POLITECNICO
MILANO 1863

Politecnico di Milano
Dipartimento di Elettronica, Informazione e Bioingegneria
LAUREA MAGISTRALE IN INGEGNERIA BIOMEDICA

BRAVE AI:
BReast cancer Assessment Via Efficient Artificial Intelligence in Histopathology

Master of Science thesis of:
Ana Bogdanovic
Matricola 914342

Advisor:
Prof. Marco D. Santambrogio

Co-advisor:
Prof. Giovanni Squillero, PoliTo

Dott. Ing. Eleonora D'Arnese

Academic Year 2019/2020

Acknowledgments

Somewhere, something incredible
is waiting to be known.

Carl Sagan

I would like to express sincere gratitude to all of the people who have helped me throughout this period; thank you for immense support, patience, and encouragement which gave me strength to always move forward.

Firstly, I would like to thank Marco Domenico Santambrogio for his guidance, and the confidence he gave me; without him this thesis would not have been possible. I am grateful for all the words said; they have helped me grow academically and personally.

A special thank goes to Eleonora D'Arnese for her support and engagement which was crucial to carry out this work from the very beginning to the end. I am thankful for our meetings which were of great value to me, and this work.

Furthermore, I want to thank the NECSTLab, and all of the people in it, for nourishing a friendly and at the same time challenging environment full of opportunities: "you never know what is next at the NECSTLab".

To my parents, who have believed in me, and in my choices, who encouraged me and were there for me without a doubt. Always.

Last but not least, I want to thank my friends for sharing this experience with me. You make everything so special.

Abstract

Breast cancer is the most common type, and the leading cause of cancer-related deaths among women worldwide. The assessment process starts with the imaging tests which provide initial diagnosis, but the reliable result of presence of cancer can be determined only with a biopsy. It requires the sectioning of a small tissue sample and its analysis under the microscope. Nowadays, the diagnosis requires several days to complete, thus, it represents the bottleneck of the dynamic of modern day hospitals. Therefore, the automation of the diagnostic process in histopathology, with accurate detection of breast cancer is one of the most addressed challenges in the recent years. Within this context, Machine Learning, and especially Deep Learning, which are more and more used for automation of decision making processes, provide large space for exploration of a possible solution. The main characteristic of this methodology is the utilization of the data as the only resource of knowledge about the underlying condition, and its characterization.

BRAVE AI presents the design and the implementation of an automated, end-to-end pipeline for Invasive Ductal Carcinoma (IDC) detection in histopathological Whole Slide Images (WSIs). The purpose of the work is to facilitate, standardize, and accelerate the breast cancer diagnosis process which would, in turn, reduce the pathologists' workload and enable higher throughput from the pathology departments in hospitals. Exploiting information contained in the patches of WSI, we train multiple Neural Network (NN) models narrowing down the space for optimal solution of the imposed problem. The approach is validated on an open-source dataset used in multiple works from the state of the art. BRAVE AI best performing model relies on DenseNet121 architecture, and obtains a balanced accuracy of 88.41%, a F1 score of 89.55%, and a sensitivity of 91.97%, achieving performance comparable to the state of the art.

Sommario

Il cancro al seno è il tipo di cancro più comune e la principale causa di morte tra le donne di tutto il mondo. Il processo di valutazione di questa patologia inizia con i test di imaging che forniscono una diagnosi iniziale, ma il risultato più attendibile sulla sua presenza può essere determinato solo con una biopsia. Questo esame richiede il sezionamento di un piccolo campione di tessuto e la sua analisi al microscopio. Al giorno d'oggi, la diagnosi richiede diversi giorni rappresentando così il collo di bottiglia della dinamica degli ospedali moderni. Pertanto, l'automazione del processo diagnostico in istopatologia, con l'obiettivo di un'accurata identificazione del cancro al seno è una delle sfide più affrontate negli ultimi anni. In questo contesto, il Machine Learning, e soprattutto il Deep Learning, sono sempre più utilizzati per l'automazione dei processi decisionali, fornendo ampi spazi d'esplorazione per una possibile soluzione. La caratteristica principale di queste metodologie è l'utilizzo dei soli dati come unica fonte di conoscenza sulla condizione sottostante e unico strumento per la sua caratterizzazione.

BRAVE AI presenta dunque la progettazione e l'implementazione di una pipeline end-to-end automatizzata per il rilevamento del Carcinoma Duttale Invasivo da immagini WSI (Whole Slide Image) istopatologiche. Lo scopo di questo lavoro è quello di facilitare, standardizzare ed accelerare il processo di diagnosi del cancro al seno che, a sua volta, ridurrebbe il carico di lavoro demandato ai patologi e consentirebbe una maggiore produttività dai reparti di patologia negli ospedali. Sfruttando le informazioni contenute in sotto aree estratte dalle immagini WSI sono stati addestrati più modelli di reti neurali al fine di identificare la soluzione ottimale e riducendo così lo spazio delle possibili soluzioni. L'approccio è stato validato su dati open-source utilizzati in diversi lavori presenti nello stato dell'arte. Il modello che raggiunge le migliori prestazioni all'interno di BRAVE AI si basa sull'architettura della rete DenseNet121 e ottiene un'accuratezza bilanciata di 88,41%,

un valore di F1 di 89,55% e una sensitività di 91,97%, ottenendo prestazioni paragonabili allo stato dell'arte.

Questa tesi è organizzata come segue:

- Il Capitolo 1 fornisce una panoramica generale del problema affrontato in questa tesi e gli obiettivi principali del lavoro;
- Il Capitolo 2 fornisce le informazioni biologiche e mediche riguardanti il tumore al seno e le informazioni necessarie per comprendere il problema e apprezzare l'impatto che l'automatizzazione del processo diagnostico porterebbe;
- Il Capitolo 3 descrive brevemente la teoria sottostante gli strumenti utilizzati per l'implementazione di BRAVE AI;
- Il Capitolo 4 discute i lavori presenti in letteratura sulle diverse soluzioni per l'analisi di Whole Slide Image;
- Il Capitolo 5 propone i dettagli sulle scelte metodologiche e implementative che stanno alla base della progettazione della soluzione proposta;
- Il Capitolo 6 segnala i punti chiave per la valutazione del sistema implementato e discute i risultati raggiunti;
- Il Capitolo 7 esamina i principali traguardi raggiunti da questo lavoro di tesi, i limiti individuati e si conclude con i possibili sviluppi e ricerche future.

