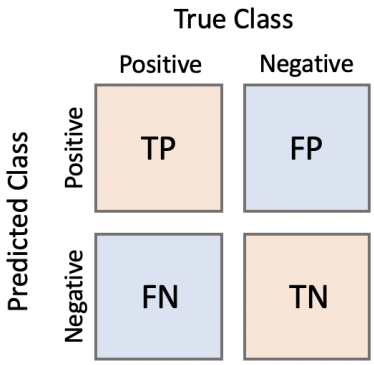


Figure 2.14: Standard Confusion Matrix (adapted from [33]). TP (TN) corresponds to a true positive (negative) and FP (FN) to a false positive (negative).

Sensitivity corresponds to the ratio between the results that were correctly classified as positive and the total of results that were actually positive. It can be written as

$$Sensitivity = \frac{TP}{TP+FN} \tag{2.2}$$

where the variables have the same meaning as in equation 2.1.

Specificity can be described as the ratio between the results that were correctly classified as negative and the total of samples in the dataset that were actually negative. It can be represented as

$$Specificity = \frac{TN}{FP+TN} \tag{2.3}$$

where the variables have the same meaning as in equation 2.1.

Finally, precision corresponds to the ratio between the results that were correctly classified as positive and all the results that were predicted to be positive. It can be written as

$$Precision = \frac{TP}{TP+FP} \tag{2.4}$$

where the variables have the same meaning as in equation 2.1.

Another important metric for evaluating the correlation between two regions is the Dice Coefficient (DC) [34]. This coefficient can be calculated as in equation 2.5, where the variables have the same meaning as in equation 2.1.

$$DC = \frac{2TP}{2TP+FP+FN} \tag{2.5}$$

In addition, the intraclass correlation coefficient (ICC) [35] is a measure of reliability and it can be used to evaluate the agreement between measurements. There are different forms to calculate this metric, but in this work we will consider the form of equation 2.6, where MS_R corresponds to the mean square for rows, MS_C for columns, MS_E to mean square error, k is the number of measurements and n is the number of testing samples. The more closer to 1 is this value, the more

reliable it is. This metric is extremely important in medicine since without a ICC that is close to 1, the results obtained cannot be trusted.

$$ICC = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)} \tag{2.6}$$

Kendall's tau-b (τ_b) metric is a rank correlation coefficient that measures the ordinal correlation between two obtained measurements. This coefficient is considered to be a nonparametric hypothesis test. This means that it aims to evaluate if the two measurements are correlated or independent between them and it is not made any kind of assumption on the probabilistic distribution of the two variables. τ_b is adjusted for ties and can be calculated as in equation 2.7, where x represents the cellularity that was estimated with a computerized method, y denotes the reference values for the estimation, n_c is the number of concordant pairs, n_d is the number of discordant pairs, n_{T_x} is the number of ties in x, n_{T_y} is the number of ties in y and $n_{T_{xy}}$ is the number of ties in both x and y. The values for τ_b can go from -1, which represents a negative association, to 1, which represents a perfect association between the measurements. If the value of this metric is 0, it means that there is no association or correlation between the measurements. The prediction probability (P_k) metric can be calculated as in equation 2.8, where the variables possess the same meaning as in equation 2.7 for τ_b. This metric also evaluates ordinal correlation.

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n - n_{Tx} - n_{T\,\mathrm{xy}})(n - n_{Ty} - n_{T\,\mathrm{xy}})}} \tag{2.7}$$

$$P_k = \frac{n_c + \frac{n_{Tx}}{2}}{n_c + n_d - n_{Tx}} \tag{2.8}$$

Moreover, another broadly used metric is the receiver operating characteristic (ROC), which consists of a plot of the sensitivity against (1- specificity). This graphic is used for binary classification and the closer the area under the curve (AUC) is to 1, the better the classifier is (Figure 2.15). When the ROC curve coincides with the diagonal dotted line, the binary classifier corresponds to a random classifier.

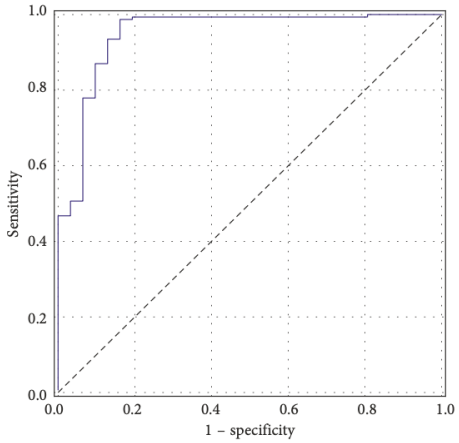


Figure 2.15: Example of a ROC curve (extracted from [9]).

2.6 Summary

The estimation of the tumor cell content between pathologists is not accurate enough and a more robust solution is required, which may rely on computational methods. Therefore, digitized images of the tissue slides should be obtained. In turn, the digitization of the slides also facilitates sharing of relevant information between experts and avoids the natural deterioration of the tissue samples. Nevertheless, since the digitized histopathology slides exhibit a very high resolution, they take up a considerable amount of storage space. In order to minimize this issue, the images are stored in a pyramidal architecture, which allows downloading only a specific part of the image.

Machine learning techniques have been used for classification and regression problems. Among these techniques, deep learning methods distinguish themselves for mimicking the functioning of the human brain and have the advantage of automatically extracting features, without the need for domain knowledge. One widely used DL method relies on the implementation of CNNs. These networks perform the convolution operation, which is particularly interesting since it provides spatial information about the input image.

Chapter 3

Literature Review

This chapter includes the literature review of some important methods for this dissertation. It starts with the review of the algorithms implemented in histopathology images of lung cancer. Moreover, some color normalization processes are compared. After that, methods that were proposed for the analysis of breast cancer WSI images are also mentioned. Even though this dissertation is focused on the analysis of histopathology images of breast cancer, the implemented methods by the different authors for lung cancer still remain relevant and are also addressed. Furthermore, some theoretical knowledge regarding the differences between different CNN architectures is presented. Finally, the results obtained by the different authors are reported.

3.1 Automated methods for Lung Cancer

In the last few years, several automated methods have been applied to the analysis of lung cancer in histopathology images. These methods are mainly based on CNNs. Some of the proposed approaches aimed at classifying the images into tumor or normal tissue, whereas others tried to classify the type of lung cancer.

Coudray *et al.* [36] proposed a deep learning algorithm to classify the WSI data into 3 different classes, more concretely normal tissue, adenocarcinoma and squamous cell carcinoma. The deep learning algorithm used consists of a Inception-v3 architecture [37], a deeper CNN where the computational cost is lower than other network models. Moreover, the authors also predicted the most frequent mutated genes in lung adenocarcinoma. The process of classification was performed with two different approaches. The first consisted in classifying the images into normal and tumor slides, along with classifying the tumor images into adenocarcinoma and squamous cell carcinoma. The second approach consisted in directly classifying the images into adenocarcinoma, squamous cell carcinoma and normal tissue. The network was also trained with transfer learning and fully training the model. Transfer learning [38] can be very useful because it applies knowledge gained from one problem and uses it to another related task. Since this method already uses knowledge that was previously acquired, it requires less data for training. In this case, the network was trained on a large-scale dataset known as ImageNet [39]. When training the network

with transfer learning, most of the weights keep their value and the fully connected layer is the one that is trained. By fully training the model, the network is completely trained by starting with random weights. The obtained results showed that fully training the network after separating the normal tissue images gave slightly better results than with transfer learning. Furthermore, the results also revealed that when separating the normal tissue from tumor in a first stage did not give better results. In fact, directly classifying the images into the 3 different classes was better, achieving an AUC of about 0.97.

Pham *et al.* [40] proposed a two-step deep learning method to detect lymph node metastases and classify them into three different classes, macrometastases, micrometastases and isolated tumor cells. The purpose of using a two-step approach was that lymphoid follicles were usually misclassified as tumor and excluding them in an initial phase could potentially lead to better results and eliminate false positives. Regarding this information, the first classifier differentiated lymphoid follicles from the rest of the tissue and the second step was responsible for detecting tumor cells. In Figure 3.1 is shown the comparison between a one-step and a two-step approach.

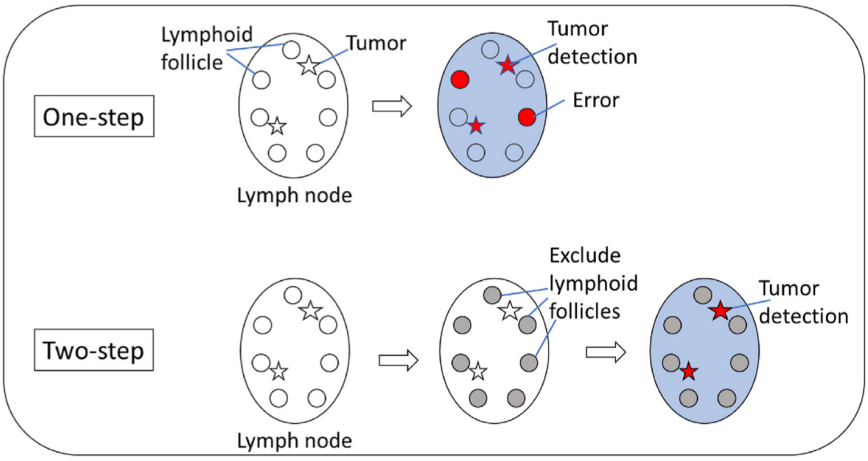


Figure 3.1: Comparing a one-step approach with a two-step. A two-step approach could potentially eliminate false positives (extracted from [40]).

Two methods were compared for lymphoid follicles detection, a random forest model and a CNN. The random forest method consists of an ensemble of decision trees classifiers and are usually trained with the bagging method, which means to train the algorithm with randomly selected subsets from the training data [25]. For the detection of lymphoid follicles, the CNN gave better results, achieving an accuracy of 94.5%, while the random forest model achieved an accuracy of 51.7%. The random forest method misclassified a lot of tumor cells as lymphoid follicles, resulting in a large number of false positives. For that reason, the CNN was chosen as the model to exclude lymphoid follicles. In the second step, a CNN with the VGG architecture [41] was used to detect tumor cells and the slides were labeled as containing metastases or not containing metastases. The algorithm performed extremely well for positive slides, identifying macrometastases, micrometastases and isolated tumor cells with 100% accuracy. As for negative slides, the two-step approach

gave poor results, since images still contained small points that were identified as isolated tumor cells or micrometastases, resulting in an overall sensitivity of 100% and specificity of 0%.

Qu *et al.* [42] proposed a deep learning method that performed both segmentation and classification in a unified framework (Figure 3.2). The prediction network is responsible for segmenting individual nuclei and classifying them into three classes, tumor, lymphocyte and stroma. It consisted of a U-Net architecture [43], widely used for segmentation in medical image, with a ResNet [44] as an encoder. The decoder part of U-Net is responsible for upsampling, which restores the feature maps to the size of the original image. ResNet introduces shortcut connections that skip one or more layers and their outputs are added to the outputs of the stacked layers (Figure 3.3). The perceptual loss part of the network compares the predicted label with the ground truth. Transfer learning was also used in order to compensate for the small dimension of the dataset.

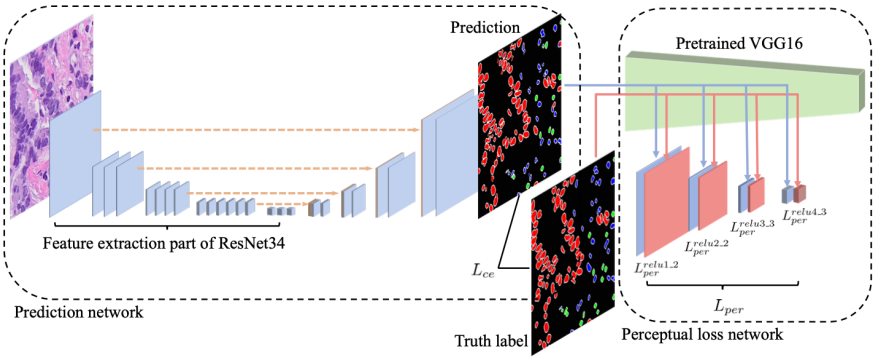


Figure 3.2: Proposed framework (extracted from [42]).

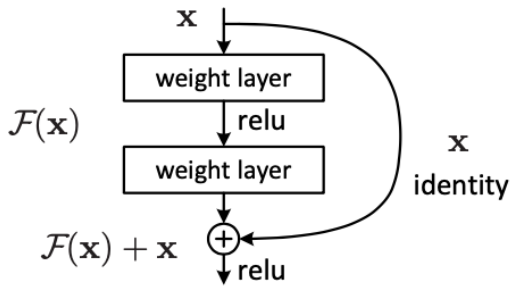


Figure 3.3: Residual Learning Block (extracted from [44]).

Furthermore, data preprocessing was also performed, consisting of color normalization and data augmentation. In the last 20 years, several color normalization methods of H&E staining in histopathology images have already been studied. Roy *et al.* [45] studied the proposed methods by different authors and evaluated their performances in terms of different metrics. There are three different types of color normalization methods, more concretely global color normalization [46], color normalization after stain separation by supervised [47] or unsupervised methods [48][49].

To test distinct methods, six of the algorithms presented in [45] were evaluated in a dataset of histopathology images of different cancers, namely liver, kidney, breast and colorectal. The results showed that a structure-preserving color normalization method [50] performed better than the other analyzed methods, maintaining the structure and brightness of the source image. It would be interesting to compare the performance of the method in [50] with others that were not evaluated in the study by Roy *et al* [45]. Additionally, data augmentation is extremely important when the training dataset is small and it is utilized to prevent the machine learning model that is being developed from overfitting the data. This process is usually done by enlarging the size of the dataset, by creating somewhat different copies of data that already exists and adding it to the dataset. The Dice coefficient for the proposed method was 0.876, which shows good concordance with the ground truth.

He *et al.* [51] proposed a deep learning model to automate the labelling of the tumor regions. It was used a DenseNet [52], which was responsible for classifying the image patches into malignant or benign. DenseNet architecture is based on dense blocks that connect all layers, which the main goal is to maximize the information that passes through the features from the different layers. In addition, it was pre-trained on the ImageNet dataset [39]. The annotations generated by the algorithm were reviewed by a pathologist who made adjustments to the annotation when it was not accurate. It was obtained a sensitivity of 87.9% and a specificity of 87.2% for image-level classification. Moreover, 38 images were selected and a comparison between the generated and reviewed annotations was made, obtaining a mean Dice coefficient for these images of 0.84.

Saric *et al.* [53] proposed a method that classifies the image patches as tumor or normal. A patch would be considered as tumor if 75% of the annotated pixels were tumor. Two different CNNs were implemented, a ResNet-50 and a VGG-16, and the results were compared. Both networks were pre-trained on ImageNet [39]. The results showed that VGG was slightly better than ResNet (Figure 3.4), even though ResNet was better on the ImageNet dataset. An important conclusion is that pre-training the networks on the ImageNet dataset in the context of histopathology images may not be a good practice, since the domains of both datasets differ greatly. A probable good solution to improve the results might be to pre-train the networks on another histopathology dataset, even from different types of cancer. The difference in the performance of the 2 networks might also be explained by the size of the training dataset. Since the quantity of the data is small and the ResNet architecture used was deeper than the VGG, the size of the training dataset probably needed to be increased in order to achieve better results. The classification accuracy at patch level for ResNet and VGG was 72.05% and 75.41%, respectively.

Wang *et al.* [14] proposed a weakly-supervised approach for classification in squamous cell carcinoma, adenocarcinoma, small cell lung cancer and normal tissue. The algorithm is divided into three different parts (Figure 3.5). The first step consists in obtaining a probability map through a patch-based CNN, more concretely a ScanNet architecture [54]. This architecture does not have an upsampling step, which is necessary for segmentation, since it is only necessary to perform classification. It is based on a VGG-16 but instead of having at the end three fully connected layers, it has three fully convolutional layers. The probability maps of each patch are stitched

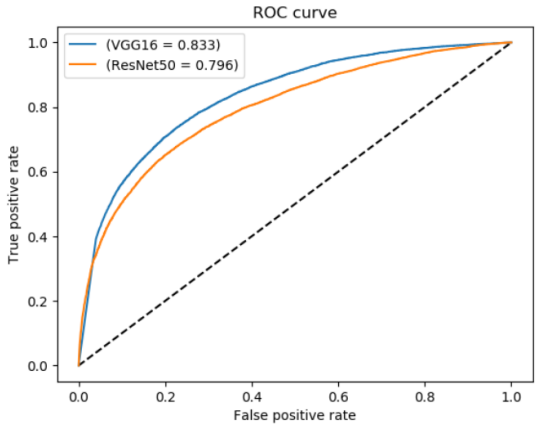


Figure 3.4: ROC curve for VGG-16 and ResNet-50 (extracted from [53]).

together in order to obtain the probability map for the whole image. In order to deal with coarse annotations, if the CNN misclassifies a region that was annotated by a pathologist, a higher penalty for that classification is considered. The second step takes into consideration spatial information, which is extremely relevant. This is the case since histopathology images present a very complex organization and heterogeneity. If spatial information would not be considered, the possibility of having patches with high probability but actually corresponding to an outlier is elevated. In order to deal with this information, a block is considered, which consists of a set of patches with some overlap. In this case, taking the average probability of a block into account, even if there is an outlier present in a patch with a high probability, the block will still be considered as normal tissue. Finally, in the third step, a specific class feature is obtained for each block and feature aggregation is performed. A random forest classifier uses the global feature descriptor to classify the images. Some preprocessing was performed, namely data augmentation and background removal by using the OTSU algorithm [55]. To evaluate the results, different feature selection and aggregation methods were used and the best result consisted of a 97.3% accuracy.

Li *et al.* [18] proposed the ACDC@LungHP 2019 challenge, which consisted of the first challenge of lung cancer detection and classification using histopathology images. The 10 methods for lung segmentation that performed better are reviewed and briefly explained. The implemented methods were mainly divided into two different groups, namely methods that only used one single model and methods that used multiple models. The results have shown that multi-model methods performed better than single-model methods with a mean Dice coefficient of 0.7966 and 0.7544, respectively. With respect to the single model methods, 5 different algorithms were described. The first algorithm used a CNN with a DenseNet architecture to classify images in tumor or normal tissue and a fully convolutional network (FCN) based on a DenseNet to segment the tumor area. The mean Dice coefficient for this method was 0.77. The other 4 algorithms based on single-model gave even inferior results.

However, the 3 methods that performed better in the challenge were all based on multi-model methods (Figure 3.6). The best performing team proposed a method that consisted of a U-Net

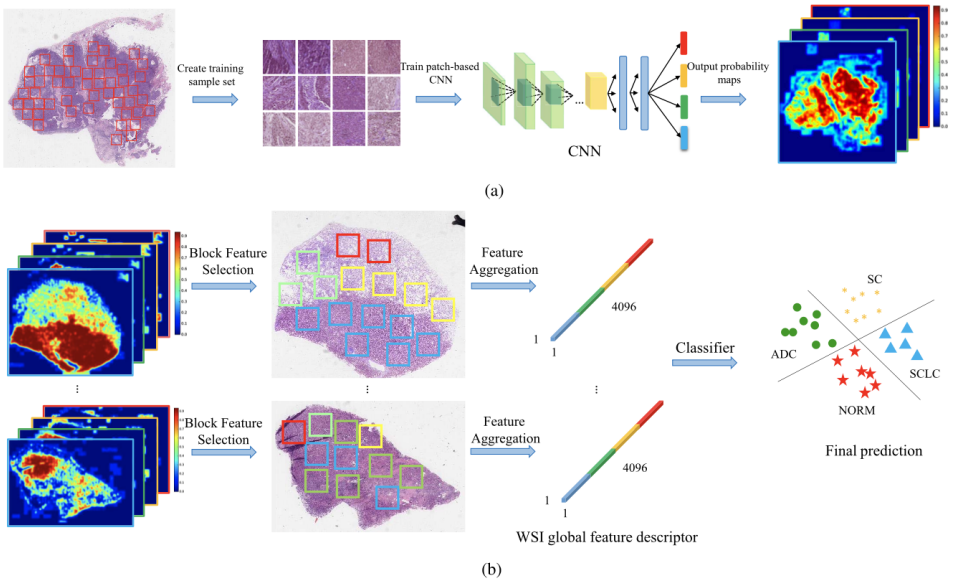


Figure 3.5: Proposed methodology. (a) Corresponds to the ScanNet architecture, which will be responsible for obtaining the probability map. (b) Spatial information is taken into account and features are aggregated. Finally, a random forest classifier is used (extracted from [14]).

combined with dense blocks and dilation blocks. The purpose of the dilation blocks was to obtain more information and acquire features at multiple scales. The best model was selected by trying different loss functions and then was ensembled. Color normalization and background removal by the OTSU algorithm were also performed as a preprocessing step. This algorithm achieved a mean Dice coefficient of 0.8372. The second best performing algorithm also implemented the OTSU algorithm to remove the background and data augmentation techniques for preprocessing. The method was based on a ResNet architecture with some nuances as an encoder. The network was pre-trained on ImageNet and the ensemble of models was also performed. The obtained mean Dice coefficient was very close to the best performing team, achieving a 0.8297 as a result. Finally, the third best performing team implemented a U-Net divided into two groups. The first group used three models where each one of them received the whole dataset but at different resolutions. The second group used three models as well but each one received a partition of the original dataset. The results were then fused with a Conditional Random Field [56] to improve the segmentation results and a mean Dice coefficient of 0.7968 was obtained. An interesting outcome was that for a particular image in the training dataset, all teams had a high performance. This is probably due to the staining process of the slide, which was well prepared, and that the cancer tissue was evident. Another conclusion was that the teams that pre-trained the models with ImageNet did not perform better than others. Interestingly, none of the teams used a dataset for pre-training the networks with similar characteristics to the histopathology images, which could potentially improve the results. In addition, algorithms using methods to remove the background of the images also presented a higher Dice coefficient.