Contents

List of Tables

List of Figures

Introduction 1

This Chapter provides an overview of the context in which is placed the work of BRAVE AI. Section 1.1 provides the details of the problem and the impacts of its resolving. In Section 1.2 we give a brief overview of the state of the art in the field, while in Section 1.3 we provide the proposed solution in the thesis. Finally, we conclude this Chapter with Section 1.4 where we summarize the outline of the thesis.

1.1 Problem Statement

Breast cancer is the most commonly diagnosed type of cancer among women in the world, with more than two million cases in 2018 reported by the World Health Organization (WHO) [1]. It accounts for around 15% of the burden of cancer mortality, even though recent research shows that early detection and effective treatment increase the five-year survival rate to 88%, and even more for the localized disease [2]. Nowadays, the process of cancer diagnosis is long and cumbersome. It starts with imaging tests like mammography, which are able to capture the mass inside the body providing initial diagnosis, but the final result of the presence of cancer can be determined only with a biopsy examination. The biopsy sample is then investigated by a skilled pathologist who is looking for the cancerous cells and making the definitive diagnosis. This process takes around one to two weeks to be completed. Furthermore, due to the aging population, there is an increasing interest in preventive and personalized medicine, requiring screening protocols, and specific testing which leads to a larger workload in the laboratories. This all

results in the increased need for automation, which is not only speeding up the process but also making it more accurate and reliable. Furthermore, this is enabling remote work which is crucial for places with no specialists, or during the time of pandemic like in 2020.

In recent years, due to the aforementioned need of the diagnostic process automation, there is an emerging field called digital histopathology. Specifically, In this thesis, we are going to address the problem of Invasive Ductal Carcinoma (IDC) identification from histopathological Whole Slide Images (WSIs), which is the last step in the diagnostic pipeline. In this domain, Machine Learning (ML), and, more recently, Deep Learning (DL), have been able to provide encouraging results in the identification of cancer present in the tissue slices. The problem here is how to generalize the decision-making process, and which models are able to perform well on tissue samples that they have not seen before. On the other hand, since the positive or the negative result of the biopsy has a huge impact on the life of the patient, and even bigger consequences if the result of the biopsy is wrong, there is a problem of estimating model performance and reliability of the results. The model must provide accurate and reliable results in order to be implemented as part of decision-making support systems for precision medicine.

1.2 Machine Learning in Breast Cancer assessment

Mammography is the first, and the most common exam done in the process of breast cancer assessment, hence, there are many works addressing the problem of detecting suspicious lesions in mammograms. Some approaches rely on feature extraction, based on morphological analysis, and more traditional preprocessing of the images followed by a Neural Network (NN) [3], while other works employ automated pipeline working

directly with images utilizing Convolutional Neural Networks (CNNs) [4] to classify the images as benign or malignant. Nevertheless, even though mass identification is very important, none of these approaches is able to investigate the cells inside the body, and analyze their structure and shape, to provide the definitive diagnosis for the detected lump.

Therefore, the method employing tissue biopsy, and the analysis of the slides is the only way to confirm the presence of cancer with high reliability. Since the manual analysis suffers from inter and intra observer variability [5], its automation can standardize the outcomes, and accelerate the whole process. There are two types of approaches, one employing shallow methods with manual feature extraction, and the other employing fully automated DL frameworks. Some shallow methods apply simple statistical classifiers obtaining lower accuracies [6], while others use classifiers like Random Forests (RFs) [7] obtaining high accuracy on their classification problems.

Nevertheless, the aforementioned methods for histopathological image classification use hand-made feature extraction which requires expertise in many sub-domains, and careful integration of engineering and medicine. It is very laborious to include new datasets in these frameworks since the feature extraction requires major adjustments each time different data is presented. On the other hand, DL models automatically perform the process in a completely end-to-end manner, learning the features ranging from simple to abstract ones in the deeper layers. Early works like [8] used a custom made CNN obtaining results of 71% for the F1 score and 84% for the balanced accuracy, while more recent works like the ones proposed in BASH [9], and CAMELYION [10] challenges use architectures already proven on classification problems like ImageNet, in order to classify images as cancerous or healthy ones. The most commonly used architectures are Residual Neural Networks (ResNets), VGG networks, and Densely connected Neural Networks (DenseNets).

1.3 BRAVE AI Solution

Analyzing the current state of the art and the challenges, the purpose of BRAVE AI is to develop and implement an automated, end-to-end pipeline for breast IDC detection in histopathology images, aiming to be a support for the pathologist during the assessment of the condition. This automated procedure can offer standardization of the diagnostic process, reliability, and reproducibility also allowing a reduction of time needed for one diagnosis and lightening the workload of the pathologist.

This thesis proposes a methodology and a software tool that, using histopathology image patches of WSIs from breast cancer, and DL based classification, provides the identification of cancerous patches. Furthermore, we provide the visualization of the cancer probability map from the reconstructed histopathology image of the patient's biopsy slide.

The contributions of this work are:

- Reproducible methodology, design, and implementation of an automated end-to-end pipeline for breast cancer assessment;
- Exploration of Deep Neural Network architectures as the space of solutions for histopathological images classification;
- Implementation of a Deep Neural Network classifier based on the histopathology images, which has performance comparable to the models from state of the art on the same dataset.

The validation of the pipeline is obtained using a public dataset obtained by Cruz-Roa et al. [8], and our best performing model using DenseNet121 architecture obtained a balanced accuracy of 88.41%, a F1 score of 89.55%, and a sensitivity of 91.97%, and it achieves performance comparable to the results from the state of the art.

1.4 Outline

The thesis is organized in seven chapters. Chapter 2 provides the basic biological and medical knowledge needed to understand the problem and the imapct of automated solutions. Chapter 3 gives a brief description of the tools used for the implementation, while Chapter 4 discusses the solutions in the state of the art. Chapter 5 explains into details the proposed pipeline with the specifics of its implementation. Chapter 6 reports the experimental setup and, provides results of the performed experiments. Finally, Chapter 7 presents the final remarks and conclusions of the work done in the thesis.

Background 2

This Chapter has the aim of providing basic theoretical knowledge that will facilitate the understanding of the faced challenges, and the reasons behind some of the choices that have been made for the development of this work. It mainly focuses on the background from biology and medicine which are crucial for understanding the problem and its proposed solution, and the future towards which the world goes when it comes to medicine, bioengineering, and emerging technologies. Section 2.1 introduces the domain of histopathology in which the work of BRAVE AI is focused. Section 2.2 goes into details of normal and pathological cell functioning. Section 2.3 explains how the breasts function, their importance and then explains into details the pathology of cancer and its types. Additionally, this section addresses the diagnostic methods and gives brief explanation of the medical procedures for its assessment. At the end, in Section 2.4 we present the possible impact which the presented work, and similar works in automation and Machine Learning (ML), may have in the development of medicine and medical diagnosis.