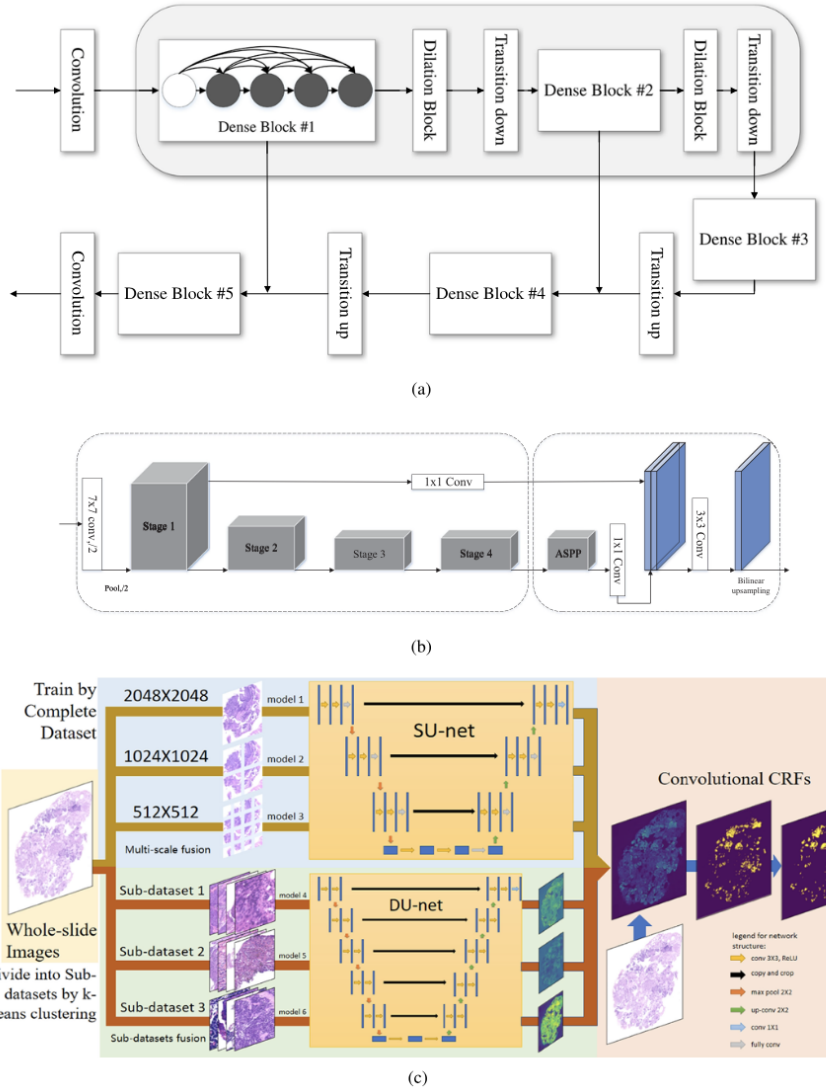


Figure 3.6: Proposed methods by the top 3 winning teams. (a) Rank 1 team. (b) Rank 2. (c) Rank 3 (extracted from [18]).

3.2 Automated methods for Breast Cancer

The analysis of breast cancer histopathology images is extremely relevant and several methods have already been proposed over the years. These approaches usually classify the images into healthy tissue or tumor, estimate the cancer cellularity or even segment the tumor area.

Chen *et al.* [33] proposed a method based on rough set (RS) theory [57] and a SVM for the classification of images into benign and malignant for breast cancer diagnosis. RS theory is mainly used for eliminating redundant features, which may lead to better results. In [57], the RS attribute reduction algorithm was used in a way to select a reduct set of features with the help of genetic algorithms (GA). To find the optimal feature subsets, the subsets of attributes that are maintained are the ones that included the most relevant attribute and the least relevant. Moreover,

a SVM is a method that consists of the representation of each data as a singular point in a space with n dimensions, where n is given by the number of features [58]. This method then tries to assign each point to one of two categories for classification (Figure 3.7). This is done by trying to find an hyperplane that maximizes the margin that separates the two classes. By maximizing the margin, it is possible to make the model less prone to overfitting and able to generalize better. Although SVMs were firstly created to only deal with classification problems, they can also be used for regression problems. In addition, this method can be used in both linear and nonlinear problems. SVMs have been applied in several fields, such as for pattern recognition [59], object detection [60] and handwritten digits recognition [61]. While this algorithm works very well for a small amount of training data, it performs poorly for large training datasets, since the training time complexity grows with the size of the dataset [25]. The training phase and the test step were performed taking into account three different divisions of the dataset. The results for the best performing subset gave an accuracy of 99.41%. Finally, another important outcome was that with the results obtained, that particular subset can be considered to be the most relevant for breast cancer classification, therefore containing the most informative features for diagnosis.

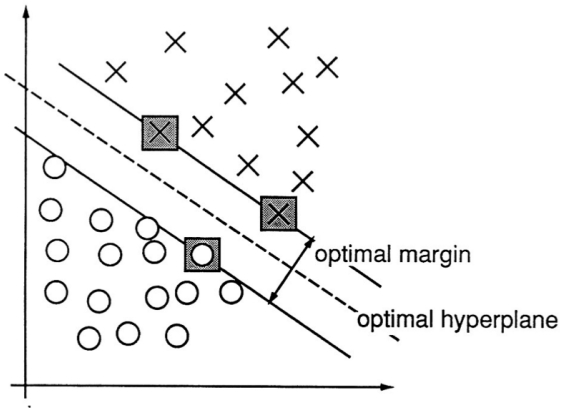


Figure 3.7: Graphical representation of an SVM classifier in a 2-dimensional space. The squares represent the support vectors that define the maximum margin between classes (extracted from [62]).

Spanhol *et al.* [63] proposed a method based on a CNN architecture for classifying the images into malignant or benign. The CNN was trained with patches of the images, in which they were selected randomly or by a sliding window approach. Different fusion rules were tested in order to obtain the final classification result. The best results were achieved by an AlexNet [64] architecture and for a magnification factor of 40x, with an Image Recognition Rate of 85.6%. The Image Recognition Rate is defined as the ratio between the number of correctly classified tumor images and the number of tumor images in the test set.

Araújo *et al.* [10] proposed a CNN based on the AlexNet architecture to classify the images into four different classes, normal tissue, benign, invasive or *in situ* carcinoma and posteriorly into two different classes, cancerous or normal tissue. For preprocessing the images, both color

normalization [65] (Figure 3.8) and data augmentation were performed. In order to obtain the classification of the image, two methods were compared, more concretely a patch-based CNN with a SVM, where the CNN works as a feature extractor and the SVM performs the classification, and an individual patch-based CNN that performs both feature extraction and classification. The algorithms were trained with three different fusion rules, maximum probability, majority voting and the sum of the probabilities. With maximum probability, the image will be classified in accordance to the patch with the maximum probability. In addition, with majority voting, the image class will be the most frequent class in the image patches. Finally, the sum of the probabilities corresponds to selecting the class with higher probability, where the sum of the probabilities of the same class is performed for each class. The obtained results for image classification with majority voting for 4 classes corresponded to an accuracy of 77.8% for both the CNN and the CNN with SVM. As for the classification with 2 classes, the accuracy with majority voting was 80.6% for the CNN and 83.3% for the CNN with SVM. Patch results were also evaluated and the accuracy values for the classification with 2 and 4 classes were lower than for the image classification. The authors mentioned that this result is probably due to considering the label of the patch image as the label of the whole image when no information about the location of the tumor was given.

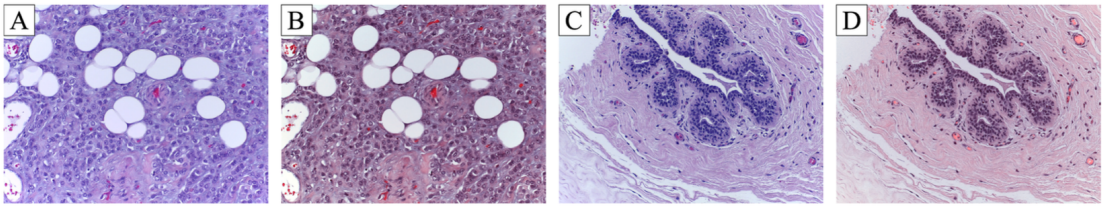


Figure 3.8: Color normalization of histopathology images. A and C correspond to the original images, while B and D are color normalized (extracted from [10]).

Peikari *et al.* [66] proposed a method for segmentation of nuclei from post-treated breast cancer tissue and classification in 4 classes, normal, low, medium or high residual cancer cellularity. The small dimension of the training dataset influenced the decision of the authors into using machine learning techniques and not deep learning methods. To annotate the images, a pathologist selected the regions of interest and classified the nuclei into 3 different groups, lymphocyte, benign or malignant. In order to perform the classification in the 3 classes, a cascaded classifier was used to train a SVM to separate in a first step the lymphocyte from benign and malignant and in a second step to separate the benign from the malignant. The cellularity was calculated as the ratio between the area of malignant and the total patch area. The results were compared with two pathologists, obtaining an ICC with a 95% confidence interval (CI) of 0.89 between the two pathologists, 0.74 between one of the pathologists and the automated method and 0.75 between the other pathologist and the automated method. The proposed method gave good results when it was analyzing medium cellularity images, with around 31% to 70% cellularity, and the accuracy decreased for low cellularity and normal patches.

Cruz-Roa *et al.* [16] proposed an algorithm based on a CNN for detecting the presence of

invasive tumor and the extent of it. The images were divided into patches and color normalization and data augmentation were performed. Patches would be considered as tumor if they presented at least an 80% overlap with the annotations made by a pathologist (Figure 3.9). CNNs with different layers were evaluated and the CNN with 3 layers performed the better with an AUC of 0.9021. The proposed method got a Dice coefficient of 75.86%, a positive predictive value of 71.62%, which corresponds to the ratio between the TP and the sum of the TP with the FP, and a negative predictive value of 96.77%, which corresponds to the ratio between the TN and the sum of the TN with the FN.

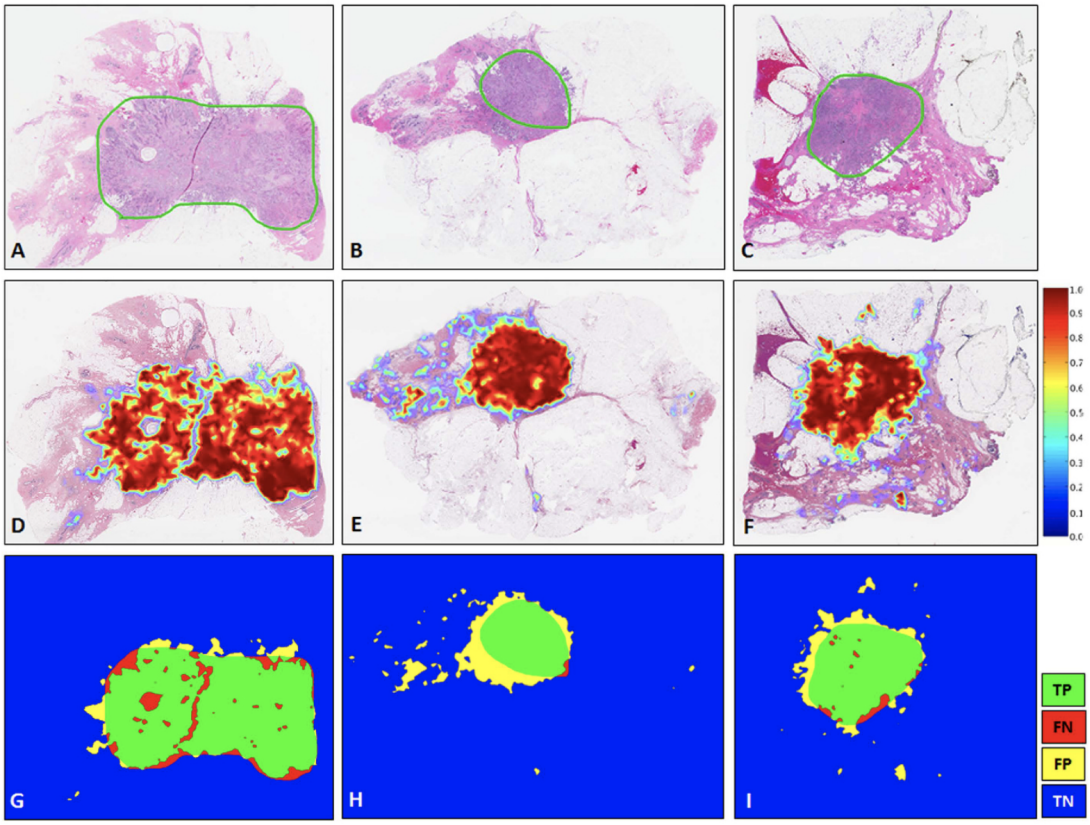


Figure 3.9: Example of the results obtained with the CNN. A-C corresponds to the annotations made by a pathologist, D-F corresponds to the probabilities obtained by the CNN, G-I shows the CNN results in terms of TP (green), FN (red), FP (yellow) and TN (blue) (extracted from [16]).

Rakhlin *et al.* [67] proposed a deep learning method for feature extraction with gradient boosting decision trees (GBDTs) for classification into 4 classes, *in situ* carcinoma, invasive carcinoma, benign and normal tissue. Results for 2 classes, benign and malignant, were also evaluated. GBDTs are based on the principle of boosting [25], which consists in having several elements added in a sequential form, where each element tries to minimize the loss of the previous elements. With GBDTs, the elements are represented as decision trees that are combined together, building a stronger model that has proven several times to be successful in classification. In addition, several different architectures of a CNN were evaluated, a ResNet-50, a Inception-v3 and a VGG-16

without the fully-connected layers. Each image was represented by 20 randomly extracted patches, which resulted in 20 extracted feature vectors from the CNNs. P-norm pooling [68] was used to obtain the final feature vector and can be calculated by equation 3.1, where N is the number of vectors, f_i is the feature vector correspondent to the patch i, p is the norm and $f_{pooling}$ is the final feature vector. In this case, 3-norm pooling was used.

$$f_{\text{pooling}} = \left(\frac{1}{N} \sum_{i=1}^{N} f_i^p \right)^{1/p} \tag{3.1}$$

For preprocessing the data, color normalization and data augmentation were performed. The results obtained for the classification into 4 and 2 different classes corresponded to an accuracy of 87.2% and 93.8%, respectively. For the 2 class classification, an AUC of 0.973 was obtained. A comparison between the results obtained by Araújo *et al.* [10] and Rakhlin *et al.* [67] can be made, where the obtained accuracy by the latter for both the 4 and 2 class classification was higher.

Pei *et al.* [9] proposed a framework for estimating the tumor cellularity by using different pre-trained CNNs, GBDTs and SVMs. Different combinations with the individual components were performed in order to evaluate which model gave the best results. The proposed method presents an initial step of preprocessing the data through color normalization and data augmentation (Figure 3.10). For data augmentation, cropping was not performed, since it could result in the loss of neoplastic or non-neoplastic cells from the image, leading to a possible incorrect diagnosis. Then, feature extraction is implemented by three CNNs, VGG-16, ResNet-50 and Inception-v3. Each patch is rotated and flipped and 8 variations of that patch are obtained, which means 8 feature vectors are obtained. The 3-norm pooling was performed to obtain the final feature vector. Moreover, for feature selection, minimum redundancy maximum relevance (mRMR) method [69] can be used. This method selects the set of features that are more relevant to the target class while the features are the least correlated between them. Principal Component Analysis (PCA) [25] is used for reducing the features that were selected. The step of feature selection grants the opportunity of increasing the speed of the training process, by eliminating unimportant data. For classification and prediction, GBDTs and SVMs can be used to determine tumor cellularity, as described in Figure 3.10. With this, a GBDT classifier was used and it was possible to separate the data that presented no neoplastic cells (tumor cell content = 0) from the ones that did contain them (tumor cell content > 0). This separation was necessary since the dataset was not balanced and contained a big number of images with zero tumor cellularity. To predict the percentage of tumor cellularity using GBDT, two different types of losses were optimized. For SVM, two methods were also proposed for the estimation. In order to evaluate the results, 3 metrics were used, ICC, τ_b and P_k. It was found that the best results were obtained for the prediction using SVM based on support vector regression (SVR) and utilizing a ResNet as a feature extractor. In this case, the values obtained for a 95% CI corresponded to an ICC of 0.95, a τ_b of 0.83 and a P_k of 0.93. Moreover, the proposed method exhibited better agreement in terms of ICC with the estimations made by a pathologist than the estimations between different pathologists.

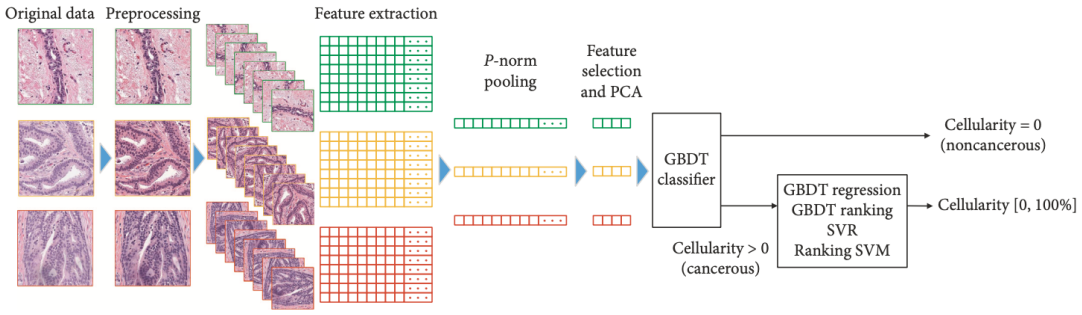


Figure 3.10: The proposed method starts with a preprocessing step, feature extraction is performed and the final feature vector is obtained. After, feature selection and dimensionality reduction are carried out to train the classification and prediction part (extracted from [9]).

Akbar *et al.* [70] proposed two different methods to estimate tumor cellularity, one based on a CNN with an Inception architecture and another based on traditional ML techniques. The proposed methods are presented in Figure 3.11. Two pathologists annotated patches in the histopathology images that they considered to be representative enough and classified the tumor cellularity into 1 of 4 classes, 0%, 1-30%, 31-70% and > 70%. For the training dataset, it was only used the annotations made by one of the pathologists, whereas for the test dataset both were used. This is the cause so it is possible to study the variability between the two annotations. The obtained ICC for a 95% CI between the pathologists was 0.89, demonstrating how the cellularity assessment between the experts can vary.

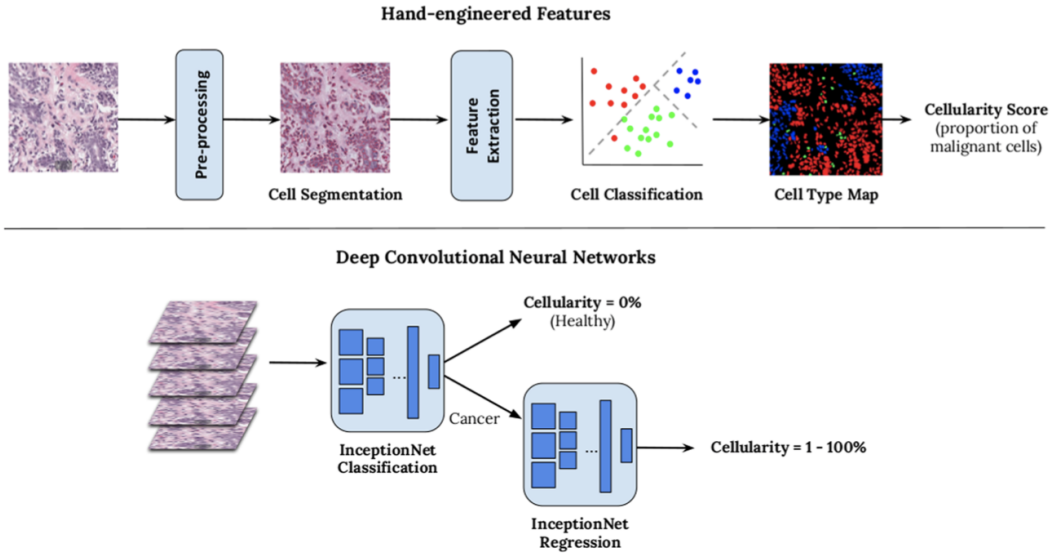


Figure 3.11: Proposed methods. The algorithm on top corresponds to the traditional ML techniques that will try to replicate the workflow of the pathologist. The method below corresponds to the deep learning approach based on CNNs (extracted from [70]).

The ML method that tried to replicate the workflow of the pathologist was the same as in

Peikari *et al.* [66], where the ICC for a 95% CI was 0.74 for one pathologist and 0.75 for the other. For the DL approach, a pre-trained InceptionNet was used to classify the patches into neoplastic cellularity > 0 or healthy tissue and another was used to estimate cellularity in the non-healthy patches. The CNN method resulted in an ICC for a 95% CI of 0.83 for one of the pathologists and 0.81 for the other. The comparison between the traditional ML approach and the DL method shows that the algorithm based on CNNs estimates the tumor cellularity in a more accurate way. Moreover, another approach was tested by fusing the 2 proposed methods, where the traditional ML techniques were used to estimate the cellularity in the cancerous tissue. With this, the ICC for a 95% CI was 0.76 for one of the pathologists and 0.79 for the other, showing that the replacement of the second CNN by a traditional ML approach does not provide better results.

3.3 Summary

A meticulous analysis of the histopathology slides by a pathologist is considered extremely relevant for the diagnosis of the patient. Most recently, several automated methods have been applied to breast cancer classification and segmentation, as well as to other types of cancer, such as lung cancer. Moreover, the application of DL-based algorithms has proven to be superior to the implementation of traditional ML techniques. In addition, the majority of the algorithms described rely on deep learning methods that are patch-based, since the WSI data has an extremely large resolution and analyzing the whole image would be impossible in terms of computational cost. A grand part of the proposed methods rely on pre-training the networks on large datasets, like the ImageNet dataset. The results indicated that the pre-training of the networks on this dataset did not exhibit better results. Therefore, it would be interesting to evaluate the results if the networks were pre-trained on a dataset that contained images with similar characteristics to the histopathology images. In addition, the approaches that removed the background of the images performed better than the others.

Chapter 4

Materials and Methods

This chapter presents a detailed description of the materials and methods used in this work. It starts with a brief description of the ImageNet dataset and its usual applications. Then, some information regarding the distribution of the PatchCamelyon dataset [71] and the SPIE-AAPM-NCI BreastPathQ dataset [72] is given. Furthermore, the steps necessary to preprocess the data of the SPIE-AAPM-NCI BreastPathQ are described and details concerning the ResNet-18 architecture are given. Finally, this chapter presents a detailed overview of the methodology used in this dissertation.

4.1 Datasets

4.1.1 ImageNet Dataset

As previously mentioned, the ImageNet dataset is a large-scale dataset with millions of images intended for object classification and detection [39]. The available data contains images with annotations in the form of binary labels and other images that contain the class label but are also associated with a bounding box representing the location of the object. Furthermore, a dataset that includes a large diversity of classes gives the opportunity to understand the effect each object class has on different algorithms and how to improve them. Examples of images from the ImageNet dataset are presented in Figure 4.1.

In addition, a very frequent and regular application of this dataset is to use it for transfer learning. Most datasets present a small number of samples and training a network from scratch, i.e. with the weights being initialized with random values, may not give the best results. It can be benefitial to pre-train the network with another dataset that is larger and using the target dataset just for training some layers. Most libraries that are used to train neural networks already offer the pre-training of the networks on the ImageNet dataset. This can be a good option to explore, since it is already implemented and it is not time consuming.

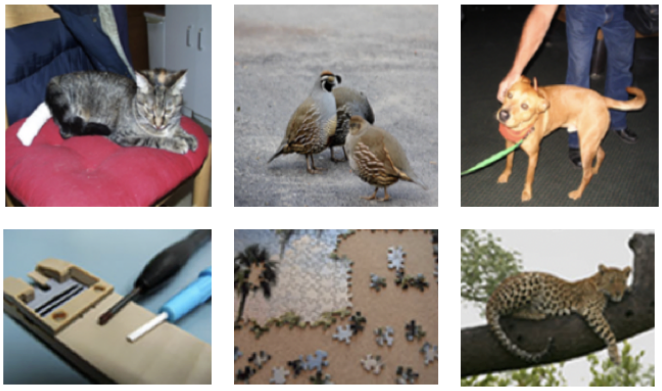


Figure 4.1: Examples of images from the ImageNet dataset (extracted from [39]).

4.1.2 PatchCamelyon Dataset

The PatchCamelyon (PCam) dataset is built using the data from the Camelyon16 challenge [73]. This competition used 399 WSI images stained with H&E of lymph nodes sections with tumor annotations to detect the presence of metastases in the lymph nodes. The PCam dataset is a large dataset, containing 327680 images [71]. These images correspond to patches of size 96×96 pixels extracted from the WSI data of Camelyon16. In addition, each patch is associated with a binary label corresponding to the presence of metastatic tissue. Positive labels in the PCam dataset mean that in the center region (32×32 pixels) there is at least 1 pixel that corresponds to tumor. Examples of images of the dataset are presented in Figure 4.2. Moreover, the dataset was created in order to be balanced in each split, i.e. each split contains a similar number of positive and negative patches (with and without metastases). The train/validation/test split is already given and the distribution is represented in Table 4.1.

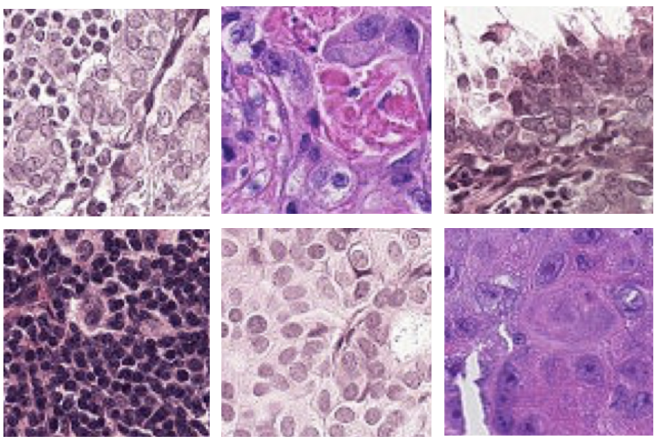


Figure 4.2: Examples of images from the PCam Dataset [71].

Furthermore, the mean and standard deviation of all training data was computed and all the

pixel values from the images were normalized accordingly. This process is done in order to facilitate the usage of the model with new images and transfer learning.

Table 4.1: Distribution of the PCam Dataset.

Dataset	Data Split (%)	No. of Images	No. of Images with Metastases	No. of Images without Metastases
Train	80	262144	131072	131072
Validation	10	32768	16399	16369
Test	10	32768	16391	16377

4.1.3 SPIE-AAPM-NCI BreastPathQ Dataset

The SPIE-AAPM-NCI BreastPathQ (BreastPathQ) dataset consists of patches of WSI data of residual invasive breast cancer of patients that underwent neoadjuvant therapy [72]. It contains 96 WSI images stained with H&E scanned with a magnification of 20× that were collected from 64 patients. The dataset already includes all images divided in patches, where each patch has a dimension of 512×512 pixels. Examples of images of the BreastPathQ dataset are presented in Figure 4.3.

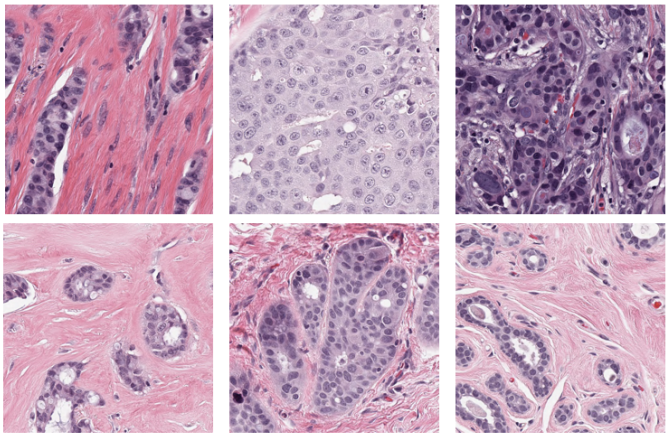


Figure 4.3: Examples of images from the BreastPathQ Dataset [72].

Furthermore, each patch was annotated by a pathologist, who gave a score between 0 and 1 representing the tumor cellularity for that patch. Additionally, the dataset split of train/validation/test is already given. However, the test dataset does not contain any labels associated to the patches since it was a requirement for the challenge to submit the scores given by the algorithms in the test set. Therefore, the test set will be discarded and the validation dataset will be considered for evaluation. In Table 4.2 is presented the distribution of the patches by the training and validation subsets that will be used. On top of that, the number of patches in each subset in relation to the tumor percentage is represented in Table 4.3. The cancer cellularity of the BreastPathQ dataset

follows a specific distribution. The tumor cellularity from 0% to 10% and from 90% to 100% varies in intervals of 1%. However, for values between 10% and 90%, the cellularity varies in intervals of 5%. It is important to notice that both the training and validation dataset have missing patches for some of the percentages, and therefore these values are not presented in the table. The missing tumor percentages are 4%, 6%, 9%, 91%, 94% and 96%.

Table 4.2: Distribution of the BreastPathQ Dataset.

Dataset	**No. of Images**
Train	2394
Validation	185

Table 4.3: Distribution of the images in the BreastPathQ dataset by each percentage value (label).

Label (%)	**No. of Images (Train)**	**No. of Images (Validation)**	**Label (%)**	**No. of Images (Train)**	**No. of Images (Validation)**
0	670	31	55	13	3
1	4	0	60	102	8
2	5	4	65	41	4
3	14	4	70	68	14
5	88	7	75	30	0
7	25	2	80	62	8
8	3	0	85	21	1
10	190	13	90	60	8
15	130	12	92	1	0
20	131	15	93	2	0
25	100	6	95	59	3
30	85	2	97	19	0
35	32	7	98	21	0
40	171	12	99	22	0
45	19	1	100	55	1
50	151	19			

4.1.3.1 Data Preprocessing

As mentioned before, the BreastPathQ presents an uneven distribution regarding tumor cellularity, which can become an additional difficulty for the selected methodology in equally differentiating

between classes. To deal with this problem, the labels were artificially changed by forcing them to have equal distances between them. In Figure 4.4 is presented the old label value with their new associated value, which is obtained by using equations 4.1 and 4.2. After that, all new label values are evenly spaced with a distance of 5 between them, resulting in labels ranging from -40 to 140. Finally, the obtained label values are converted to the range between 0 and 1, by adding 40 to each label value and dividing the result by 180. These new labels are used to train the model. However, to calculate the evaluation metrics is necessary to convert them back into the old labels by using equations 4.1 and 4.2 but solving them in order to *old_label*. Then, it is only necessary to round the value to the nearest possible bin.

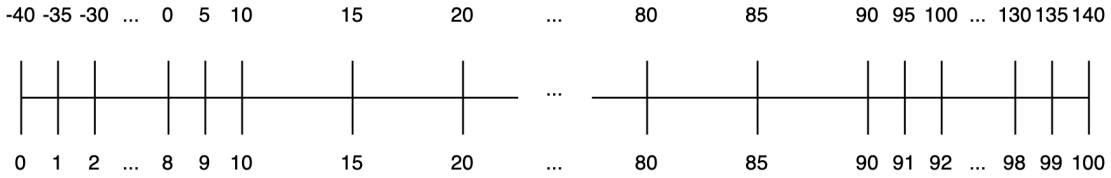


Figure 4.4: Old label values (bottom) with their corresponding new label values (top).

$$new\_label = 10 + 5(old\_label - 10), \quad old\_label < 10 \tag{4.1}$$

$$new\_label = 90 + 5(old\_label - 90), \quad old\_label > 90 \tag{4.2}$$

At last, to deal with the problem of stain inconsistency and the usage of different scanners, the implementation of two different color normalization techniques was performed by using the *StainTools* package [74]. This tool offers the implementation of the methods suggested by Macenko *et al.* [65] and Vahadane *et al.* [50]. Macenko *et al.* [65] proposed a method based on singular value decomposition and does not make any assumption regarding the fact that the pixel values cannot be negative. Vahadane *et al.* [50] proposed a solution based on the method developed by Macenko et al. [65], but already considered that a stain density for each pixel cannot present a negative value, since that would mean emitting light. In addition, the authors assumed that for a given image, the proportions of the stains in relation to a specific biological structure are the same and that more than one stain can tint the structure. They also assume that for each pixel there can only be one biological structure. Both methods use a chosen target image as the color base and change the source image accordingly, but maintaining the initial stain concentrations. In Figure 4.5 are represented some patches with and without color normalization using the two described techniques. From the observation of this Figure it is not straightforward to see clear differences between these two methods. However, a close inspection of this Figure shows that the pink stain associated with the application of the Macenko et al. [65] approach is slightly more saturated than in the other method. It is important to notice that both these methods depend greatly on the choice of the target image.

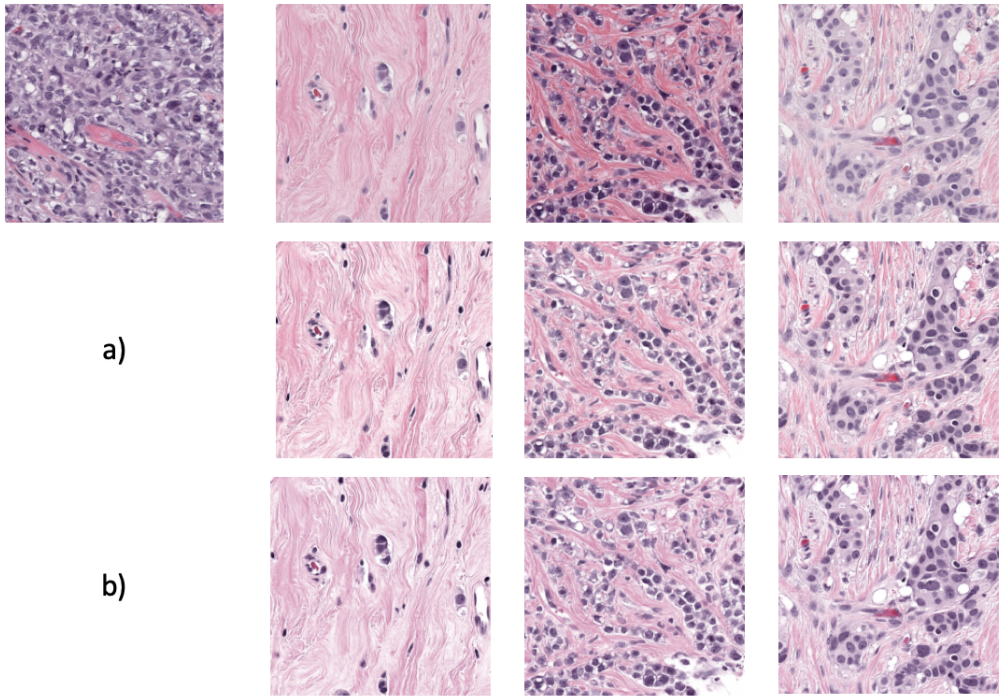


Figure 4.5: Effect of color normalization on some patches. On the top left corner is presented the target image. The top row represents the images without any type of color normalization. In a) is represented the same images normalized with respect to the target image using the Macenko *et al.* [65] approach. In b) images are normalized with the Vahadane *et al.* [50] approach.

4.2 Network Architecture

As mentioned before, several architectures could be used for the main objective of this work, i.e. the quantification of the cancer cells in WSI data, such as ResNets, VGGs and Inceptions. The performance of these 3 different architectures was compared by Pei *et al.* [9] using the Breast-PathQ dataset. In that study, it was found that the best performing model was the one that used a ResNet-50. Therefore, we have selected a ResNet based architecture for this work. Moreover, in order to investigate the performance when using a simpler version of this type of networks, we will be using a ResNet-18.

Figure 4.6 presents a diagram that exemplifies the architecture of the ResNet-18 [44]. As with the others ResNet networks, this network was introduced in the context of the ImageNet classification task. This network has 18 layers with 4 basic blocks that repeat themselves. It is possible to calculate the size of the image after each layer by using equation 4.3, in which W stands for the image width, F for the size of the filter, P to the size of the padding and S to the size of the stride. In each block the first layer has a stride of 2, except for the first block. Each convolutional layer is followed by a Rectified Linear Unit (ReLU) activation function. In the end, there is a fully connected layer with 1000 neurons that represent the 1000 classes of the classification task of the ImageNet dataset.

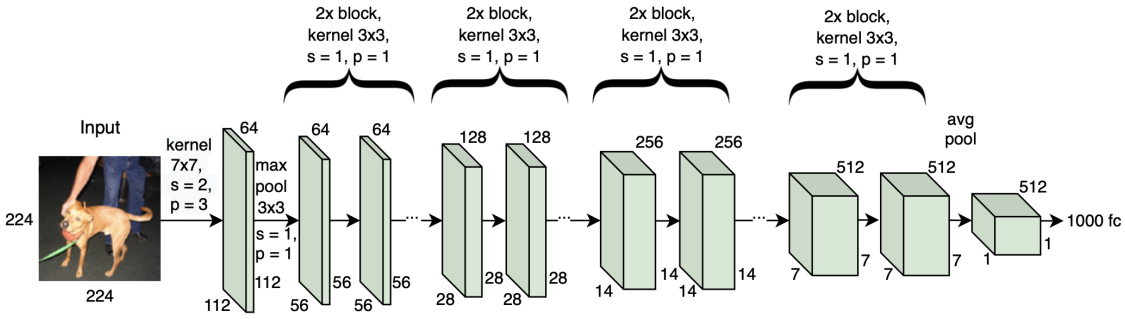


Figure 4.6: Diagram of the ResNet-18 network architeture [44] for images of 224x224 px of the ImageNet dataset. The letters s and p stand for stride and padding, respectively.

$$new\_width = (W - F + 2P)/S + 1 \tag{4.3}$$

4.3 Methodology

Figure 4.7 presents the overview of the workflow that will be followed to investigate the prediction of the percentage of tumor cells in the BreastPathQ dataset. This pipeline consists in comparing 3 different approaches:

- using the ResNet-18 architecture without any pre-training to predict the percentage, i.e. the network weights are initially with random values;
- using the ResNet-18 pre-trained on the ImageNet dataset, which is already available within the utilized framework;
- pre-training the network on the PCam dataset and only then using the ResNet-18 for predicting the cancer cells in the BreastPathQ patches.

In order to avoid overfitting, two different data augmentation techniques were applied to both the PCam and the BreastPathQ dataset. Image patches in the training dataset could be vertically flipped with a probability of 50% and a random perspective with a 0.5 distortion factor with a 50% probability of being applied was also implemented. This random perspective uses a bilinear interpolation. In Figure 4.8 is represented four possible variants of a patch of the training dataset. Both the network, the data augmentation process and the training phase were all implemented by using the *PyTorch* framework [75].

For the three approaches mentioned above, several tests will be made with different conditions regarding distinct factors. The training data in the BreastPathQ dataset is heavily unbalanced as it presented a large number of patches with zero cellularity (670 samples). Regarding this information, in order to verify the influence of this class during the training phase, tests were made with and without the presence of these patches in the training dataset. Additionally, it will be studied the influence of the images not being color normalized, being color normalized according

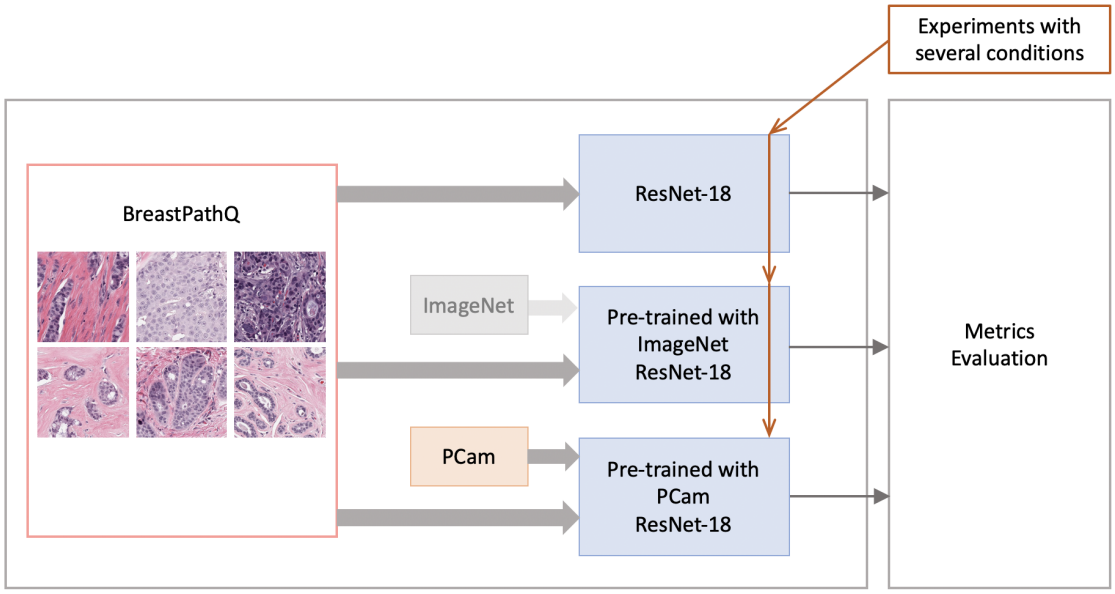


Figure 4.7: Overview of the workflow established here to study the prediction of the cancer cells percentage in the BreastPathQ dataset.