2.1 Histopathology

Pathology is the study of the causes and effects of a disease, injury, or any changes in cells, tissues, and organs that are associated with the disease. The name itself comes from the Greek language and it means the study (logos) of suffering (pathos).

Principally, there are two major facets of pathology: *etiology* and

pathogenesis. Etiology encompasses the knowledge of the underlying causes and factors that are initiators of the disease or they are related to its progression. In short, it attempts to explain why the disease occurs. For example, diabetes, hypertension, and cancer are caused by a combination of inherited genetic susceptibility and various environmental triggers. On the other hand, pathogenesis is related to the mechanisms that lead to the state of the disease. In other words, it describes how a disease develops. These mechanisms are responsible for structural, functional, and morphological changes on cellular and molecular levels. These changes can be observed and investigated to characterize different diseases [11]. Knowing the disease, its etiology, and pathogenesis is crucial for defining effective treatments and prevention measures. Moreover, the investigation of these abnormalities requires a deep understanding of how the healthy and normal organism functions, which requires entering the field of *histology*. In particular, we can define histology as the study of microanatomy of cells, tissues, and organs of the body. It analyzes tissue biology, especially focusing on the correlation between structure, organization, and arrangement of cells which produces different organs' functions [12].

The intersection of histology and pathology is giving rise to the discipline called *histopathology*. It refers to the examination of tissue structural changes with the aim of studying a particular disease. Recognizing and evaluating changes in the tissue, and providing diagnostic information is a manual process which requires analysis by highly skilled medical practitioners. To perform the histopathology examination the patient must undergo a tissue biopsy, the process which involves the extraction of cells and pieces of tissues, to determine the diagnosis [13].

2.2 The cell - living unit of health and disease

The cell is the basic structural, functional, and biological unit of all living organisms [14]. It is the smallest unit of life. The cell is surrounded by the plasma membrane inside which are different types of organelle, cytoplasm, and, optionally, nucleus depending on the type of the cell. The nucleus is the largest organelle in the cell, and it contains the Deoxyribonucleic Acid (DNA) which is the genetic material of the cell [15].

The development of a multicellular organism involves cellular replication, growth, and functional differentiation. Almost all cells replicate through mitosis producing two genetically identical daughter cells. The daughter cells continue with replication and, sometimes, they evolve to specialize in a specific function, thus creating differentiated cells with particular functions like the cells of the skin or muscles. The only cells that do not follow this type of division are male and female germ cells which divide by meiosis. Cell division is happening throughout the whole life of the organisms, and it depends on the function and the homeostatic machinery, therefore it is tightly controlled to meet the needs of the organism. An important part of the normal cycle is the programmed death of the old cells and that mechanism is called apoptosis [16]. When normal regulatory influences that protect the body, and keep the organism balanced, are broken down the cells may become cancerous.

Cancer is a pathology caused by a rapid division of abnormal cells within the body which destroys or replaces the normal tissues. Cancer can spread from the inception site to other organs, and this process is called metastasizing [17]. For example, it is very common for breast cancer to spread to the lungs or liver, and cause complications like a respiratory failure which can result in death [16]. Cancer is a genetic disease that can be traced to specific gene modifications in the DNA of the cell. It is not an inheritable disease, but arises during its lifetime due to a variety of risk factors [14]. Cancer-causing environmental exposures

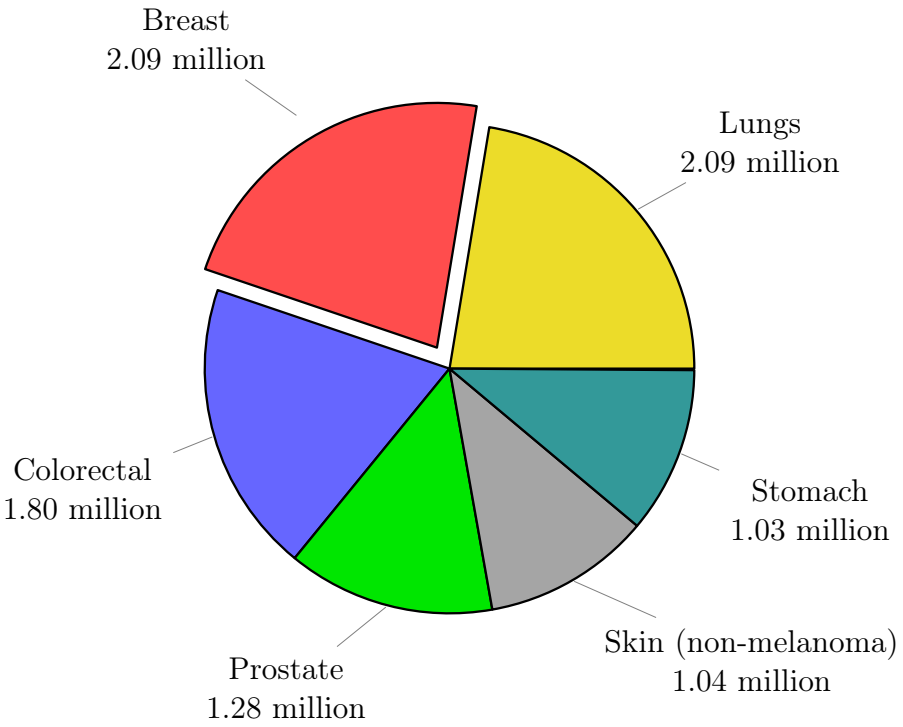


Figure 2.1: The most common types of cancer as reported by WHO.

include substances, such as the chemicals in tobacco smoke, radiation, and ultraviolet rays from the sun [18].

Cancer is a leading cause of death worldwide as reported by WHO [17]. Its most common types and their absolute number of occurrence are shown in Figure 2.1. According to another WHO report on cancer, in 2018, there were an estimated 18 million new cases of cancer and 10 million deaths from cancer worldwide. The predicted global burden will double to about 29–37 million new cancer cases by 2040. Of the 15 million premature deaths between the ages of 30 and 69 in the same year, 4.5 million are due to cancer. In fact, cancer develops in 1 out of 5 people before they reach the age of 75 [1].

2.3 The Pathology of Breasts

The work of the thesis focuses on histopathology of breast cancer as it is the second most common type of cancer in the world, and the first among the female part of the population. That is why this Section will firstly provide basics for understanding the functions of breasts, their importance on various levels and continue with explanations of the pathology and its types. Finally, it also addresses the methods for diagnosing the disease and gives brief explanations of the medical procedures.