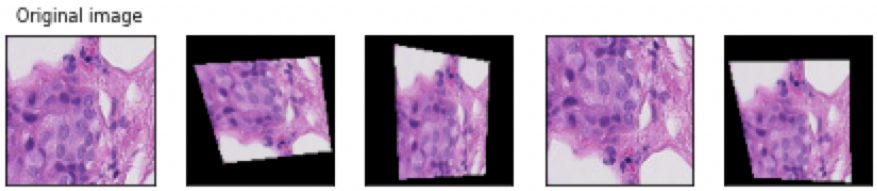


Figure 4.8: Original patch and 4 different variants that are possible to appear in the training dataset. From left to right: the fourth image was vertically flipped, but it did not undergo any random perspective transformation. The third and fifth images were not vertically flipped, but a random perspective was applied. Finally, the second image was both vertically flipped and a random perspective transformation was applied.

to the Macenko *et al.* [65] approach and to the Vahadane *et al.* [50] approach. Furthermore, an investigation regarding the classifier at the end of the network, i.e. the number of fully connected layers will be made, with two different scenarios considered: one fully connected layer, from 512 neurons to 1 with a sigmoid at the end, or two fully connected layers, from 512 neurons to 64 to 1 with a sigmoid at the end. The used optimizer for all cases is the Stochastic Gradient Descent.

The combination of all these possibilities results in twelve different tests, which will be applied to the ResNet-18 trained from scratch and to the ResNet-18 pre-trained with the ImageNet dataset. Since the training process in the ResNet-18 pre-trained on ImageNet or on PCam consists in only training the classifier part and having the rest of the network frozen, an initial investigation concerning unfreezing some convolutional layers will also be conducted. The performance of all these tests will be compared according to several metrics: τ_b, P_k and ICC (see Section 2.5).

In order to simplify the analysis of cancer cells quantification using the ResNet-18 pre-trained

on the PCam dataset, only a selected number of tests were conducted. Therefore, from the results obtained with the ResNet-18 pre-trained with the ImageNet, the four cases that show the best performance and one with zero patches in the training dataset will be investigated with the ResNet-18 pre-trained with the PCam.

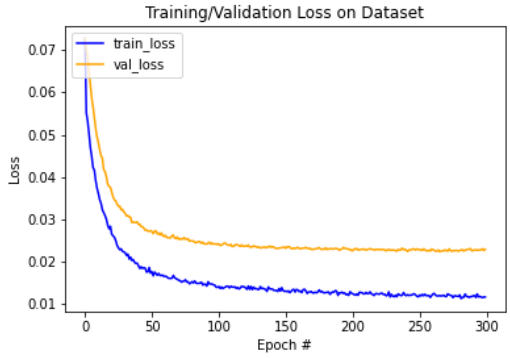
Finally, since implementing all twelve tests considered is time-consuming, an attempt to try to make a first choice of the hyperparameters (learning rate, number of epochs and batch size) of the model was made. Additionally, to investigate the influence of the loss function, a simple test considering two different loss functions, the mean squared error (MSE) and the mean absolute error (MAE), was conducted.

Learning Rate

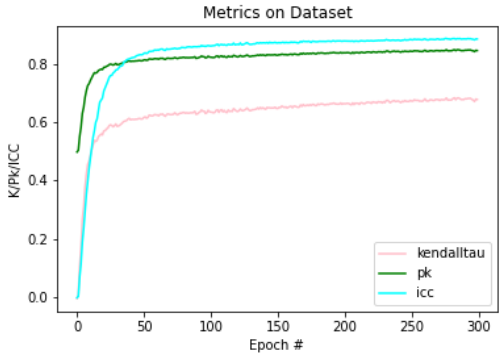
The choice of an adequate learning rate is very important since it affects the performance of the model. As an example, the evolution of the training and validation curves, and of the evaluation metrics on the validation dataset for 300 epochs are represented in Figure 4.9 for three different learning rates, 0.003, 0.004 and 0.006. These results were obtained using the ResNet-18 pre-trained on ImageNet when the loss function was the MSE, the classifier consisted in 512 neurons to 1 with a sigmoid at the end and the batch size during training was 114. Furthermore, the model was trained by removing the image patches of zero tumor cellularity from the training dataset and no color normalization technique was employed. Note that the learning rate affects the general evolution of the represented curves and also the final values at 300 epochs. In particular, an oscillatory behaviour is visible in Figure 4.9e, hinting that the ideal learning rate is lower than 0.006. Although the loss curves during training and the evolution of the metrics are good for having a general idea of their progress, it is not always straightforward to select the adequate learning rate from just the plot of these curves. Therefore, in order to make this selection, one could easily compare the values of the evaluation metrics at a specific point, for example, at the point where the validation loss is minimum. Table 4.4 presents the epoch for the minimum validation loss and the corresponding values of the evaluation metrics for the learning rates of 0.003, 0.004 and 0.006. It is possible to observe that the learning rate that results in higher values of the evaluation metrics is 0.003. Therefore, from the three values of the learning rates considered, we should select 0.003. However, the differences between the evaluation metrics corresponding to each of the three learning rates is not large.

Table 4.4: Evaluation metrics obtained for the smaller value of the validation loss curves for the learning rates of 0.003, 0.004 and 0.006.

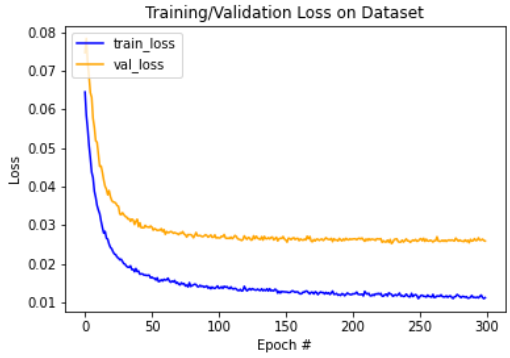
Learning Rate	**Best Epoch**	**Val Loss**	τ_b	P_k	**ICC**
0.003	266	0.0224	0.6762	0.8449	0.8838
0.004	230	0.0252	0.6280	0.8200	0.8640
0.006	158	0.0235	0.6595	0.8352	0.8726



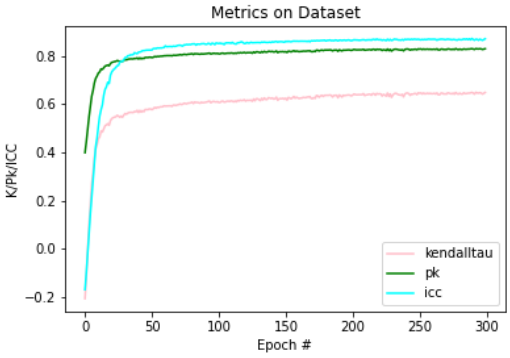
(a) Training and Validation Curves for the LR of 0.003.



(b) Metrics for the Validation Dataset for the LR of 0.003.

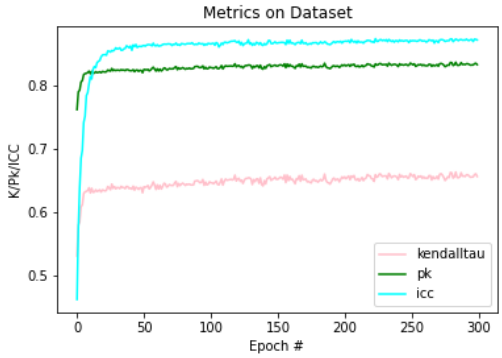


(c) Training and Validation Curves for the LR of 0.004.



(d) Metrics for the Validation Dataset for the LR of 0.004.

(e) Training and Validation Curves for the LR of 0.006.



(f) Metrics for the Validation Dataset for the LR of 0.006.

Figure 4.9: Training and Validation Curves for 300 epochs (left) and Evaluation Metrics on the Validation Dataset (right) for a Learning Rate (LR) of 0.003 (top), 0.004 (middle) and 0.006 (bottom).

Even though a learning rate of 0.003 would be adequate to study the evolution of the model under the considered conditions, the distinct scenarios studied in this work, such as, the two loss functions and the usage of color normalization techniques, influence the behaviour of the model, resulting in different ideal learning rates. Hence, in all the performed tests, the value of 0.003 was initially used for the learning rate. This value was subsequently adjusted if necessary. For

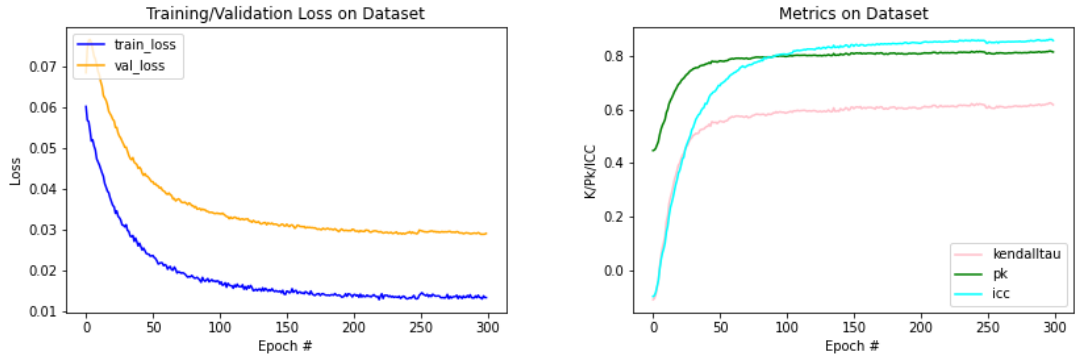
example, if the loss curves are decreasing at a very slow rate, the learning rate could be increased, whereas it should be decreased in case of oscillatory behaviour of the curves.

Number of Epochs

Since 300 epochs was enough for the study in Figure 4.9, the tests that follow consisted in training the network for 400 epochs most of the times, except for when it was considered to be necessary more epochs.

Batch Size

The batch size used during training also affects the model behaviour. Figure 4.10 presents the evolution of the training and validation curves over 300 epochs and the evaluation metrics on the validation dataset for the same conditions of Figure 4.9a, but with a batch size during training of 228 and not 114. It is possible to observe that in Figure 4.10 the evolution of the curves is slower than in Figure4.9a and, therefore, every test requires more epochs until reaching a point where the validation loss is practically steady. Hence, the batch size of 114 during training was selected to be maintained throughout this work for all tests done.



(a) Training and Validation Curves.

(b) Metrics for the Validation Dataset.

Figure 4.10: Training and Validation Curves for 300 epochs (left) and Evaluation Metrics on the Validation Dataset (right) for a Learning Rate (LR) of 0.003 and a Batch Size of 228.

Loss Function

In order to choose a loss function to use throughout this dissertation, a single test was made by using two different loss functions, the MSE and the MAE. Table 4.5 presents the values of the evaluation metrics for the ResNet-18 trained from scratch when using the MSE and MAE loss function. These results were obtained when the classifier consisted in 512 neurons to 1 with a sigmoid at the end and the batch size during training was 114. Furthermore, the model was trained by removing the image patches of zero tumor cellularity from the training dataset and no color normalization technique was employed. It is possible to observe in Table 4.5 that the values of

the evaluation metrics are very similar for both loss functions, although slightly higher for the MSE loss function. Therefore, the MSE loss function was selected as the loss function to be used throughout this work.

Table 4.5: Evaluation metrics at the epoch for the minimum validation loss for the MSE and MAE loss function.

	τ_b	P_k	**ICC**
MSE	0.640	0.825	0.838
MAE	0.613	0.811	0.802

4.4 Summary

A brief description of the two datasets that will be used throughout this work was given. One of these datasets is the PCam and will be employed in the pre-training of the network. The images in this dataset correspond to WSI images with similar characteristics to the images of the dataset that will be used to train and quantify the cancer cells (BreastPathQ). This chapter also describes the chosen network architecture and the methodology that was defined to evaluate the influence of different factors in the performance of this model, such as the removal of the images with zero cellularity from the training dataset, the usage of color normalization techniques and the classifier. Preliminary studies were conducted to define the value of the hyperparameters that will be used in the following chapter, that is, the initial learning rate will be 0.003, the minimum number of epochs is 400 and the batch size during training is 114. Additionally, an investigation regarding the loss function was made, and the MSE was selected as the loss function to be used in this work.

Chapter 5

Cancer Cells Quantification

This chapter presents the results obtained with the methodology defined in the previous chapter. It also includes a discussion of the results and compares the different network and training dataset configurations.

5.1 ResNet-18 trained from Scratch

This section presents the results obtained with a ResNet-18 network that was trained from scratch using the BreastPathQ dataset. Several tests are made regarding the presence of zeros in the training dataset, the number of fully connected layers at the end of the network and the effect of two different color normalization techniques.

5.1.1 Tests

As previously mentioned in Chapter 4, experiments changing three different parameters will be made, which results in twelve different tests. In this work, in order to simplify the identification of a particular test, a numerical code was defined, where each digit has a different meaning. Table 5.1 presents the three conditions that will be altered, the position of the digit in the code of the test (starting from the left), the code numbers that they can have and respective meaning. The code numbers have three digits, where the first digit corresponds to the zeros code, the second to the classifier and the third to the color normalization technique used. For example, the digit corresponding to zeros can have two different values, 0 or 1, where the former means that the training dataset will not have patches with zero cellularity and the latter to the training with the full dataset. Thus, the test with the number code 002 means that the training of the network was with the dataset without zeros, the classifier had 512 neurons to 1 with a sigmoid at the end and that the color normalization used was the Vahadane approach.

Table 5.2 summarizes all the tests that were performed. In total, twelve tests were implemented. The Table 5.2 contains the total number of epochs for which the model was trained, the learning rate that was used, the epoch that resulted in the minimum value of the validation loss and its corresponding value, the value of the three evaluation metrics, τ_b, P_k and ICC, for the minimum

validation loss epoch and their maximum values. For each one of these six metrics, the highest value across all cases is highlighted in bold. Finally, the Table 5.2 also contains the equation of the linear regression that better approximates the estimations of the model concerning the validation dataset. It also contains the R^2 value that measures the level of agreement between the linear regression and the estimations on the validation dataset. Ideally, if the model was perfect, the equation of the linear regression would be of the type $y = x$ and the R^2 value would be 1. Detailed plots of the evolution of the training and validation loss, evaluation metrics and scatter plots are presented in Appendix A.

From the observation of the Table 5.2, it is possible to verify that the maximum values of the evaluation metrics are similar to the values corresponding to the minimum loss. Therefore, since the minimum validation loss is considered to be a convenient stopping point, in the following analysis, the evaluation criteria at this epoch will be used.

Table 5.1: Variable parameters during the tests, their attributed code and meaning.

Variable Parameters	**Position**	**Value**	**Meaning**
Zeros	1	0	No images with zero cellularity
		1	Images with zero cellularity
Classifier	2	0	512→1
		1	512→64→1
Color Normalization	3	0	No color normalization
		1	Macenko approach [65]
		2	Vahadane approach [50]

Table 5.2: Summary of the Results for the ResNet-18 trained from scratch. For detailed plots of the evolution of the training and validation loss, evaluation metrics and scatter plots see Appendix A.

Test	**Epochs**	**LR**	**Best Epoch**	**Val Loss**	τ_b	P_k	**ICC**	τ_b **max**	P_k **max**	**ICC max**	**y = mx + b**	R^2
000	400	0.001	357	0.024	**0.640**	**0.825**	**0.838**	**0.647**	**0.829**	**0.842**	y = 0.78x + 0.06	0.65
001	350	0.004	338	0.029	0.578	0.795	0.802	0.635	0.822	0.802	y = 0.76x + 0.07	0.58
002	400	0.004	362	0.029	0.567	0.788	0.768	0.587	0.800	0.807	y = 0.69x + 0.05	0.45
010	400	0.001	325	0.025	0.636	0.823	0.830	0.638	0.824	0.838	y = 0.78x + 0.07	0.63
011	250	0.002	242	0.038	0.607	0.811	0.768	0.608	0.811	0.768	y = 0.82x + 0.16	0.51
012	400	0.004	318	0.031	0.559	0.785	0.777	0.590	0.802	0.800	y = 0.7x + 0.08	0.50
100	400	0.001	222	0.028	0.585	0.799	0.701	0.593	0.802	0.709	y = 0.66x + 0.06	0.33
101	250	0.001	119	0.026	0.587	0.799	0.717	0.599	0.803	0.751	y = 0.59x + 0.08	0.24
102	250	0.001	158	0.025	0.602	0.807	0.749	0.608	0.810	0.749	y = 0.68x + 0.09	0.44
110	400	0.001	357	0.028	0.583	0.798	0.712	0.598	0.806	0.725	y = 0.7x + 0.06	0.41
111	400	0.001	181	0.024	0.616	0.814	0.764	0.629	0.820	0.767	y = 0.64x + 0.08	0.39
112	400	0.001	220	0.025	0.607	0.809	0.735	0.618	0.814	0.748	y = 0.63x + 0.07	0.34

5.1.2 Discussion of Results

5.1.2.1 Effect of Image Patches with Zero Cellularity

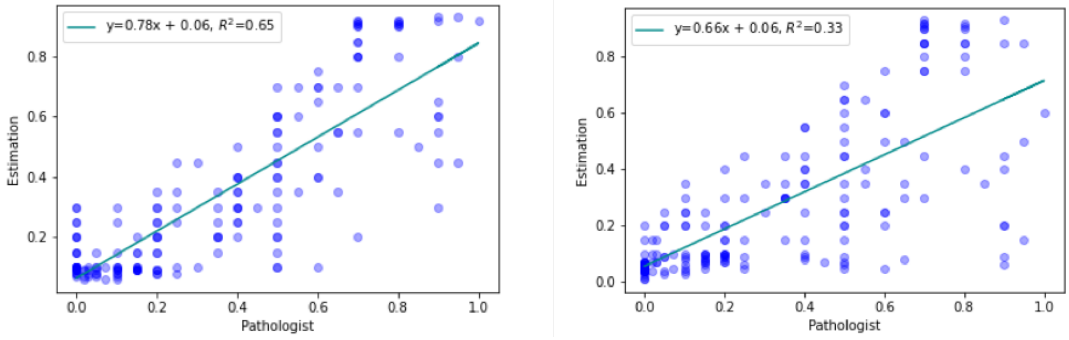
Although all parameters may impact the performance of the model, removing the patches with zero cellularity from the training data may alter the results significantly, since the dataset is highly unbalanced.

Table 5.3 presents the minimum and maximum values of the evaluation metrics corresponding to the epoch with minimum validation loss for the study of the cases with and without zero patches for the ResNet-18 trained from scratch. It is possible to observe that the ranges of τ_b and P_k are similar in both cases. However, the values of the ICC are significantly larger in the case where the dataset does not contain images with zero cellularity. The difference in the behaviour of these metrics could be explained by their nature: both τ_b and P_k are metrics that evaluate ordinal correlation, whereas ICC evaluates the level of agreement between measurements. In effect, since the dataset with zero patches is heavily unbalanced, containing a large amount of images with zero cellularity, the training will be naturally biased towards the classification of those images. On the contrary, the dataset without the zero patches is more balanced and allows to train a model that will perform better for all the classes. This difference in the classification can be observed in Figure 5.1, which presents the estimations made by the model on the validation dataset for the tests 000 and 100. For visualisation purposes, in Figure 5.1 each dot corresponds to an image patch from the validation dataset, where different shades of blue were used to represent the level of overlapping. Darker colors are related to a higher number of overlapping patches, i.e., the model classified several images with the same percentage. In Figure 5.1b it is possible to observe that more patches from the validation dataset are classified with a low tumor percentage than in Figure 5.1a. Moreover, it is also visible in Figure 5.1b that for images classified by the pathologist with tumor cellularity around 50% or larger, the patches are more spread, suggesting a poor capacity of the model in classifying images with large cellularities.

Table 5.3: Minimum and maximum values of the evaluation metrics at the epoch for the minimum validation loss in relation to the training dataset for the ResNet-18 trained from scratch.

	τ_b	P_k	**ICC**
Without zero patches	0.559– 0.640	0.785 – 0.825	0.768 – 0.838
With zero patches	0.583 – 0.616	0.798 – 0.814	0.701 – 0.764

Additionally, it is important to notice that the slopes of the linear regression and the R^2 values in Figure 5.1 are consistent with the model trained without the zero patches performing better. This behaviour is also visible in all the remaining studied cases, as it can be observed in Table 5.4. Both the values of the slope and the R^2 are always larger for the case where the training dataset did not contain images with zero cellularity, suggesting that these values have a similar behaviour to the ICC metric.



(a) Test 000. (b) Test 100.

Figure 5.1: Scatter Plot showing the level of agreement between the estimations of the model with the predictions of the pathologist on the validation dataset. A linear regression that better fits the data is also shown. The different shades of blue dots intend to show the overlapping of patches.

Table 5.4: Minimum and maximum values of the linear regression slope and the R^2 value at the epoch for the minimum validation loss for the ResNet-18 trained from scratch.

	Slope	R^2
Without zero patches	0.69 – 0.82	0.45 – 0.65
With zero patches	0.59 – 0.7	0.24 – 0.44

5.1.2.2 Effect of the Classifier

The classifier at the end of the network can be composed of one or several fully connected layers and, since in this work we are dealing with a regression problem, it will always end in a single neuron. Here, the effect of using two different classifiers will be investigated: one from 512 neurons to 1 with a sigmoid at the end, and another that consists in 512 neurons to 64 to 1 neuron with a sigmoid at the end.

Table 5.5 presents the minimum and maximum values of the metrics for the minimum validation loss in relation to the classifier. It is possible to observe that the values of the three evaluation metrics are very similar for both classifiers, suggesting that the addition of a second layer to the classifier does not improve the performance of the model.

Table 5.5: Minimum and maximum values of the evaluation metrics at the epoch for the minimum validation loss in relation to the classifier for the ResNet-18 trained from scratch.

Classifier	τ_b	P_k	**ICC**
$512 \rightarrow 1$	0.567 – 0.640	0.788 – 0.825	0.701 – 0.838
$512 \rightarrow 64 \rightarrow 1$	0.559 – 0.636	0.785 – 0.823	0.712 – 0.830

5.1.2.3 Effect of the Color Normalization

As previously mentioned, color normalization techniques may be useful to deal with stain inconsistencies and the usage of different scanners. In order to study the influence of these techniques, two different color normalization approaches were implemented, Macenko *et al.* [65] and Vahadane *et al.* [50], and compared to the case where no normalization was used.

In Table 5.6 the minimum and maximum values of the evaluation metrics in the epoch corresponding to the minimum validation loss are presented for the three cases: (i) without color normalization, (ii) with the Macenko approach and (iii) with the Vahadane approach.

Table 5.6: Minimum and maximum values of the evaluation metrics at the minimum validation loss epoch for different color normalization approaches and for the ResNet-18 trained from scratch.

	τ_b	P_k	**ICC**
No color normalization	0.583 – 0.640	0.798 – 0.825	0.701 – 0.838
Macenko approach	0.578 – 0.616	0.795 – 0.814	0.717 – 0.802
Vahadane approach	0.559 – 0.607	0.785 – 0.809	0.735 – 0.777

It is possible to observe that the highest values of the three evaluation metrics occur for the case where no color normalization technique was applied. This could be explained by the choice of the target image used for the normalization process, i.e., a different target image may yield a better result.

However, a different conclusion would be obtained if only the results corresponding to the training dataset containing the images with zero cellularity were used. Effectively, from Table 5.2 it is possible to observe that the evaluation metrics for tests 100, 101 and 102 have a different behaviour. The metrics have their lower values for the case 100, where no color normalization was applied, followed by the ones with the Macenko approach (test 101), and, finally, the best results are obtained for the Vahadane approach (test 102). A similar behaviour is visible when comparing the cases 110 with 111 and 112, where the value of the metrics is higher for the cases where color normalization was used, with the Macenko approach outperforming the other two.

5.2 ResNet-18 pre-trained on ImageNet

This section addresses the task of cancer cells quantification using a ResNet-18 network that was pre-trained on the ImageNet dataset.

5.2.1 Tests

Table 5.7 summarizes all the tests that were performed. Detailed plots of the evolution of the training and validation loss, evaluation metrics and scatter plots are presented in Appendix A.

Similarly to what happens with the ResNet-18 trained from scratch, the maximum values of the evaluation metrics are close to the values corresponding to the minimum loss. Therefore, the evaluation criteria at this epoch will also be used here.

Table 5.7: Summary of the Results for the pre-trained ResNet-18 on the ImageNet Dataset. For detailed plots of the evolution of the training and validation loss, evaluation metrics and scatter plots see Appendix A.

Test	**Epochs**	**LR**	**Best Epoch**	**Val Loss**	τ_b	P_k	**ICC**	τ_b **max**	P_k **max**	**ICC max**	**y = mx + b**	R^2
000	400	0.003	428	0.023	**0.703**	**0.858**	**0.895**	**0.707**	**0.860**	**0.896**	y = 0.93x + 0.06	0.79
001	200	0.006	144	0.024	0.639	0.825	0.873	0.642	0.827	0.878	y = 0.87x + 0.05	0.75
002	600	0.006	578	0.024	0.628	0.820	0.878	0.634	0.823	0.878	y = 0.91x + 0.03	0.76
010	700	0.003	571	0.022	0.678	0.845	0.890	0.684	0.849	0.892	y = 0.91x + 0.05	0.78
011	200	0.006	139	0.025	0.633	0.822	0.874	0.638	0.825	0.879	y = 0.88x + 0.06	0.75
012	500	0.006	481	0.024	0.630	0.821	0.876	0.637	0.824	0.880	y = 0.88x + 0.05	0.75
100	200	0.003	47	0.032	0.617	0.811	0.619	0.633	0.820	0.631	y = 0.56x	-0.27
101	350	0.003	45	0.031	0.587	0.798	0.628	0.602	0.806	0.654	y = 0.56x + 0.02	-0.15
102	350	0.003	42	0.037	0.534	0.770	0.560	0.543	0.774	0.583	y = 0.49x + 0.02	-0.61
110	400	0.003	81	0.026	0.655	0.832	0.712	0.661	0.835	0.715	y = 0.59x	-0.1
111	150	0.003	72	0.025	0.666	0.838	0.702	0.668	0.839	0.722	y = 0.61x + 0.01	0.07
112	150	0.003	66	0.027	0.625	0.816	0.690	0.638	0.822	0.708	y = 0.6x + 0.01	0.04

5.2.2 Discussion of Results

5.2.2.1 Effect of Image Patches with Zero Cellularity

Table 5.8 presents the minimum and maximum values of the evaluation metrics corresponding to the epoch with minimum validation loss for the study of the cases with and without zero patches in the training dataset.

Table 5.8: Minimum and maximum values of the evaluation metrics at the epoch for the minimum validation loss in relation to the training dataset for the ResNet-18 pre-trained on ImageNet.

	τ_b	P_k	**ICC**
Without zero patches	0.628 – 0.703	0.820 – 0.858	0.873 – 0.895
With zero patches	0.534 – 0.666	0.770 – 0.838	0.560 – 0.712

Similarly to what happened with the ResNet-18 trained from scratch, the evaluation metrics present higher values when the training dataset does not include the images with zero cellularity. Actually, the difference between the evaluation metrics for the cases with and without zero patches is considerably more pronounced now than with the ResNet-18 trained from scratch. For example, the ICC metric for the case of the ResNet-18 trained from scratch ranges between 0.757 - 0.838 (without zero patches) and between 0.701 - 0.764 (with zero patches), whereas for the ResNet-18 pre-trained with ImageNet this metric ranges between 0.873 - 0.895 (without zero patches) and between 0.560 - 0.712 (with zero patches).

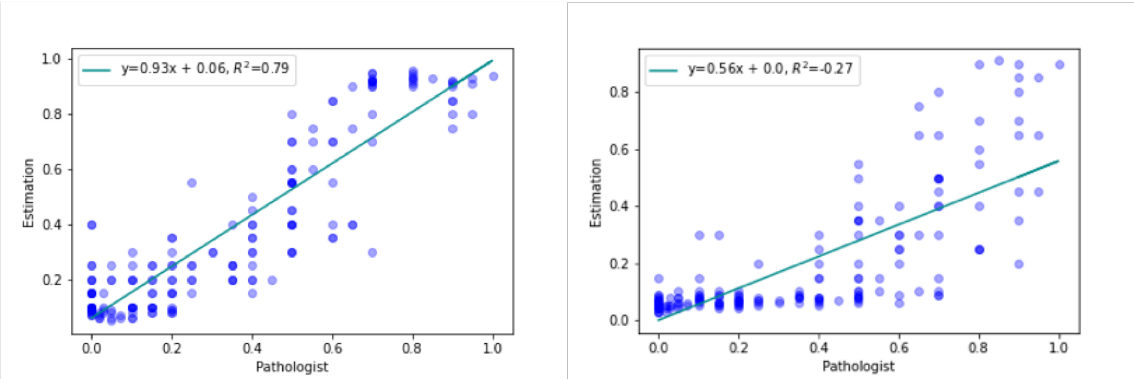
The difference between the network behaviour for the cases with and without the zero patches in the training dataset can be easily observed in Figure 5.2, which presents the estimations made by the model on the validation dataset for the tests 000 and 100. Effectively, in Figure 5.2b it is possible to observe a notorious bias of the estimations towards the lower tumor cellularities. This bias is not visible in Figure 5.2a. Furthermore, similarly to the ResNet-18 trained from scratch, the predictions made by the model trained with the zero patches tend to misclassify the larger cellularities.

The range of values of the linear regression and R^2 for all the studied cases are presented in Table 5.9 and are consistent with the model trained without zero patches performing better.

Table 5.9: Minimum and maximum values of the linear regression slope and the R^2 value at the epoch for the minimum validation loss for the ResNet-18 pre-trained on ImageNet.

	Slope	R^2
Without zero patches	0.87 – 0.93	0.75 – 0.79
With zero patches	0.49 – 0.61	-0.61 – 0.07

Finally, it is important to mention that the comparison of the results in Table 5.8 with those in Table 5.3 for the training dataset without zero patches suggests that the ResNet-18 pre-trained with the ImageNet outperforms the one trained from scratch.



(a) Test 000.

(b) Test 100.

Figure 5.2: Scatter Plot showing the level of agreement between the estimations of the model with the predictions of the pathologist on the validation dataset. A linear regression that better fits the data is also shown. The different shades of blue dots intend to show the overlapping of patches.

5.2.2.2 Effect of the Classifier

Table 5.10 presents the minimum and maximum values of the metrics for the minimum validation loss in relation to the classifier. It is possible to observe that the values of the three evaluation metrics are very similar for both classifiers. An identical conclusion was obtained for the ResNet-18 trained from scratch.

Table 5.10: Minimum and maximum values of the evaluation metrics at the epoch for the minimum validation loss in relation to the classifier for the ResNet-18 pre-trained on ImageNet.

Classifier	τ_b	P_k	**ICC**
$512 \rightarrow 1$	0.534 – 0.703	0.770 – 0.858	0.560 – 0.895
$512 \rightarrow 64 \rightarrow 1$	0.625 – 0.678	0.816 – 0.845	0.690 – 0.890

5.2.2.3 Effect of the Color Normalization

In Table 5.11 the minimum and maximum values of the evaluation metrics in the epoch corresponding to the minimum validation loss are presented for the three cases: (i) without color normalization, (ii) with the Macenko approach and (iii) with the Vahadane approach.

Once again, it is possible to observe that the highest values of the three evaluation metrics occur for the case where no color normalization technique was applied, which could be explained by the choice of the target image used for the normalization process.

Table 5.11: Minimum and maximum values of the evaluation metrics at the minimum validation loss epoch for different color normalization approaches for the ResNet-18 pre-trained on ImageNet.

	τ_b	P_k	**ICC**
No color normalization	0.617 – 0.703	0.811 – 0.858	0.619 – 0.895
Macenko approach	0.587 – 0.666	0.798 – 0.838	0.628 – 0.874
Vahadane approach	0.534 – 0.630	0.770 – 0.821	0.560 – 0.878

5.3 ResNet-18 pre-trained on PCam

This section is dedicated to cancer cells quantification using a network that is pre-trained with the PCam dataset. It starts with the task of metastases classification in the PCam dataset and the selection of the model with the best performance. Afterwards, this pre-trained model is used in the quantification of cancer cells and results are compared with those obtained with the model that was pre-trained with the general dataset.

5.3.1 Proposed Method

The pipeline implemented to predict the percentage of cancer cells is represented in Figure 5.3. This method consists in two distinct parts. The first part comprises a binary classification task to classify image patches of the PCam dataset as containing metastatic tissue or not. This will be done by using a specific CNN, the ResNet-18 network [44]. This task is used in this pipeline to make use of transfer learning. As mentioned before, the PCam dataset consists of 96×96 px patches and contains 262144 images for the training dataset and 32768 for the validation and test dataset, which is much larger than the 2394 training images of the BreastPathQ dataset. Additionally, both problems present a lot of similarities in terms of image patches. The second step consists in a regression task to predict the percentage of cancer cells in the BreastPathQ dataset by using the pre-trained ResNet-18. Furthermore, results will be compared by using an already pre-trained ResNet-18 model but in the ImageNet dataset.

5.3.1.1 Lymph Nodes Metastases Classification

The first task of the proposed method consists in using the selected architeture (ResNet-18) for differentiating between metastatic and non-metastatic tissue in the PCam dataset in order to pre-train the network. With this, it is necessary to adjust the network for a classification task and replace the last fully connected layer to contain just 2 neurons and not 1000 neurons. In Figure 5.4 is presented the ResNet-18 adjusted for only 2 classes and with the correct image sizes throughout the layers given the size of the PCam images.

Finally, it was used the Cross-Entropy loss as the loss function to minimize and the Stochastic Gradient Descent as the optimizer for training the model. In Table 5.12 are presented the values of the hyperparameters that were used for training the model. Other attempts with other values

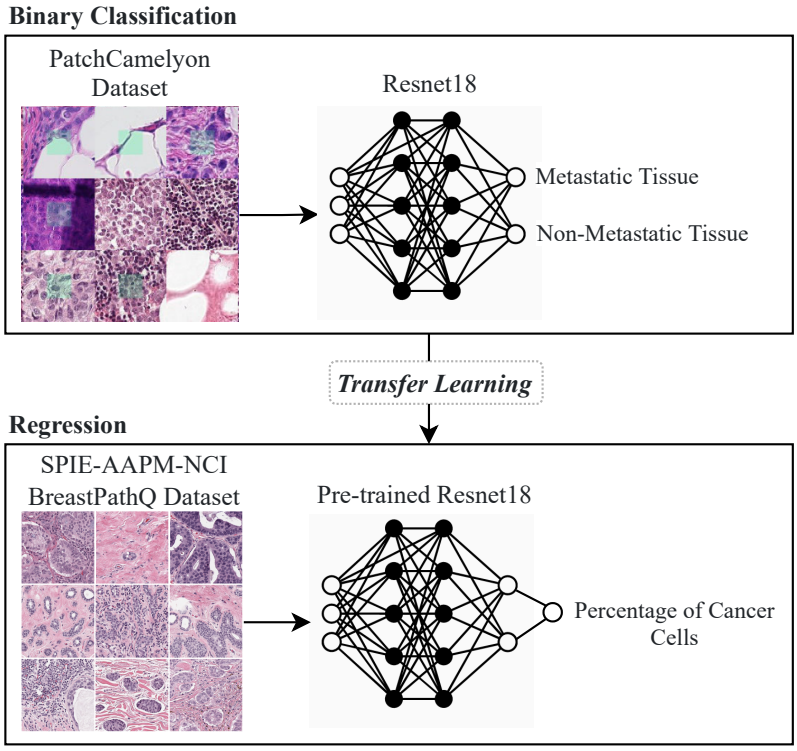


Figure 5.3: Pipeline developed to predict the percentage of cancer cells in the BreastPathQ dataset, based on the pre-trained Resnet18 model using whole-slide images from the PCam dataset.

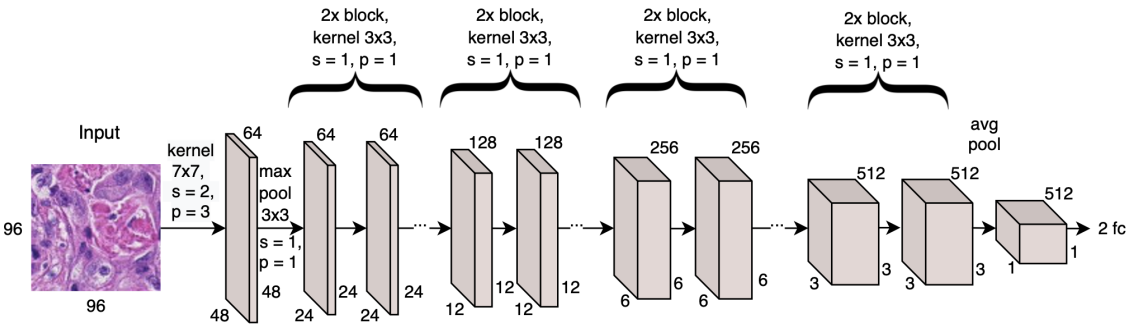


Figure 5.4: Diagram of the ResNet-18 network architeture for images of 96×96 px of the PCam dataset. The letters *s* and *p* stand for stride and padding, respectively.

for the batch size, learning rate and even data augmentation techniques were performed, but since each epoch was very time consuming, due to the large amount of data, for the small schedule of this work it was not possible to perform an exhaustive amount of tests.

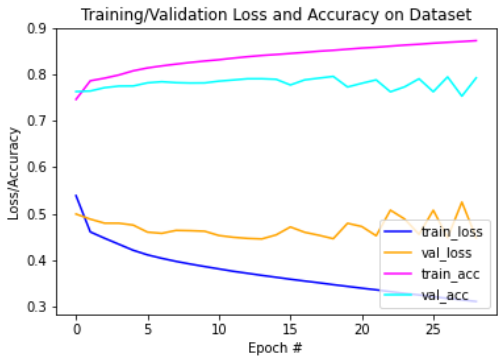
Results and Discussion

As previously mentioned, data augmentation techniques were used in order to reduce overfitting. Due to time constraints, it was only possible to test the influence of two approaches. The first method consisted in applying vertical flips to an image with a 50% probability. The training

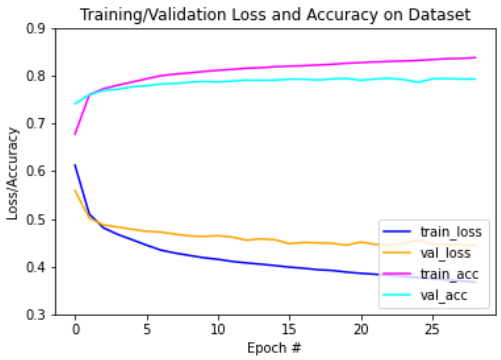
Table 5.12: Selected Hyperparameters for the metastases classification.

Selected Hyperparameters	**Value**
Number of Epochs	70
Learning Rate	0.001
Batch Size	1024

process was run with a learning rate of 0.001 and a batch size of 1024. This process was stopped after 29 epochs, since at that point the validation loss shows an increasing trend. The obtained accuracy and loss on the training and validation dataset is presented in Figure 5.5a.



(a) Only vertical flips as data augmentation.



(b) Not only vertical flips as data augmentation but also random perspectives.

Figure 5.5: Training of the ResNet-18 network for 29 epochs with 2 different approaches regarding data augmentation.

It is possible to observe that the overfitting happens quickly since that the training loss continues to decrease and that the validation loss has stagnated or that it even increases. In other words, it is possible to observe that the distance between the training and validation curves keeps getting larger. The second approach consisted in not only applying vertical flips with a 50% probability but also applying random perspectives with a distortion factor of 0.5 and 50% probability. This was the selected data augmentation technique, as described above, and the model was trained for 70 epochs. However, in Figure 5.5b is presented the training phase for only the first 29 epochs to directly compare with Figure 5.5a. It is possible to observe that with this approach the overfitting reduced significantly and that the validation accuracy at epoch 29 is still increasing, which was not the case in Figure 5.5a. Although these results are better, they could probably be further improved by applying some color normalization technique throughout the dataset and by diminishing the learning rate when no significant improvement is seen in the validation accuracy after several epochs.

Furthermore, in Figure 5.6 is presented the selected metrics to evaluate the model during the training of the network, which correspond to the sensitivity, precision, accuracy and AUC. It is

possible to observe that the behaviour of the precision, accuracy and AUC curves are much more stable than the sensitivity curve, which suggests that there is a lot of false negatives, i.e. the network misclassifies patches with metastases as not having any. Additionally, the oscillation of the curves suggests that the learning rate might have been too high and that the training could have benefitted from a lower learning rate. Table 5.13 was created by searching for the maximum value in each curve and obtaining the corresponding epoch. Then, for each epoch, the metrics were gathered from all curves and rounded to 2 decimal places. From Table 5.13 is possible to observe that sensitivity is the metric with the lowest values of them all and that epoch 64 results in the epoch with the most maximum values for all metrics. Therefore, the chosen model that will be used for pre-training and in the second task of this work will be the ResNet-18 network trained for 64 epochs. In this case, the model is capable to classify lymph node metastases with a sensitivity of 0.70, a precision of 0.85, an accuracy of 0.79 and an AUC of 0.88.