2.3.1 Breasts

Breasts are milk-producing organs of mammals that provide appropriate nourishment to their offspring. They are essential for the survival of infants. The act of nursing has two important benefits: physiologically, it helps to involute the uterus; psychologically, it helps to "bond" the mother and the baby [2]. Furthermore, there are other, non-biological functions of this organ. Breasts are visible and as such have a social, cultural, and personal significance for the individual. This characteristic is not shared by other organs, except the skin which defines the race, and with it impacts the society at large. These features play a role when considering the origins and treatment of breast diseases [19].

On the other hand, biologically, breasts are defined as highly modified apocrine sweat glands (mammary glands). They develop embryologically along two lines, known as the milk lines, extending from the axillae to the groin. Mammals can have multiple glands, but in humans, only two glands are developed - one gland on each side of the thorax. The breasts of both sexes follow a similar course of development until puberty, after which the female breasts undergo visible changes in size and function as a result of the influence of pituitary, ovarian, and other hormones. Until menopause, the breasts go through cyclical changes in activity which are controlled by the hormones of the ovarian cycle. The

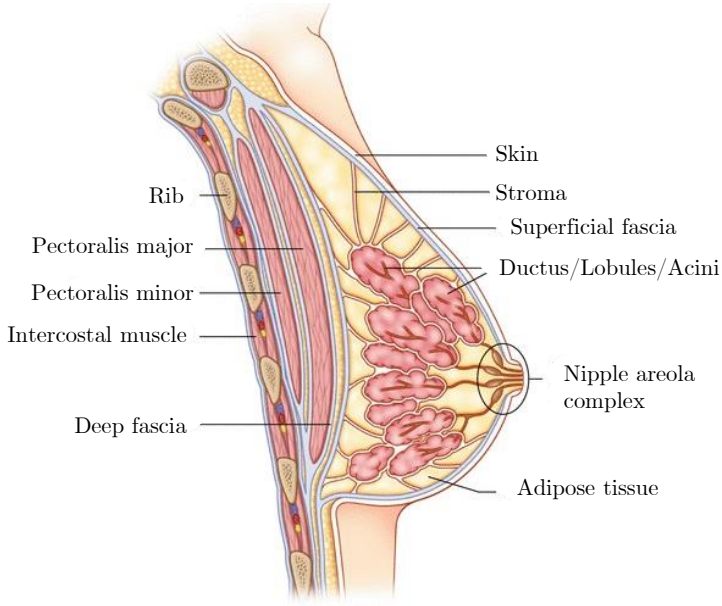


Figure 2.2: Schematic representation of the women's breast

most visible changes can be seen during and after pregnancy when they are producing milk for the baby. After menopause, the breasts, like the other female reproductive tissues, undergo progressive atrophy and involutional change [16]. A highly schematic representation of the anatomy of the breast is shown in Figure 2.2.

Considering everything mentioned above, it is obvious that breasts are very important for human development and survival, and they should be healthy. However, there are many pathologies that can affect their development, functions, and structure which have been studied throughout the years in order to give proper diagnosis and treatment and even prevent the occurrence of the diseases. The focus of this work is one of the most severe pathologies, that can even lead to death if left untreated, breast cancer.

2.3.2 Breast Cancer

Breast cancer is the most frequently diagnosed cancer, and the leading cause of cancer-related death among females worldwide, with an estimated 2.1 million new cases in 2018 which is the 24.2% of all detected cancer conditions in women part of the population [1]. The incidence of this disease increases with higher life expectancy, urbanization, and modern lifestyle which is characteristic of western countries, but nonetheless, breast cancer is the most common cancer among women both, of the developed and the developing world [20]. In the USA, it ranks second to lung cancer in terms of mortality. There are an estimated $41,000$ breast cancer deaths among women annually, which accounts for 15% of the burden of cancer mortality. In recent years, the mortality from this type of pathology has decreased due to early detection and advancement of treatments. Five-year survival rate is 88%, and five-year survival is 98% for women diagnosed with localized disease [2].

Tumor stage and type are remaining the most important determinants of the outcome, and an early detection of breast cancer allows more effective and less aggressive therapy assuring lower mortality rates, and better quality of life [1].

Depending on how the cells behave, there are two types of breast tumors: those that are non-cancerous, or 'benign', and those that are cancerous, generally called 'malignant'. Benign tumors are usually not aggressive towards neighbors, i.e. they lack the ability to invade surrounding tissue. Benign tumors are not removed unless they continue to grow and cause pain, pressure, or other problems to surrounding organs. On the other hand, malignant ones are cancerous and aggressive, which means that they invade and damage surrounding tissue. When a tumor is suspected to be malignant, the doctor will perform a histopathological analysis (biopsy) to determine the severity of the tumor [21].

Malignant tumors are classified morphologically according to whether

they have penetrated the basement membrane. Those that remain within this boundary are called in situ carcinomas, and they are non-invasive. On the other hand, those that have spread beyond it are called invasive carcinomas [19]. Breast cancer can spread when the cancer cells get into the blood or lymph system and are carried to other parts of the body [22]. There are many types and subtypes of breast cancer depending on which part of the breast and tissue it is affecting. The most common non-invasive types are [23]:

- Ductal carcinoma in situ (DCIS) which grows in the milk ducts of the breast.
- Lobular carcinoma in situ (LCIS) which develops in the milk-producing glands at the end of breast ducts called lobules.

The most common invasive types of breast cancer are [23]:

- Invasive ductal carcinoma (IDC) which affects the milk ducts, but it has broken through the lining of the duct and spread to surrounding tissue. It can also spread to the other parts of the body. About 80% of detected breast cancers are IDC.
- Invasive lobular carcinoma (ILC). The cancerous cells appear firstly in the lobules and spread into surrounding tissue.
- Paget's disease of the nipple. Cancerous cells grow in the nipple or in the area around it called areola. It is usually the indication of invasive cancer somewhere else in the breast.
- Inflammatory breast cancer. This is rare and aggressive form of invasive cancer that affects the blood vessels and/or lymphatic vessels of the breast.
- Phyllodes tumours of the breast. They can be benign or malignant. They develop quickly inside the connective tissue of the breast.

- Locally advanced breast cancer. It is usually large and spreads to skin, chest wall or muscles and may have involved lymph nodes.

- Metastatic breast cancer. This is stage IV breast cancer, and it has spread to lungs, bones, liver or other more distant parts of the body.

2.3.3 The Process of Diagnosis

The process of diagnosis usually starts with a woman noticing a small lump under the arm, or swollen breast during-self examination. Based on the symptoms, the doctor can request a different type of test to determine if the lesion is cancerous, and give a suitable diagnosis and treatment.