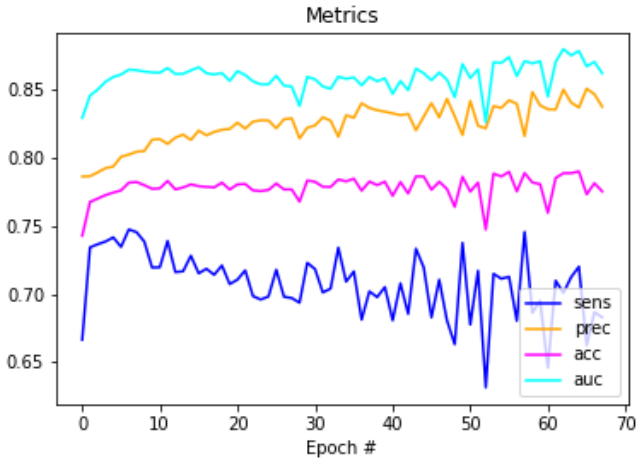
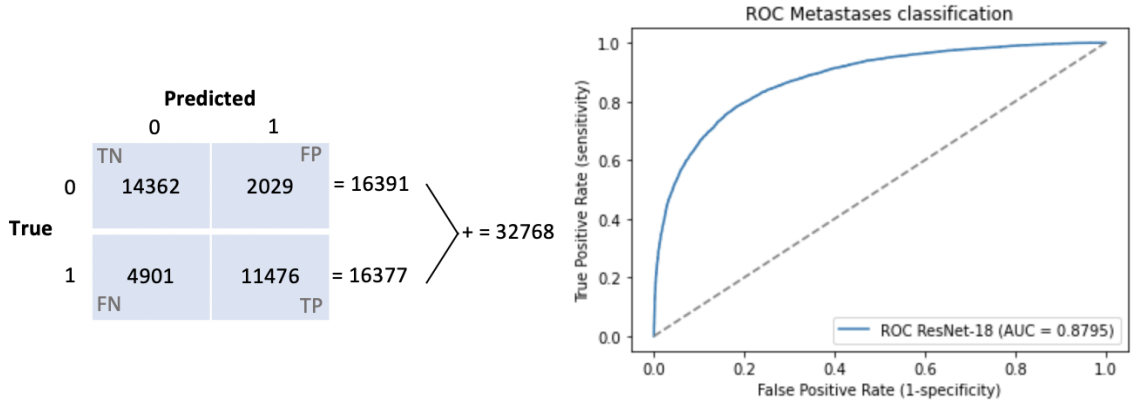


Figure 5.6: Obtained metrics for the pre-training of the ResNet-18 for 70 epochs.

Table 5.13: Evaluation metrics for the selected epochs.

Epoch	Sensitivity	Precision	Accuracy	AUC
8	**0.75**	0.8	0.78	0.86
64	0.70	**0.85**	**0.79**	**0.88**
66	0.72	0.84	**0.79**	**0.88**
67	0.66	**0.85**	0.77	0.87

Additionally, in Figure 5.7a is presented the confusion matrix for the model at epoch 64 regarding the test dataset. It is possible to observe that the number of false positives is smaller than the number of false negatives, which is compatible with what was mentioned before regarding the lower values for the sensitivity. Figure 5.7b refers to the ROC curve on the test dataset for the lymph nodes metastases classification at epoch 64.



(a) Confusion matrix.

(b) ROC curve for lymph node metastases classification. The dashed line corresponds to the ROC curve of a random classifier.

Figure 5.7: Confusion matrix and ROC curve for the ResNet-18 network trained for 64 epochs on the test dataset.

The PCam dataset was presented in [71], where the authors proposed a model to classify the image patches as healthy tissue or tumor. They reported an accuracy of 89.8 and an AUC of 96.3. The results obtained in [71] are significantly better than the ones obtained in this work, which can be explained by two factors. Firstly, the network used was not the same, since the authors explored the fact that the WSI data exhibits properties of translation, rotation and reflection symmetry. Therefore, they implemented CNNs that are equivariant to 90º rotations and reflections. Moreover, due to the large size of the PCam dataset, the hardware and time constraints did not allow the testing of different data augmentation techniques and the training of our network for more than 70 epochs, which may not be sufficient.

5.3.2 Tumor Cells Prediction

In order to investigate the usage of the pre-training with the PCam dataset, five different cases were considered, which correspond to codes 000, 002, 010, 012 and 100. These cases were selected because four of them exhibit the higher values of the ICC evaluation metric in Table 5.7, and the other one was included to assess the effect of having images with zero cellularity in the training dataset. Note that the ICC metric translates the correlation between the estimations made by the model with the predictions made by the pathologist better than the other two evaluation metrics considered in this work.

Table 5.14 summarizes all the tests that were performed. The highest value of each one of the six metrics across all cases is highlighted in bold. Unlike what was obtained with the ResNet-18 trained from scratch or pre-trained with the ImageNet, where the highest value of all the evaluation metrics occurred for the same case (000), the highest value of the different metrics does not occur for a single case. Furthermore, for the tests corresponding to the training dataset without zeros,

the value of the evaluation metrics is quite constant across these cases, suggesting once again that the effect of classifiers and color normalization are not very relevant.

The results presented in Table 5.14 are in general slightly worst than the corresponding results in Table 5.7, indicating that the model performs better when it is pre-trained with ImageNet.

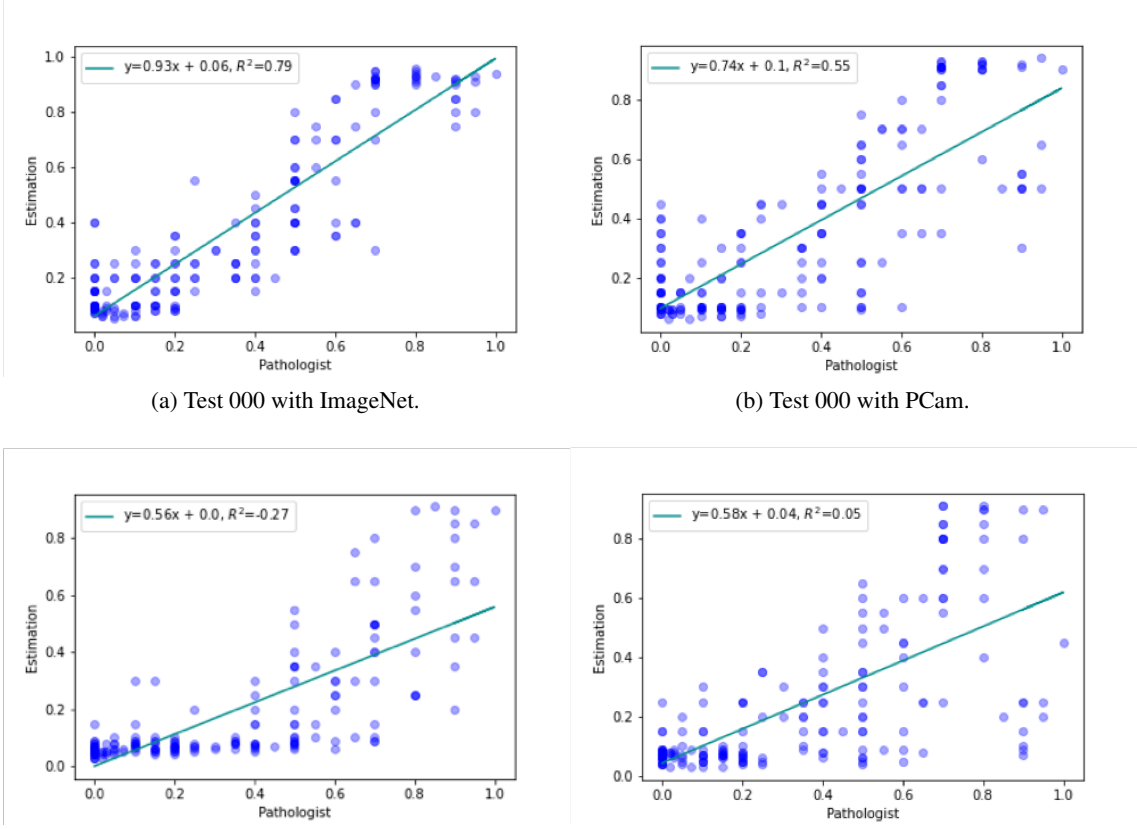
It is also interesting to compare the estimations made by the models pre-trained with the ImageNet and with the PCam dataset, which are depicted in Figure 5.8 for cases 000 (top row) and 100 (bottom row). As expected, the 100 case exhibits a biasing in the estimation towards the class zero in both models. Moreover, the estimations made by the model pre-trained with the PCam (right column) show a visible gap in the scatter plot, as if the model was trying to classify the images as 0 or 1. This may be explained by the fact that the model was pre-trained on a dataset where it only learned how to distinguish between healthy tissue and tumor.

Table 5.15 compares the evaluation metrics at the minimum loss epoch for the three approaches used for the training of the model. It is possible to observe that the results for the model trained from scratch are slightly better than those of the network pre-trained on the PCam dataset, and that the network pre-trained on the ImageNet dataset outperforms the other two. Furthermore, it should be mentioned that for case 100, the ICC value for the ResNet-18 trained from scratch (0.701) is higher when compared to the one obtained with the pre-training on the ImageNet dataset (0.619) and the PCam dataset (0.630).

Lastly, Figure 5.9 presents the estimations made by the models for case 100. It is interesting to notice that the predictions made by the model trained from scratch are somewhat similar to the ones made by the model pre-trained on the PCam dataset. Indeed, unlike Figure 5.9a, both Figures 5.9b and 5.9c exhibit a notorious gap in the scatter plots. Although only the 100 case is represented, this behaviour is also present in the other studied cases. This suggests that the reason behind the existence of this gap is not the one previously pointed (the pre-training with the PCam favoured the estimation as being healthy or tumor, that is, 0 or 1), but instead it may be associated with an insufficient training of the network.

Table 5.14: Summary of the Results for the pre-trained ResNet-18 on the PCam Dataset. For detailed plots of the evolution of the training and validation loss, evaluation metrics and scatter plots see Appendix A.

Test	**Epochs**	**LR**	**Best Epoch**	**Val Loss**	τ_b	P_k	**ICC**	τ_b **max**	P_k **max**	**ICC max**	**y = mx + b**	R^2
000	700	0.006	640	0.030	0.558	0.783	0.788	0.566	0.787	0.792	y = 0.74x + 0.1	0.55
002	400	0.01	118	0.032	0.564	0.787	0.775	**0.607**	0.797	0.775	y = 0.74x + 0.14	0.51
010	1100	0.01	1006	0.028	0.576	0.792	**0.789**	0.584	0.797	**0.806**	y = 0.75x + 0.06	0.55
012	400	0.01	209	0.032	**0.583**	**0.797**	0.787	0.602	**0.801**	0.787	y = 0.77x + 0.14	0.54
1000	850	0.003	768	0.035	0.499	0.753	0.630	0.507	0.757	0.630	y = 0.58x + 0.04	0.05



(a) Test 000 with ImageNet.

(b) Test 000 with PCam.

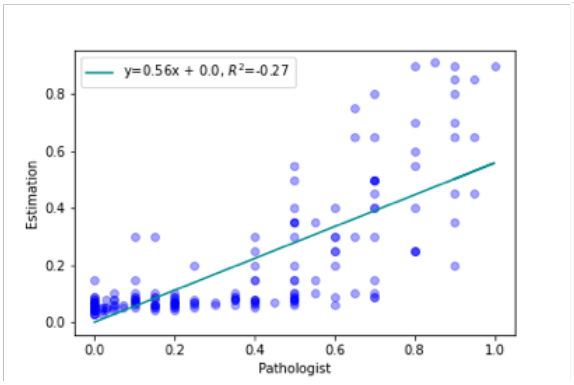
(c) Test 100 with ImageNet.

(d) Test 100 with PCam.

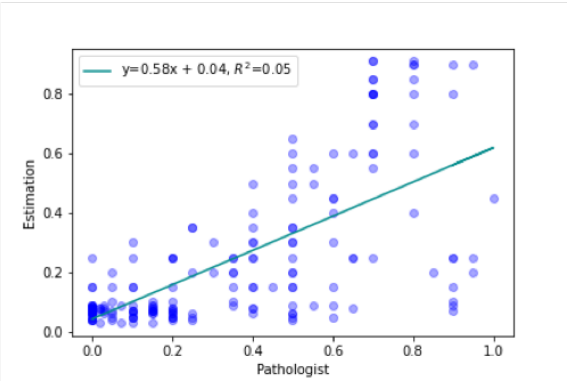
Figure 5.8: Scatter Plot showing the level of agreement between the estimations of the model with the predictions of the pathologist on the validation dataset for the case 000 and 100.

Table 5.15: Comparison of the evaluation metrics for the three approaches used for the training of the model.

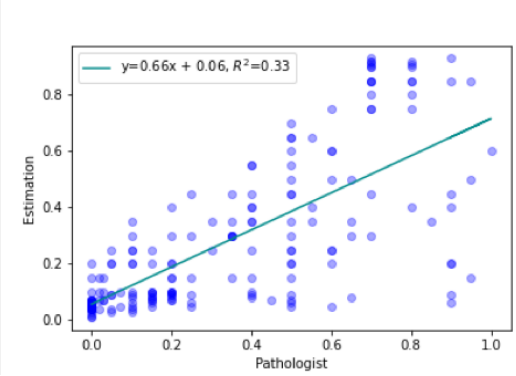
Training Type	**Test**	τ_b	P_k	**ICC**
Trained from Scratch	000	0.640	0.825	0.838
	002	0.567	0.788	0.768
	010	0.636	0.823	0.830
	012	0.559	0.785	0.777
	100	0.585	0.799	0.701
Pre-Training with ImageNet	000	0.703	0.858	0.895
	002	0.628	0.820	0.878
	010	0.678	0.845	0.890
	012	0.630	0.821	0.876
	100	0.617	0.811	0.619
Pre-Training with PCam	000	0.558	0.783	0.788
	002	0.564	0.787	0.775
	010	0.576	0.792	0.789
	012	0.583	0.797	0.787
	100	0.499	0.753	0.630



(a) Test 100 with ImageNet.



(b) Test 100 with PCam.



(c) Test 100 from scratch.

Figure 5.9: Scatter Plot showing the level of agreement between the estimations of the model with the predictions of the pathologist on the validation dataset for the case 100.

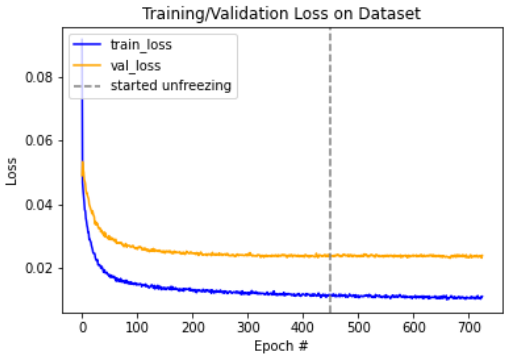
5.4 Unfreezing the Layers of the ResNet-18

Previous studies have shown that progressively unfreezing the layers of the network is a good approach for fine-tuning models [76]. In an attempt to improve the results that were previously obtained when the ResNet-18 was pre-trained with the ImageNet and the PCam dataset, a preliminary study on the influence of this fine-tuning approach is also conducted in this work. For that, only the case 000 was considered. Ideally, this process would start from the point corresponding to the minimum validation loss epoch, but since the state of the model was saved every 50 epochs and not at that point, it was decided that the unfreezing would start at the nearest saved checkpoint. Each time a certain layer was unfreezed, the model was trained for 25 epochs before unfreezing the next one. Furthermore, at every 25 epochs the learning rate was decreased by a factor of 1.1[1], with the initial value being the one of the saved model.

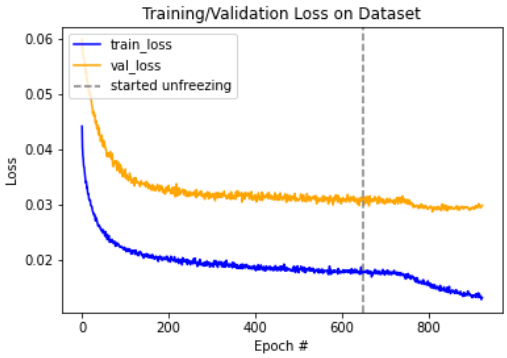
Figure 5.10 presents the loss curves and the evaluation metrics for test 000 for the pre-trained ResNet-18 with the ImageNet (left) and the PCam (right) datasets when unfreezing the layers. The start of the process of unfreezing is represented by the dashed vertical lines. It is possible to observe that in Figure 5.10b there is a clear decrease in the training and validation loss after the process of unfreezing the layers started. On the other hand, Figure 5.10a presents no significant change in the behaviour of the loss curves, suggesting it had no effect for the case where the network was pre-trained with the ImageNet dataset. As expected, the evaluation metrics for the model pre-trained with ImageNet do not show a visible modification after unfreezing, whereas the evaluation metrics for the model pre-trained with PCam slightly increase. In effect, for the model pre-trained with PCam without unfreezing, the value of τ_b was 0.558, P_k was 0.783 and ICC was 0.788 and for the model pre-trained with PCam with unfreezing the value of τ_b was 0.565, P_k was 0.786 and ICC was 0.795 (values at the minimum validation loss epoch).

In this preliminary analysis, the unfreezing of the layers of the network showed promising results in fine-tuning the model. Nonetheless, a more thorough investigation is required to fully understand the potential of this approach.

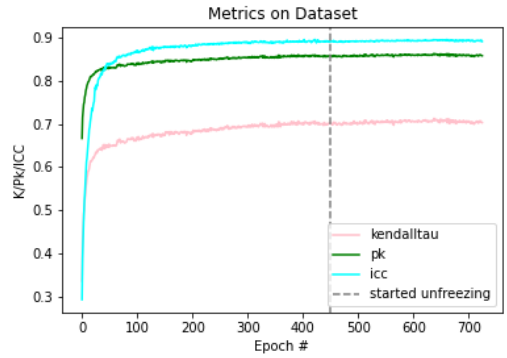
[1] In some preliminary analyses, it was found that this value performed the best.



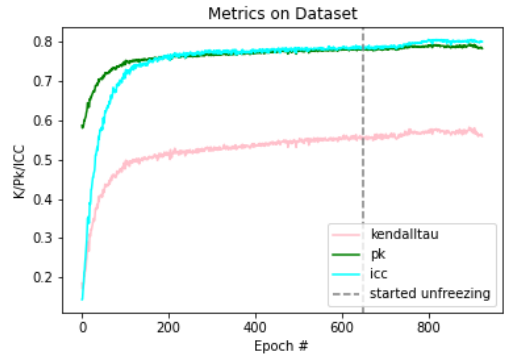
(a) Loss Curves for the Pre-Trained ResNet-18 with ImageNet.



(b) Loss Curves for the Pre-Trained ResNet-18 with PCam.



(c) Metrics for the Pre-Trained ResNet-18 with ImageNet.



(d) Metrics for the Pre-Trained ResNet-18 with PCam.

Figure 5.10: Training and Validation Curves (top) and Evaluation Metrics on the Validation Dataset (bottom) for the ResNet-18 pre-trained on the ImageNet (left) and the PCam (right) datasets when unfreezing the layers.

5.5 Summary

Several tests were performed in order to investigate the influence of different factors in the performance of a ResNet-18: transfer learning, presence of images with zero patches in the training dataset, classifier at the end of the network and usage of color normalization techniques. The effect of transfer learning is investigated by comparing the performance of the network trained from scratch with that of the network pre-trained with a general dataset (ImageNet) or with a context specific dataset (PCam). The influence of the classifier is addressed by considering two different classifiiers: one from 512 neurons to 1 with a sigmoid at the end, and another that consists in 512 neurons to 64 to 1 neuron with a sigmoid at the end. Finally, two different color normalization approaches were implemented, Macenko *et al.* [65] and Vahadane *et al.* [50], and the results were compared to the case where no normalization was used.

In order to pre-train the network on the PCam dataset, the necessary adjustments to the ResNet-18 are described. A small study concerning two different data augmentation approaches is done and the usage of random vertical flips and random perspectives is selected. Furthermore, the selected model corresponds to epoch 64, which presents the most maximum values for all evaluation metrics. The pre-trained model on the PCam dataset obtained a sensitivity of 0.70, a precision of 0.85, an accuracy of 0.79 and an AUC of 0.88. These results might have been further improved, but it was not possible to train the network for more than 70 epochs due to time constraints, which may have not been sufficient.

Firstly, the removal of the images with zero cellularity from the training dataset resulted in a higher overall performance for the three networks, hinting at the importance of removing these patches prior to training. Also, for the case which corresponds to having the training dataset with zero patches, the biasing towards the zero class is evident. The two selected classifiers considered presented similar results, indicating that the addition of the second layer does not improve the predictions made by the three models. Then, the effect of the color normalization was studied, and it was found that in general the best results were obtained for the cases where no color normalization technique was used. This could be associated to the choice of the target image in the color normalization techniques, which may not have been ideal.

Overall, the results with the network pre-trained on the PCam dataset are slightly worst than for the case where the network is trained from scratch, and the best results were obtained with the network that was pre-trained with the ImageNet. Furthermore, a visual gap is also visible in the scatter plots corresponding to both the newtwork trained from scratch and pre-trained with the PCam, which might be linked to a non-existent or insufficient pre-training of the network.

To sum up, a fine-tuning approach based on unfreezing the layers of the pre-trained ResNet-18 was investigated. The preliminary results obtained showed that there was an improvement of the evaluation metrics for the case where the network was pre-trained on the PCam dataset, whereas no significant improvement was observed for the pre-trained ResNet-18 on the ImageNet.

Chapter 6

Conclusions and Future Work

6.1 Main Conclusions

Cancer is one of the main causes of death all over the world. Early detection increases considerably the survival rate of the cancer patients. Nowadays, automated methods used for the detection of tumor cells already show performances comparable to those obtained by highly trained pathologists. Among these methods, the ones based on deep learning techniques are the most promising ones.

This dissertation addresses the development of a methodology based on ML to predict the tumor cells percentage in WSI data of breast cancer in order to assist pathologists. The literature review already conducted has shown that most of the algorithms used so far rely on patch-based DL techniques. CNN algorithms are the DL methods most suitable for this kind of application.