One set of tests that can be done belongs to the group of imaging tests. They are producing images of the inside of the body, particularly, in this case, around the suspicious area in the breast. The three prevalent ones are: mammography, ultrasound, and MRI [24].

Mammography. The resulting image is an x-ray of the breast. Screening mammography is used to detect breast cancer in women without apparent symptoms. On the other hand, diagnostic mammograms are more detailed and time-consuming since the images are taken from multiple vantage points. They are used after screening or based on some signs of the pathology. The reliability of a mammogram depends on the size of the tumor, the density of the breast tissue, and the skill of the radiologist [25].

Ultrasound. This type of examination involves penetrating sound waves to create an image of the breast tissue. The tissue is not affected nor damaged during this process. Ultrasound distinguishes between cancer which is a solid mass and a cyst that is filled with liquid. The picture generated by the ultrasound is called a sonogram, and its reliability depends on the size of the mass [26].

Magnetic Resonance Imaging (MRI). This method uses a magnetic field to produce a detailed image of the body. It is usually used after the patient has been diagnosed with cancer to determine how much the disease has spread in the breast or throughout the body.

The second group of tests used for the diagnosis of breast cancer is called biopsies. A biopsy is the removal of a small amount of tissue for examination under the microscope. The aforementioned types of tests can suggest if the mass is present, but only a biopsy can make a definitive diagnosis. After the extraction of tissue a pathologist analyzes the sample. The type of biopsy depends on the size of the needle used to collect the sample and can be classified as surgical, core needle, sentinel lymph node, and image-guided biopsy [24].

Surgical biopsy removes the largest amount of tissue but the patient has to undergo surgery. Since most of the examinations are not diagnosed as cancer this process is not recommended because it means that a person takes unnecessary surgical operation.

Core needle biopsy is used for the extraction of a larger sample of the tissue. It is performed to diagnose whether the cancer is invasive or not, and what the cancer biomarkers are. Biomarkers are substances that are produced by the tumor or by the body in response to cancer.

Sentinel lymph node biopsy is used to determine whether cancer has spread to the lymph nodes around it. The lymph node which is the first to get infected is called the sentinel lymph node, and in breast cancer that is usually the node under the arm.

Image-Guided biopsy exploits additional imaging techniques like ultrasound or MRI which are used to guide the needle to the location of the mass. It is also called fine needle biopsy and it is used when the lump is likely to be filled with fluid [27].

After one of the aforementioned biopsies, the sample undergoes the technical processes of preparation and histological staining and then is

used for the diagnostic procedure done by a skilled pathologist who is looking for cancerous or abnormal cells. The pathology report takes 1-2 weeks to complete [27]. The preparation procedure is described in the following section.

2.3.4 Histological tissue preparation and staining

Histological tissue preparation and staining are the series of processes conducted to visually label the tissue which is to be used in the microscope study [28]. The tissue section which is examined is otherwise colorless because the fixed protein and the glass have the same refractive index. Different dyes are linking to different tissue proteins and this helps to understand its morphology [29]. Methods of staining make various tissue components not only non-transparent but also distinguishable from one another. Dyes stain the sample behaving like acidic or basic compounds and forming electrostatic (salt) linkages with ionizable radicals of macromolecules in the tissues. If we observe the inside of the cell we can see components with different net charges: negative ones called aions, and positive ones called cations. For example, nucleic acids have a net negative charge and they have an affinity for basic dyes, and are termed basophilic. On the other hand, proteins that have a net positive charge with many ionized amino groups, stain more readily with acidic dyes and are termed acidophilic [30]. Histological tissue preparation is the multi-step process composed of fixation, processing, embedding, sectioning, and staining [31].

Fixation has the goal of preserving the natural tissue structure and delaying the degradation of the cell structure. The most used chemical for this process is formalin. After the fixation, **processing** has the objective to dehydrate the sample, i.e. remove water from the selected tissue to solidify them and facilitate the cutting of thin sections of slides. This is done by transferring sample through a series of alcohol solutions end-

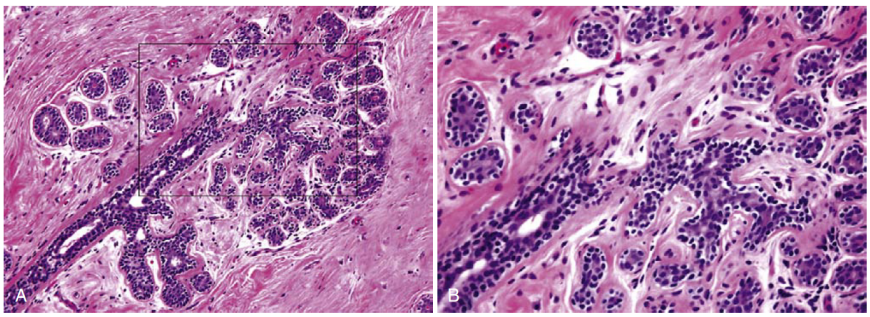


Figure 2.3: On the left is example of tissue stained with Hematoxylin and Eosin (H&E) - terminal duct lobular unit (breast tissue). On the right is the part of left image zoomed in to show more clearly the structure of the tissue.

ing with 100% ethanol. The result of **embedding** step is to secure the specimen in a block of paraffin wax for section cutting and storage without changing the morphology of the tissue. Since this process may cause changes in structure due to prolonged heating, and, thus, modifications of the Ribonucleic Acid (RNA) on high temperatures. Before concluding the procedure with the staining, there is the **sectioning** which produces 'ribbon'-like microtomes of a tissue which are later mounted on a microscope slide for examination [31]. Finally, the last step of the process is **staining**, which is utilized to apply the dye color on the posterior and anterior border of the sample tissues to locate the pathological cells (e.g. tumorous cells). The combination of Hematoxylin and Eosing (H&E) is the most commonly used for the staining process. Hematoxylin, the basic dye, stains acid structures like DNA in the cell nucleus, RNA-rich portions of the cytoplasm, and the matrix of cartilage, producing a dark blue or purple color. On the other hand, eosin is an acidic dye and it stains other cytoplasmic structures and collagen in red or pink [30]. An example is shown in Figure 2.3. In addition, this method is quick to execute, cheap and can be altered.

Slide preparation, from tissue fixation to observation with a light

microscope, may take from 12 hours to $2\frac{1}{2}$ days, depending on the size of the tissue, the embedding medium, and the method of staining [30]. After this process, all the samples are manually analyzed by the pathologist in the laboratory, who is the expert with knowlege of cell behaviour and structure, and also relying on previous experience, giving a final diagnosis on whteher the pathology is present or not, and its type.