For this work, the influence of three different factors on the cancer cells quantification done by a ResNet-18 was investigated. One of these factors is directly related to the CNN, i.e., it concerns the classifier at the end of the CNN. The other two factors are associated with the dataset used to train and test the developed methodology. One factor is the removal of the images with zero cellularity from the training dataset, and the other is the usage of color normalization techniques. Furthermore, in this work, the use of pre-training was also investigated, as it has been questioned if pre-training with a very general dataset is an effective way of improving the performance of the networks.

In order to study the effect of the classifier at the end of the network, two different scenarios were considered. One of them with one fully connected layer, from 512 neurons to 1 with a sigmoid at the end, and the other one with two fully connected layers, from 512 neurons to 64 to 1 with a sigmoid at the end. The results obtained with the two different classifiers are pretty much identical, hinting that a classifier with more layers may not be necessary in order to improve the model performance.

In this work, the dataset used to train and validate the results is the BreastPathQ dataset. This dataset contains images from 0% to 100% cellularity and is unbalanced, having a large number of patches of the class zero. In order to investigate the influence of these images, two different

settings were considered. The first one consists in the removal of the images with zero cellularity from the training dataset, and in the other one the whole training dataset is used. From the obtained results, the value of the evaluation metrics is significantly higher for the cases where the training dataset that was used did not contain images with zero cellularity. Furthermore, the scatter plots clearly show that there is a bias towards the zero class when the training data contains the images with zero cellularity. Therefore, the obtained results demonstrate the importance of removing these patches from the training data.

The effect of using color normalization techniques was investigated by comparing three different cases, more specifically the usage of two distinct color normalization approaches, the Macenko [65] and the Vahadane [50] approach, with that of not using any color normalization technique. The results obtained showed that when using the training dataset without the zero patches, the color normalization did not improve the results, and in fact the higher values of the evaluation metrics are for the cases where no color normalization is used.

In addition, the effect of pre-training the network was addressed by comparing the results obtained in three different scenarios. The first consisted in using a ResNet-18 without pre-training, that is, the network is trained from scratch. The second involved using the same network architecture and pre-training it with a general dataset, the Imagenet. The third scenario comprises the ResNet-18 pre-trained with a context specific dataset, the PCam dataset. Surprisingly, the results obtained when using the network trained from scratch are slightly better than when using the network pre-trained on the PCam dataset. Nevertheless, they are worst than with the ImageNet, which outperformed the other two cases. The fact that the network pre-trained with the ImageNet provides better results points out to the relevance of transfer learning. In effect, although the ImageNet dataset contains general images, it is a very large dataset and the pre-training of the ResNet-18 was done for an adequate amount of epochs. On the contrary, even though the pre-training with the PCam dataset used context specific images, the pre-training procedure did not occur for an extensive amount of epochs due to hardware and time constraints. As a result, the network that was pre-trained is not sufficiently pre-trained. It is interesting to refer that when comparing the scatter plots, there are similar visual gaps for the plots of the network pre-trained with the PCam and the one that was not pre-trained, clearly suggesting that the pre-training with the PCam dataset was not performed for a sufficient number of epochs.

Finally, a fine-tuning technique based on unfreezing the weights of the pre-trained ResNet-18 was explored. This approach was adopted in an attempt of improving the results obtained with the network pre-trained with the Imagenet and the PCam datasets. It was observed that while the unfreezing improved the performance of the model pre-trained with the PCam, it did not show a significant influence in the results for the model pre-trained with the ImageNet.

6.2 Future Work

As already discussed, the pre-training with the PCam dataset was insufficient and, therefore, it was not possible to assess the relevance of pre-training the network with a context specific dataset.

Since this is an important aspect, it is proposed as future work to conduct a longer pre-training of the ResNet-18 with the PCam dataset. Concerning this pre-training, it would also be interesting to further investigate the usage of more data augmentation techniques and the influence of using different hyperparameters, such as the batch size and the learning rate. Furthermore, it would be worth analyzing the impact of a pre-training with a context specific dataset in other network architectures.

Additionally, as it was mentioned before, the removal of the images with zero cellularity from the training dataset improves the performance of the model. These patches were removed manually prior to training. Therefore, for future work, it would be interesting to develop a pipeline where the first step would be a CNN to classify images as 0 or 1, that is, in order to separate the zero patches from the rest of the images and classify them in advance as healthy tissue. Then, the second step would be to use a CNN for regression and classify the remaining image patches in a scale of 0% to 100%, which was already implemented.

In the preliminary analysis of the fine-tuning technique based on unfreezing the layers of the network, promising results were obtained. Nevertheless, a more thorough investigation is required to fully understand the potential of this approach, namely by studying: (i) the influence of the learning rate decay, (ii) the number of epochs necessary to train the model after unfreezing a layer, (iii) the layers that have more impact in the performance of the model, etc.

Appendix A

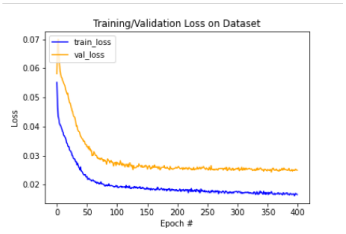
Cancer Cells Quantification - Detailed Results

This Appendix contains the detailed plots for all the tests referred in this dissertation. For each case, three plots are presented:

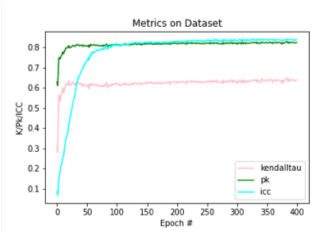
- training and validation loss as function of the number of epochs;
- evaluation metrics as a function of the number of epochs;
- scatter plot showing the relation between the estimations made by the model and the prediction of the pathologist.

A.1 ResNet-18 trained from Scratch

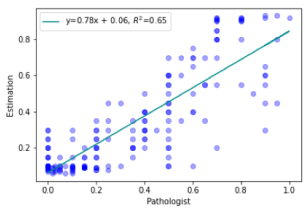
Test 000 - learning rate of 0.001



(a) Training and Validation Curves.

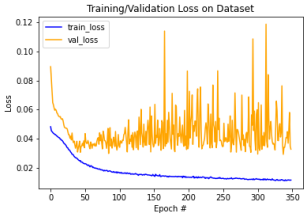


(b) Metrics for the Validation Dataset.

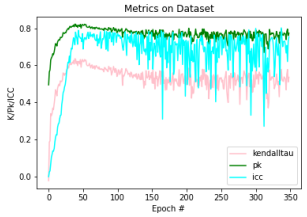


(c) Scatter Plot.

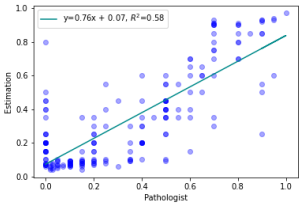
Test 001 - learning rate of 0.004



(a) Training and Validation Curves.

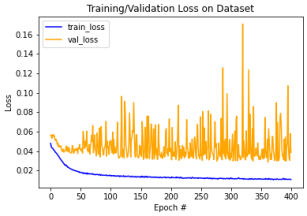


(b) Metrics for the Validation Dataset.

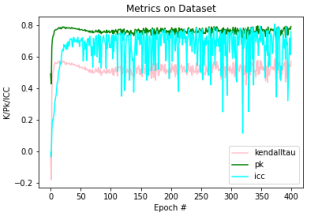


(c) Scatter Plot.

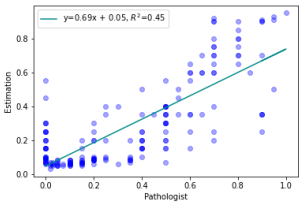
Test 002 - learning rate of 0.004



(a) Training and Validation Curves.



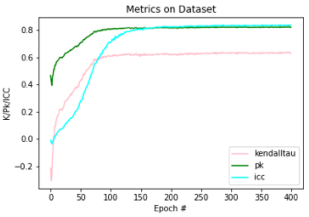
(b) Metrics for the Validation Dataset.



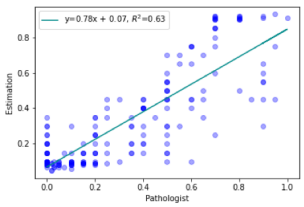
(c) Scatter Plot.

Test 010 - learning rate of 0.001

(a) Training and Validation Curves.



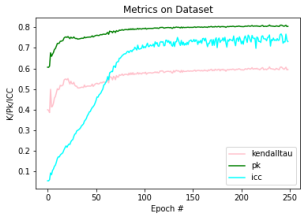
(b) Metrics for the Validation Dataset.



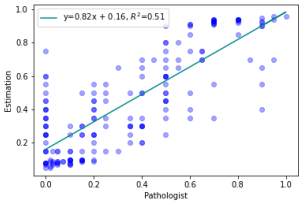
(c) Scatter Plot.

Test 011 - learning rate of 0.002

(a) Training and Validation Curves.

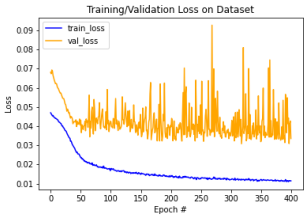


(b) Metrics for the Validation Dataset.

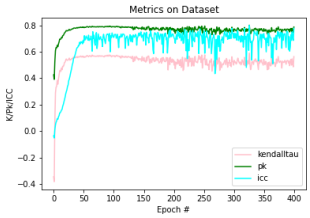


(c) Scatter Plot.

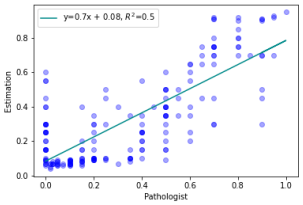
Test 012 - learning rate of 0.004



(a) Training and Validation Curves.



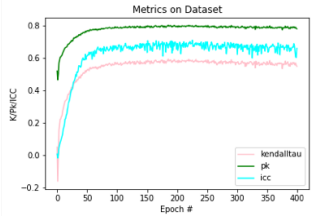
(b) Metrics for the Validation Dataset.



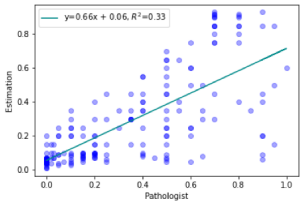
(c) Scatter Plot.

Test 100 - learning rate of 0.001

(a) Training and Validation Curves.

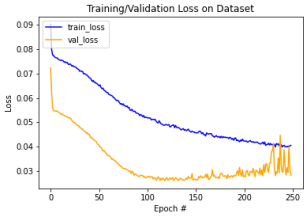


(b) Metrics for the Validation Dataset.

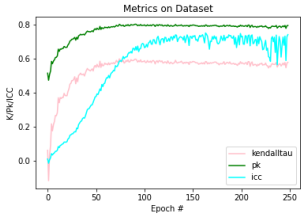


(c) Scatter Plot.

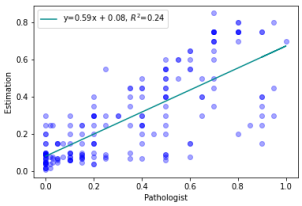
Test 101 - learning rate of 0.001



(a) Training and Validation Curves.

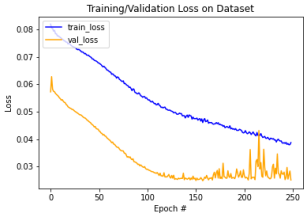


(b) Metrics for the Validation Dataset.

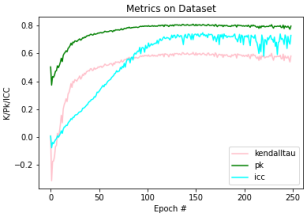


(c) Scatter Plot.

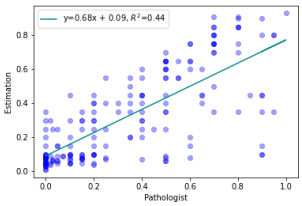
Test 102 - learning rate of 0.001



(a) Training and Validation Curves.

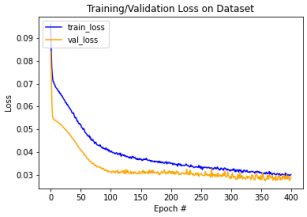


(b) Metrics for the Validation Dataset.

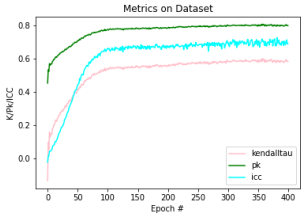


(c) Scatter Plot.

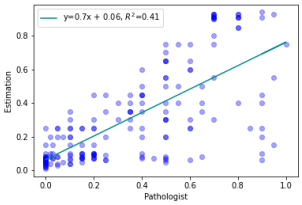
Test 110 - learning rate of 0.001



(a) Training and Validation Curves.

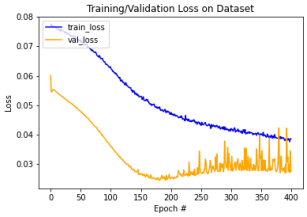


(b) Metrics for the Validation Dataset.

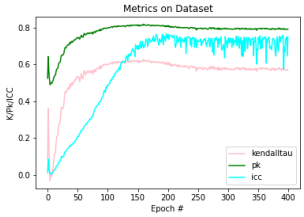


(c) Scatter Plot.

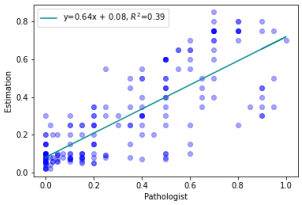
Test 111 - learning rate of 0.001



(a) Training and Validation Curves.

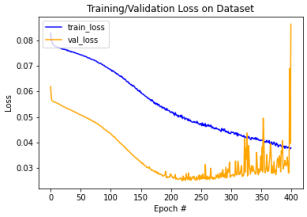


(b) Metrics for the Validation Dataset.

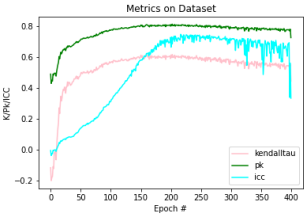


(c) Scatter Plot.

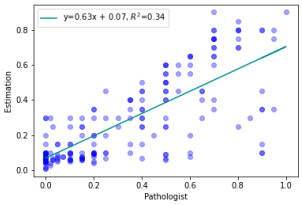
Test 112 - learning rate of 0.001



(a) Training and Validation Curves.



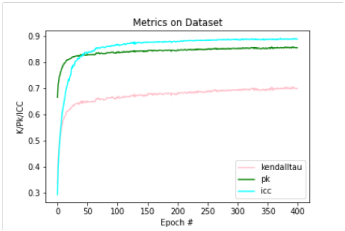
(b) Metrics for the Validation Dataset.



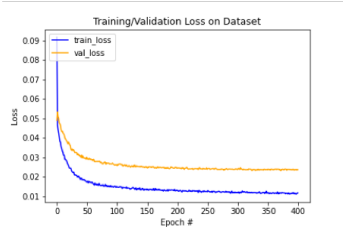
(c) Scatter Plot.

A.2 ResNet-18 pre-trained on ImageNet

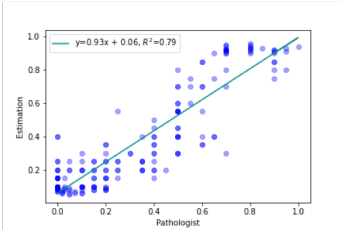
Test 000 - learning rate of 0.003



(a) Training and Validation Curves.

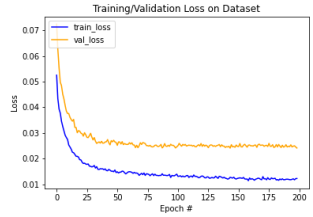


(b) Metrics for the Validation Dataset.

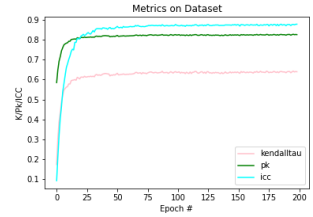


(c) Scatter Plot.

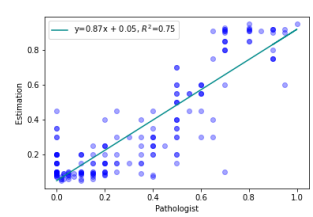
Test 001 - learning rate of 0.006



(a) Training and Validation Curves.



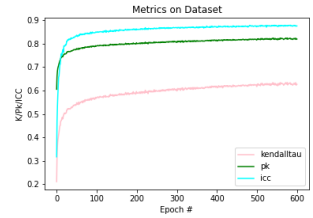
(b) Metrics for the Validation Dataset.



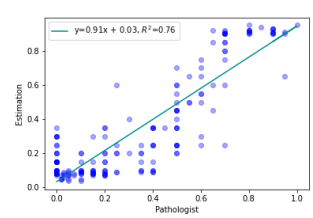
(c) Scatter Plot.

Test 002 - learning rate of 0.006

(a) Training and Validation Curves.

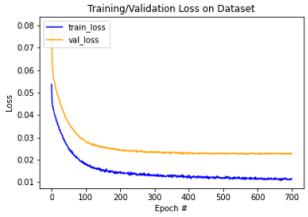


(b) Metrics for the Validation Dataset.

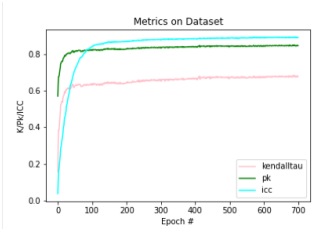


(c) Scatter Plot.

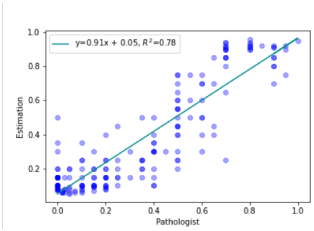
Test 010 - learning rate of 0.003



(a) Training and Validation Curves.

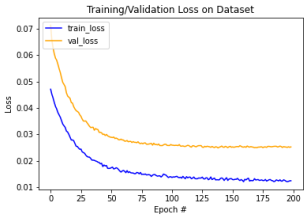


(b) Metrics for the Validation Dataset.

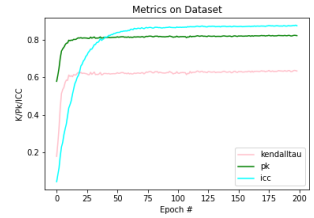


(c) Scatter Plot.

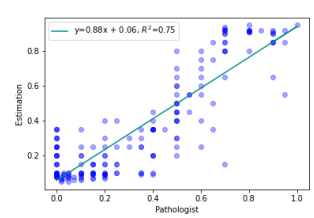
Test 011 - learning rate of 0.006



(a) Training and Validation Curves.

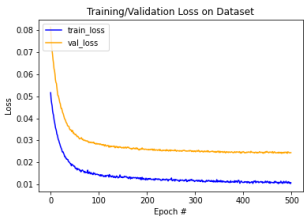


(b) Metrics for the Validation Dataset.

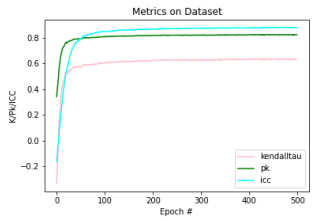


(c) Scatter Plot.

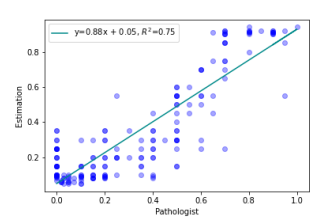
Test 012 - learning rate of 0.006



(a) Training and Validation Curves.



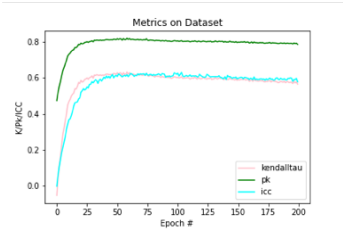
(b) Metrics for the Validation Dataset.



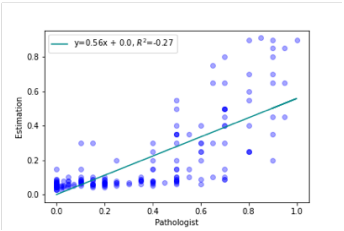
(c) Scatter Plot.

Test 100 - learning rate of 0.003

(a) Training and Validation Curves.

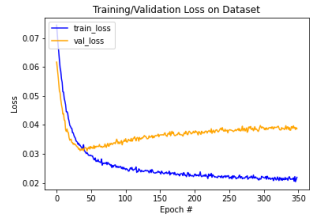


(b) Metrics for the Validation Dataset.

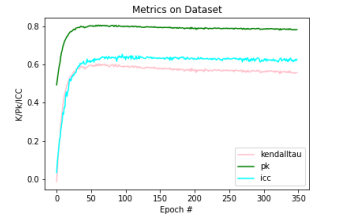


(c) Scatter Plot.

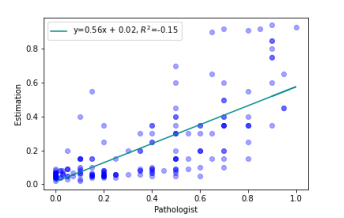
Test 101 - learning rate of 0.003



(a) Training and Validation Curves.



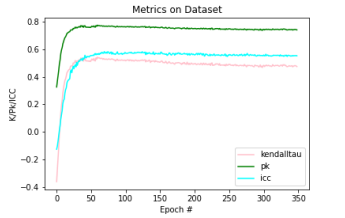
(b) Metrics for the Validation Dataset.