2.4 The Future of Pathology is Digital

In the past years, alongside the improvements in technology and innovation, comes the modernization and digitalization of the hospital. The waiting time for the results of the laboratory tests is becoming shorter every day, and, nowadays, we can get hormone or blood test results in just a couple of hours to our email address.

Unfortunately, histological and histopathological analyses are still not able to keep the pace with the requirements and norms posted by the dynamic present-day hospitals, and thus, they are labeled the slowest of all the laboratory departments. The process which includes the preparations and steps mentioned in the sections above is taking a couple of days or even weeks to provide the results.

The reasons to automate and speed up the process of histology processing come from both - financial and healthcare points of view. We should consider that, due to the aging population, there is an increase of interest in preventive and personalized medicine, requiring screening protocols and specific testing which lead to the larger workload in the laboratories. On the other hand, there is also the need for standardization and verification of all the steps and processes used which can be achieved with automation. Automation, in this domain, has been recognized as the need of society and is the main drive that leads to the development of the field of Digital Pathology (DP) [32].

DP is a broad and general term that refers to the development of dig-

ital workflow and imaging solutions in pathology. The framework goes towards creating a digital image-based practice environment in which a Whole Slide Image (WSI) or another digital image is acquired, managed, interpreted, and searched for specific content [33].

The main advantages of DP can be summarized as follows [34]:

- Improved standardization of the methods.
- Simultaneous and rapid examination of several regions, and staining procedures at any magnification.
- Simplification of morphological findings through digital tools (counting, annotations, measurements).
- Support of the assessment process by use of Artificial Intelligence (AI) applications, ML systems for decision making, and quantification of diagnostic and predictive markers.
- Provision of detailed clinical information in one dataset.
- Easy and quick access to digital archives with the data.
- Error-free slice preparations.
- Enabling remote diagnosis and consultation with pathologists around the world, which is crucial for places where only a few specialists are available.
- Flexible job opportunities - remote work.

Here should be mentioned that all the steps of the histological analysis process mentioned in Section 2.3.4 can be automated. They are highly repetitive and there are already working solutions and machines approved by Federal Drug Association (FDA) that are speeding up the process, like for example The Leica ASP200S/ASP300S Tissue Processor [35]. Note that after the staining process comes to the analysis done

by a skilled pathologist, but with the introduction of WSI it can be automated, or at least quickened, with the improvements in ML, especially Deep Learning (DL).

Automatization and/or assistance to the pathologist's analysis is the domain of research considered in this thesis and the presented work is going to focus on the automatization of the analysis of WSI when it comes to predicting the presence of breast cancer with the novel methods in DL.

Machine Learning 3

This Chapter aims to provide to the reader the basics of Machine Learning (ML), which is a trend in computer science, both in research and in industry. This approach is used more and more to solve all big data problems as well as any issues that have observable patterns and repetitions in the data that contribute to decision-making in the domain. In Section 3.1 we focus on a general overview, trying to clarify common misconceptions about it, while in Section 3.2 we provide a basic overview of Convolutional Neural Networks (CNNs) needed to understand the methodology and complex ML architectures used in the thesis which are going to be explained later in details.

3.1 Artificial intelligence, Machine Learning and Deep Learning

Since concepts of ML, Deep Learning (DL) and Artificial Intelligence (AI) are going to be mentioned interchangeably throughout the thesis, we should clarify their relationship at the beginning.

AI enables computers and machines to mimic the perception, learning, problem-solving, and decision-making capabilities of the human mind. Therefore, AI is the most general concept in this domain. It includes expert systems or any application that makes decisions based on complex rules. ML is a subset of AI which can learn automatically by itself from given data. As the amount, diversity, and quality of data increases - the accuracy increases too. The last concept is DL, and it

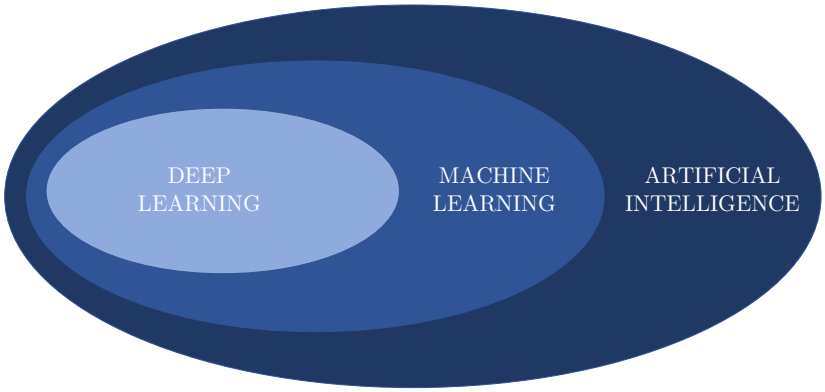


Figure 3.1: Relationship among DL, ML and AI

is the subset of both ML and AI, which is focusing on problems in the world of big data and Deep Neural Networks and the idea that the system can learn from raw data without feature engineering or any human assistance [36]. The relationship among the three concepts can be seen in Figure 3.1.

When we say that the system is able to learn from the data we usually talk about two types of learning: supervised and unsupervised. In supervised learning, the goal is to learn a mapping from input x to output y, when given a labeled set of input-output pairs. When the output y is categorical the problem is known as classification, whereas when y is a real-value variable the problem is called regression. On the other hand, in unsupervised learning, we are only given the input x and the goal is to find patterns in the data. This problem is not a well-defined one, and we are not told what kind of patterns to look for, thus this method is usually associated with knowledge discovery [37].

Even though pattern recognition and statistical ML are very powerful tools, they require proficiency and expertise from the data domain

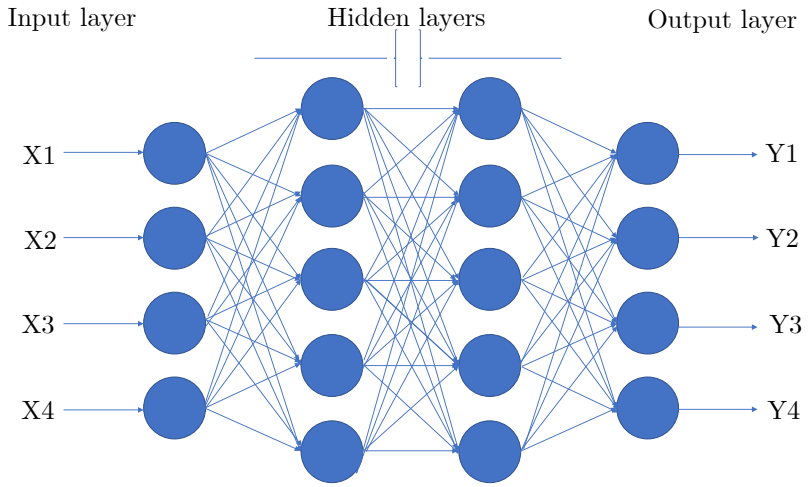


Figure 3.2: Schematic representation of Deep Neural Network (NN).

to solve the problem imposed in the thesis - providing diagnosis from breast tissue biopsies. In other words, to do this type of ML with data coming from histopathology, one must have specific knowledge in the mentioned field and be able to transform the Whole Slide Image (WSI) into meaningful features for further processing of the classifier or any other learning system. Such feature engineering can be labor-intensive and cannot scale well in general. On the other hand, it is obvious that when it comes to processing data in its raw form these techniques fail to produce meaningful results [38].

Therefore, the solution to this problem is coming with the growth of computational power and resources and consequentially the rise of DL. DL encompasses ML methods that are based on representation learning. It allows computational models to learn representations of data with several stages of abstraction obtained by stacking several non-linear modules that, starting from the raw input, transform data on multiple levels. Different layers capture different motifs, such as edges and orientations at the beginning and more complex objects as body parts or objects in

the deeper layers. Therefore, the model is able to learn the important features and internal structure of the dataset without the need for the assistance of humans nor expertise and domain knowledge [38].

Usually, when we talk about DL we talk about Artificial Neural Networks (ANNs) or simply Neural Networks (NNs). NN is, simply said, a computational model which transforms inputs into outputs through a series of nonlinear computations. Every NN consists of at least one, but usually many, basic computational units called neurons, which are stacked in interconnected layers. They are represented with blue circles in Figure 3.2. Each unit has inputs coming from previous layers, which are processed in a specific way, and it has one output that is propagated to the units in the following layer. The output of a neuron does not have to be propagated to all of the neurons in the following layer, nor just to the first following layer, but also the second, third, etc. There are many ways in which the signals can be propagated and they are defining the type of the NN [39]. Some of the examples are: fully connected NN (Figure 3.2), CNNs, U-nets, Autoencoders, etc. The details about some of them are going to be discussed later. All the layers which have the units that propagate the outputs to the following layers are called hidden layers. If the NN has more than one hidden layer it is considered as a deep NN, but the definition is not strict and it varies in the literature. The main point is that deep NNs solve the problem hierarchically through stacked layers of neurons performing different tasks.

The neuronal computational model is based on the work by Rosenblatt from 1958 [40]. Defining the parameters and their mathematical relationships of one neuron can be seen in Figure 3.3. They are the weights of the connections ω, the bias b, and the activation function φ. We can have many different activation functions. The examples of activation functions are sigmoid, hyperbolic tangent, and Rectified Linear Unit (ReLu), with the last one being the most used among them.

The reason is the fact that it does not have the problem of saturation of derivative which accelerates the convergence of the training of a NN, and it is computationally very simple to calculate.

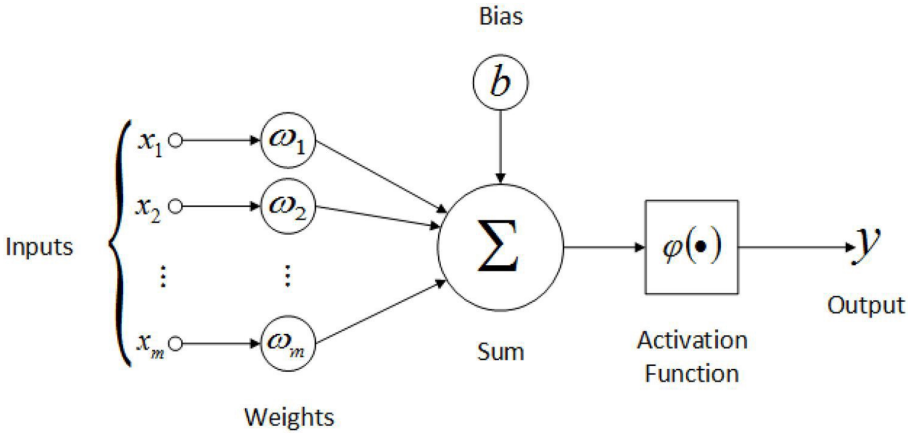


Figure 3.3: Schematic representation of a neuron in NN.

Assuming that the neuron has m inputs, we can define the output y of the neuron as:

$$y = \varphi(\sum_{i=1}^{m} \omega_i \times x_i + b) \tag{3.1}$$

When we say that NNs are able to learn from the data through training, it means that the weights and the biases of all the neurons in the NN are iteratively updated as the NN attempts to minimize the certain objective function. The objective function is usually the error, and it is also called the cost or loss function. Learning is an optimization problem and we can use different algorithms to search through the space of all the possible parameters in order to obtain good enough predictions. Typically, a NN is trained using the gradient descent optimization algorithm and the weights are updated using the error backpropagation algorithm [41].

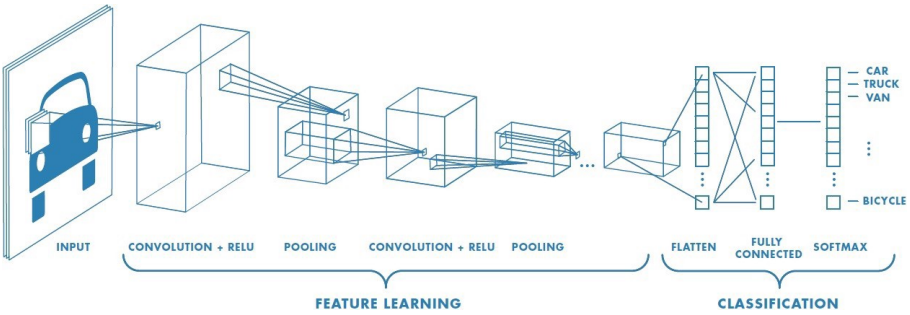


Figure 3.4: Schematic representation of a CNN and its layers.

3.2 Convolutional Neural Networks

CNNs are a very popular and powerful method for solving computer vision problems and perform image analysis. The architecture of CNNs is inspired by the receptive field structures which are found in the human primary visual cortex in the brain. This means that individual neurons respond only to the small region of the visual field, but the larger collection of neurons covers the whole visual area [42]. As a consequence, the main characteristic of these networks is not to do the mapping of input to output, but to learn the internal structure of the data.