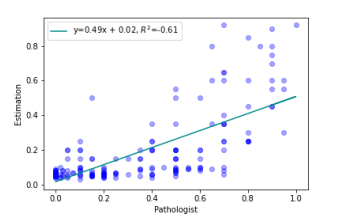
(c) Scatter Plot.

Test 102 - learning rate of 0.003

(a) Training and Validation Curves.

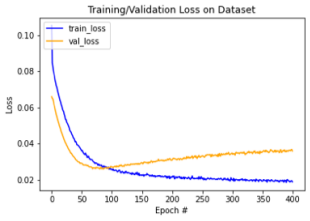


(b) Metrics for the Validation Dataset.

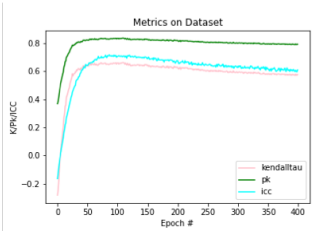


(c) Scatter Plot.

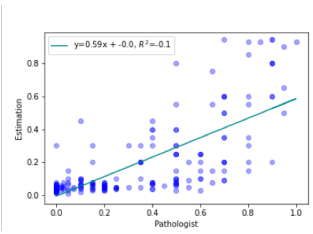
Test 110 - learning rate of 0.003



(a) Training and Validation Curves.

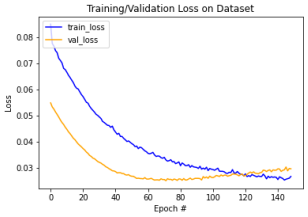


(b) Metrics for the Validation Dataset.

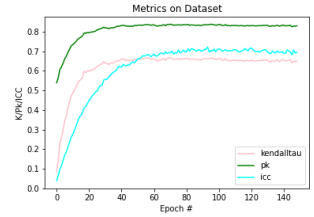


(c) Scatter Plot.

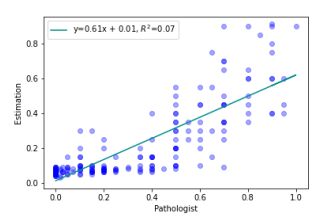
Test 111 - learning rate of 0.003



(a) Training and Validation Curves.

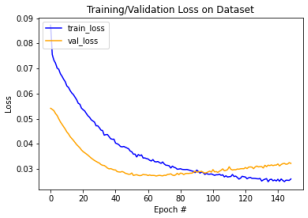


(b) Metrics for the Validation Dataset.

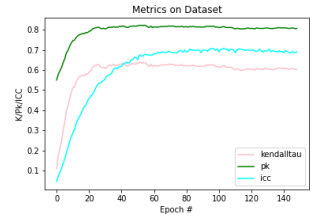


(c) Scatter Plot.

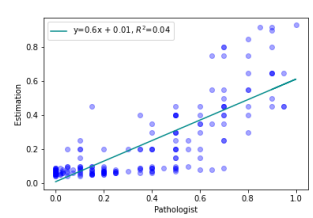
Test 112 - learning rate of 0.003



(a) Training and Validation Curves.



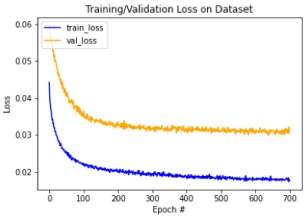
(b) Metrics for the Validation Dataset.



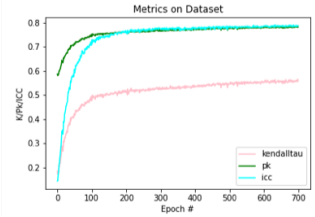
(c) Scatter Plot.

A.3 ResNet-18 pre-trained on PCam

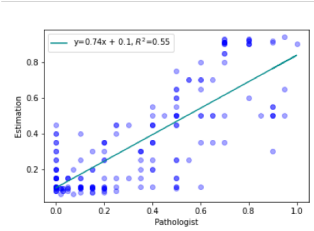
Test 000 - learning rate of 0.006



(a) Training and Validation Curves.



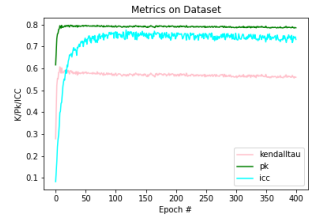
(b) Metrics for the Validation Dataset.



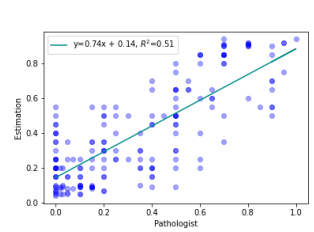
(c) Scatter Plot.

Test 002 - learning rate of 0.01

(a) Training and Validation Curves.

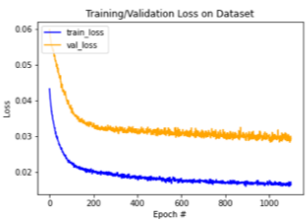


(b) Metrics for the Validation Dataset.

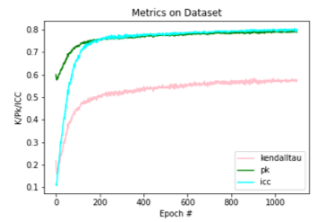


(c) Scatter Plot.

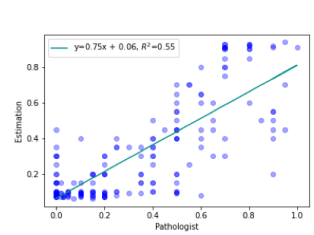
Test 010 - learning rate of 0.01



(a) Training and Validation Curves.



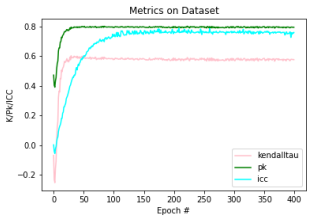
(b) Metrics for the Validation Dataset.



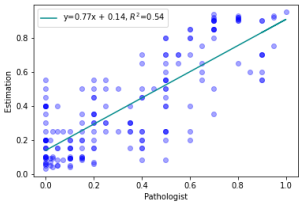
(c) Scatter Plot.

Test 012 - learning rate of 0.01

(a) Training and Validation Curves.

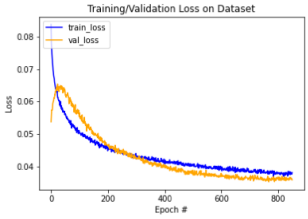


(b) Metrics for the Validation Dataset.

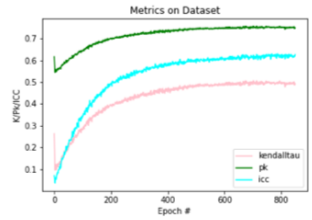


(c) Scatter Plot.

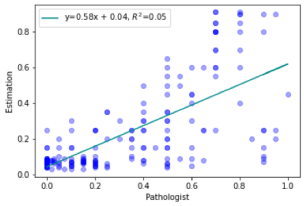
Test 100 - learning rate of 0.003



(a) Training and Validation Curves.



(b) Metrics for the Validation Dataset.



(c) Scatter Plot.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer Journal for Clinicians*, 71(3):209–249, 2021.

[2] B. Lhermitte, C. Egele, N. Weingertner, D. Ambrosetti, B. Dadone, V. Kubiniek, F. Burel-Vandenbos, J. Coyne, J. F. Michiels, M. P. Chenard, E. Rouleau, J. C. Sabourin, and J. P. Bellocq. Adequately defining tumor cell proportion in tissue samples for molecular testing improves interobserver reproducibility of its assessment. *Virchows Archiv*, 470(1):21–27, 2017.

[3] A. J. J. Smits, J. A. Kummer, P. C. De Bruin, M. Bol, J. G. Van Den Tweel, K. A. Seldenrijk, S. M. Willems, G. J. A. Offerhaus, R. A. De Weger, P. J. Van Diest, and A. Vink. The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. *Modern Pathology*, 27(2):168–174, 2014.

[4] P. W. Hamilton, Y. Wang, C. Boyd, J. A. James, M. B. Loughrey, J. P. Hougton, D. P. Boyle, P. Kelly, P. Perry Maxwell, D. David McCleary, J. Diamond, D. G. McArt, J. Tunstall, P. Bankhead, and M. Salto-Tellez. Automated tumor analysis for molecular profiling in lung cancer. *Oncotarget*, 6(29):27938–27952, 2015.

[5] F. Aeffner, K. Wilson, N. T. Martin, J. C. Black, C. L. L. Hendriks, B. Bolon, D. G. Rudmann, R. Gianani, S. R. Koegler, J. Krueger, and G. D. Young. The gold standard paradox in digital image analysis: Manual versus automated scoring as ground truth. *Archives of Pathology and Laboratory Medicine*, 141(9):1267–1275, 2017.

[6] K. Dufraing, J. H. van Krieken, G. De Hertogh, G. Hoefler, A. Oniscu, T. P. Kuhlmann, W. Weichert, C. Marchiò, A. Ristimäki, A. Ryška, J. Y. Scoazec, and E. Dequeker. Neoplastic cell percentage estimation in tissue samples for molecular oncology: recommendations from a modified delphi study. *Histopathology*, 75(3):312–319, 2019.

[7] F. Aeffner, M. D. Zarella, N. Buchbinder, M. M. Bui, M. R. Goodman, D. J. Hartman, G. M. Lujan, M. A. Molani, A. V. Parwani, K. Lillard, O. C. Turner, V. N. P. Vemuri, A. G. Yuil-Valdes, and D. Bowman. Introduction to digital image analysis in whole-slide imaging: A white paper from the digital pathology association. *Journal of Pathology Informatics*, 10(1), 2019.

[8] P. W. Hamilton, P. Bankhead, Y. Wang, R. Hutchinson, D. Kieran, D. G. McArt, J. James, and M. Salto-Tellez. Digital pathology and image analysis in tissue biomarker research. *Methods*, 70(1):59–73, 2014.

[9] Z. Pei, S. Cao, L. Lu, and W. Chen. Direct cellularity estimation on breast cancer histopathology images using transfer learning. *Computational and Mathematical Methods in Medicine*, 2019, 2019.

[10] T. Araujo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polonia, and A. Campilho. Classification of breast cancer histology images using convolutional neural networks. *PLoS ONE*, 12(6), 2017.

[11] C. Greene, E. O'Doherty, F. Abdullahi Sidi, V. Bingham, N. C. Fisher, M. P. Humphries, S. G. Craig, L. Harewood, S. McQuaid, C. Lewis, and J. James. The potential of digital image analysis to determine tumor cell content in biobanked formalin-fixed, paraffin-embedded tissue samples. *Biopreservation and Biobanking*, 19(4):324–331, 2021.

[12] Jon C. Aster Vinay Kumar, Abul K. Abbas. *Robbins Basic Pathology*. Robbins Pathology. Elsevier, 10 edition, 2017.

[13] L. He, L. R. Long, S. Antani, and G. R. Thoma. Histology image analysis for carcinoma detection and grading. *Computer Methods and Programs in Biomedicine*, 107(3):538–556, 2012.

[14] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, E. Tsougenis, Q. Huang, M. Cai, and P. A. Heng. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE Transactions on Cybernetics*, 50(9):3950–3962, 2020.

[15] L. Kim and M. S. Tsao. Tumour tissue sampling for lung cancer management in the era of personalized therapy: What is good enough for molecular testing? *European Respiratory Journal*, 44(4):1011–1022, 2014.

[16] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. N. C. Shih, J. Tomaszewski, F. A. González, and A. Madabhushi. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific Reports*, 7, 2017.

[17] D. Kazdal, E. Rempel, C. Oliveira, M. Allgäuer, A. Harms, K. Singer, E. Kohlwes, S. Ormanns, L. Fink, J. Kriegsmann, M. Leichsenring, K. Kriegsmann, F. Stögbauer, L. Tavernar, J. Leichsenring, A. L. Volckmar, R. Longuespée, H. Winter, M. Eichhorn, C. P. Heußel, F. Herth, P. Christopoulos, M. Reck, T. Muley, W. Weichert, J. Budczies, M. Thomas, S. Peters, A. Warth, P. Schirmacher, A. Stenzinger, and M. Kriegsmann. Conventional and semiautomatic histopathological analysis of tumor cell content for multigene sequencing of lung adenocarcinoma. *Translational Lung Cancer Research*, 10(4):1666–1678, 2021.

[18] Z. Li, J. Zhang, T. Tan, X. Teng, X. Sun, H. Zhao, L. Liu, Y. Xiao, B. Lee, Y. Li, Q. Zhang, S. Sun, Y. Zheng, J. Yan, N. Li, Y. Hong, J. Ko, H. Jung, Y. Liu, Y. C. Chen, C. W. Wang, V. Yurovskiy, P. Maevskikh, V. Khanagha, Y. Jiang, L. Yu, Z. Liu, D. Li, P. J. Schuffler, Q. Yu, H. Chen, Y. Tang, and G. Litjens. Deep learning methods for lung cancer segmentation in whole-slide histopathology images - the acdc@lunghp challenge 2019. *IEEE Journal of Biomedical and Health Informatics*, 25(2):429–440, 2021.

[19] V. Suzanne Klimberg William J Gradishar Kirby I. Bland, Edward M. Copeland III. *The Breast: Comprehensive Management of Benign and Malignant Diseases*. Elsevier, 5 edition, 2017.

[20] J. R. Molina, P. Yang, S. D. Cassivi, S. E. Schild, and A. A. Adjei. Non-small cell lung cancer: Epidemiology, risk factors, treatment, and survivorship. *Mayo Clinic Proceedings*, 83(5):584–594, 2008.

[21] K. A. Ban and C. V. Godellas. Epidemiology of breast cancer. *Surgical Oncology Clinics of North America*, 23(3):409–422, 2014.

[22] M. Untch, G. E. Konecny, S. Paepke, and G. von Minckwitz. Current and future role of neoadjuvant therapy for breast cancer. *Breast*, 23(5):526–537, 2014.

[23] S. V. Liu, L. Melstrom, K. Yao, C. A. Russell, and S. F. Sener. Neoadjuvant therapy for breast cancer. *Journal of Surgical Oncology*, 101(4):283–291, 2010.

[24] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. Peter Campbell. Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science and Technology*, 9(2), 2020.

[25] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 1 edition, 2017.

[26] Andriy Burkov. *The hundred-page machine learning book*. 2019.

[27] Y. Tian. Artificial intelligence image recognition method based on convolutional neural network algorithm. *IEEE Access*, 8:125731–125744, 2020.

[28] Y. Wang, Y. Chen, and H. Yu. Agent's activity recognition: A focus on comparison of automatically-learned and hand-crafted features. In *International Conference on Advanced Mechatronic Systems, ICAMechS*, volume 2019-August, pages 241–244.

[29] Seth Weidman. *Deep Learning from Scratch: Building with Python from First Principles*. O'Reilly Media, 1 edition, 2019.

[30] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[31] F. Milletari, N. Navab, and S. A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, pages 565–571.

[32] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[33] H. L. Chen, B. Yang, J. Liu, and D. Y. Liu. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 38(7):9014–9022, 2011.

[34] W. R. Crum, O. Camara, and D. L. G. Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, 2006.

[35] T. K. Koo and M. Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016.

[36] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, 2018.

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, pages 2818–2826.

[38] R. Raina, A. Y. Ng, and D. Koller. Constructing informative priors using transfer learning. In *ACM International Conference Proceeding Series*, volume 148, pages 713–720.

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[40] H. H. N. Pham, M. Futakuchi, A. Bychkov, T. Furukawa, K. Kuroda, and J. Fukuoka. Detection of lung cancer lymph node metastases from whole-slide histopathologic images using a two-step deep learning approach. *American Journal of Pathology*, 189(12):2428–2439, 2019.

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

[42] H. Qu, G. Riedlinger, P. Wu, Q. Huang, J. Yi, S. De, and D. Metaxas. Joint segmentation and fine-grained classification of nuclei in histopathology images. In *Proceedings - International Symposium on Biomedical Imaging*, volume 2019-April, pages 900–904.

[43] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9351, pages 234–241. 2015.

[44] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, pages 770–778.

[45] S. Roy, A. kumar Jain, S. Lal, and J. Kini. A study about color normalization methods for histopathology images. *Micron*, 114:42–61, 2018.

[46] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[47] A. C. Ruifrok and D. A. Johnston. Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology*, 23(4):291–299, 2001.

[48] B. E. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Höller, A. Homeyer, N. Karssemeijer, and J. A. W. M. Van Der Laak. Stain specific standardization of whole-slide histopathological images. *IEEE Transactions on Medical Imaging*, 35(2):404–415, 2016.

[49] F. G. Zanjani, S. Zinger, B. E. Bejnordi, J. A. W. M. Van Der Laak, and P. H. N. De With. Stain normalization of histopathology images using generative adversarial networks. In *Proceedings - International Symposium on Biomedical Imaging*, volume 2018-April, pages 573–577.

[50] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Transactions on Medical Imaging*, 35(8):1962–1971, 2016.

[51] Y. He, J. Wei, S. Che, S. Liu, and P. Luo. Computer-aided pathological annotation framework: A deep learning-based diagnostic algorithm of lung cancer. In *Proceedings - 2019 International Conference on Information Technology and Computer Application, ITCA 2019*, pages 110–113.

[52] H. Zhang and V. M. Patel. Densely connected pyramid dehazing network. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3194–3203.

[53] M. Saric, M. Russo, M. Stella, and M. Sikora. Cnn-based method for lung cancer detection in whole slide histopathology images. In *2019 4th International Conference on Smart and Sustainable Technologies, SpliTech 2019*.

[54] H. Lin, H. Chen, Q. Dou, L. Wang, J. Qin, and P. Heng. Scannet: A fast and dense scanning framework for metastastic breast cancer detection from whole-slide image. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 539–546.

[55] Nobuyuki Otsu. Threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*, SMC-9(1):62–66, 1979.

[56] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[57] Z. Pawlak. Rough sets. *International Journal of Computer Information Sciences*, 11(5):341–356, 1982.

[58] A. T. Azar and S. A. El-Said. Performance analysis of support vector machines classifiers in breast cancer mammography recognition. *Neural Computing and Applications*, 24(5):1163–1177, 2014.

[59] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[60] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[61] X. X. Niu and C. Y. Suen. A novel hybrid cnn-svm classifier for recognizing handwritten digits. *Pattern Recognition*, 45(4):1318–1325, 2012.

[62] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[63] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte. Breast cancer histopathological image classification using convolutional neural networks. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2016-October, pages 2560–2567.

[64] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 2, pages 1097–1105.

[65] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas. A method for normalizing histology slides for quantitative analysis. In *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009*, pages 1107–1110.

[66] M. Peikari, S. Salama, S. Nofech-Mozes, and A. L. Martel. Automatic cellularity assessment from post-treated breast surgical specimens. *Cytometry Part A*, 91(11):1078–1087, 2017.

[67] A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin. Deep convolutional neural networks for breast cancer histology image analysis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10882 LNCS, pages 737–744. 2018.

[68] Y. Xu, Z. Jia, L. B. Wang, Y. Ai, F. Zhang, M. Lai, and E. I. C. Chang. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinformatics*, 18(1), 2017.

[69] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[70] S. Akbar, M. Peikari, S. Salama, A. Y. Panah, S. Nofech-Mozes, and A. L. Martel. Automated and manual quantification of tumour cellularity in digital slides for tumour burden assessment. *Scientific Reports*, 9(1), 2019.

[71] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling. Rotation equivariant cnns for digital pathology. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11071 LNCS, pages 210–218.

[72] Nofech-Mozes S. Salama S. Akbar S. Peikari M. (2019). Martel, A. L. Assessment of residual breast cancer cellularity after neoadjuvant chemotherapy using digital pathology [data set]. *The Cancer Imaging Archive.* doi:https://doi.org/10.7937/TCIA.2019.4YIBTJNO.

[73] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, O. Geessink, N. Stathonikos, M. C. R. F. Van Dijk, P. Bult, F. Beca, A. H. Beck, D. Wang, A. Khosla, R. Gargeya, H. Irshad, A. Zhong, Q. Dou, Q. Li, H. Chen, H. J. Lin, P. A. Heng, C. Haß, E. Bruni, Q. Wong, U. Halici, M. U. Öner, R. Cetin-Atalay, M. Berseth, V. Khvatkov, A. Vylegzhanin, O. Kraus, M. Shaban, N. Rajpoot, R. Awan, K. Sirinukunwattana, T. Qaiser, Y. W. Tsang, D. Tellez, J. Annuscheit, P. Hufnagl, M. Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvuori, K. Liimatainen, S. Albarqouni, B. Mungal, A. George, S. Demirci, N. Navab, S. Watanabe, S. Seno, Y. Takenaka, H. Matsuda, H. A. Phoulady, V. Kovalev, A. Kalinovsky,

V. Liauchuk, G. Bueno, M. M. Fernandez-Carrobles, I. Serrano, O. Deniz, D. Racoceanu, and R. Venâncio. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA - Journal of the American Medical Association*, 318(22):2199–2210, 2017.

[74] Staintools - staintools documentation. https://staintools.readthedocs.io/en/latest/index.html. Online; Accessed 27-June-2022.

[75] Pytorch - pytorch documentation. https://pytorch.org/docs/stable/index.html. Online; Accessed 11-July-2022.

[76] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 4, pages 3320–3328.