In order to use CNNs, we do not need to extract hand-crafted features, because the network will do this for us by learning the filters which are applied to the input. CNNs are always deep NNs because they learn the structure of data on different levels. Firstly, they learn how to extract simple details like vertical and horizontal edges, and then from them pass to the extraction of colors, part of the objects, or more complex patterns. The structure of this type of NNs typically consists of two types of layers that are repeated several times (in-depth) called convolutional and pooling layers. Convolutional layers are working as filters to the input image, and there can be many filters applied at once producing many processed images as a result of the output of the layer.

The convolution function is defined as follows [43]:

$$S(i,j) = (f * g)_{i,j} = \sum_{k=0}^{p-1} \sum_{l=0}^{q-1} f_{i-k,j-l} \times g_{k,l} \tag{3.2}$$

where f and g are two matrices of dimensions m x n and p x q, respectively. In case of CNNs the matrix f is the input image and the matrix g is a filter (kernel). The output is sometimes called the feature map. Convolutional layers are followed by the activation function. After the convolutional layers, the CNN typically has a pooling layer. It is a form of non-linear downsampling which, after partitioning the image in rectangles, chooses a specific value from each rectangle, usually maximum, and produces it as the output. The main advantages of this operation are the elimination of noise, reduction of computation times for upper layers, and translation invariance. It should be noted that pooling reduces the size of the feature maps. After that, the network usually has a couple of fully connected layers which are processing the features produced by the convolutional backbone part of the NN. An example of CNN architecture is shown in Figure 3.4.

The main motivations behind the use of convolutional layers are sparse interactions, parameter sharing, and equivariant representations. In simple feedforward NN every output of a hidden layer is interacting with every neuron in the following layer. On the other hand, CNNs have sparse connections obtained by making the filter smaller than the input, thus allowing us to compute the output quickly and store a smaller number of parameters. This means that, in order to obtain the same performance as feedforward NN, CNN needs a smaller number of layers. Figure 3.5 shows the difference between fully connected layers seen in Figure 3.2 and layers with sparse connections. Parameter sharing refers to using the same parameter for more than one function in a model as seen in Figure 3.6 where the dark blue line represents weight with the same value. Simply said, we are doing the convolution of the input image

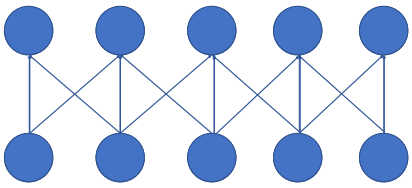


Figure 3.5: Sparse connections

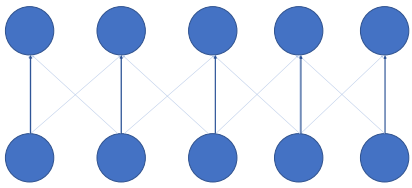


Figure 3.6: Parameter sharing.

part by part with the same filter, so each element of the filter is used at every position of the input. On the contrary, in a traditional NN, each element of the weight matrix is used exactly once when computing the output of a layer. It is multiplied by one element of the input and then never revisited. This is further reducing the storage requirements and lessening the time needed for training. The particular form of parameter sharing causes the layer to have a property called equivariance to translation. This means that if the input changes, the output changes in the same way [43].

To summarize, these were the main characteristics of CNNs, which are the motivation for their use in the problems of classification of histopathological images.

Related Work 4

Breast cancer detection is a challenge addressed from many points of view, and using different methodologies. This Chapter presents the main branches of this challenge, and goes through the related works from the literature, which are important for the research. We are going to explain different types of data that can be used, and different methods from more traditional roads, to the Machine Learning (ML) ones, explaining their advantages and disadvantages for solving the proposed problem. We will mainly focus on the problem of classification, and on works related to breast cancer datasets, since they are the topic of this thesis. Section 4.1 summarizes works related to the mammogram analysis since it is the first step towards breast cancer detection. Section 4.2 presents the state of the art in histopathological image analysis for the same task. Finally, in Section 4.3 we address the problem of datasets, data availability, and computational resources needed for achieving the results, and how these factors affect the research and implementation of given solutions in real hospitals.

4.1 Mammography Images Analysis

Mammography is the gold standard for breast cancer screening. In fact, this is usually the first test that the person undergoes when comes to an examination. It should be noted that a false negative in this context means that the person would be left untreated, and probably sentenced to a fatal end because of the disease. That is why many works are addressing the problem of breast cancer detection from mammograms.

These works usually aim to identify suspicious masses, calcification or any sort of abnormalities [44] [3].

Some works use more traditional preprocessing with Gaussian filters, Otsu thresholding, and region growing techniques from which the features are extracted and fed into a Neural Network (NN), or another type of classifier, that predicts the image as benign or malignant [34], [45]. Several other papers successfully employ Convolutional Neural Networks (CNNs) to obtain the classification of the lumps using extensive augmentations on the mammography images showing that Deep Learning (DL) can be used for this sort of problems [4]. In order to obtain more detailed images, some of the researchers use Magnetic Resonance Imaging (MRI) modality to get the information about the breast lumps [46]. These can be particularly useful for younger women with denser breast tissue, and for the detection of very small lesions that are missed on the mammograms because of their lower resolution.

Nevertheless, even though least stressful and non-invasive, this type of examination is not able to confirm whether there are cancerous cells or not, or whether the lump seen on the image is benign or malign. Therefore, the biopsy method is currently considered the only way to confirm the presence of cancer with higher assurance, but also the only method that is able to differentiate among sub-types of tumors.

4.2 Histological Images Analysis

Histological images are the most accurate resource of data for cancer detection. Moreover, manual analysis that is nowadays done in the laboratories by experts, experienced pathologists, is time-consuming, and besides that, it has problems of inter and intra observer discordance [5]. Therefore, Computer Aided Diagnosis (CAD) systems, that can automatically process histopathological images, are going to accelerate and standardize the analysis, and improve the outcomes. In the next

sections, we go through some of the ways to automate these analyses through ML methods proposed by researchers.

4.2.1 Shallow methods

Shallow methods aim to classify the images based on the features extracted from them. That is why feature engineering is a very important step of these methods since it is used as the first step to reduce the number of dimensions and to provide relevant information related to the task. Features are usually based on histograms, textures, or counts of different elements, and they require specialized domain knowledge. They also require specific image preprocessing which includes different filters and transforms to exaggerate particular characteristics of the image. For example, one of the works by Marugame et al. is using morphological operations and Gabor wavelet transform on the dataset. They report a 66% true classification rate using Bayes classifier on three-class classification (cancerous, normal/benign, and precancerous) [6]. The other work by Chen et al. proposed a method based on a pixel-wise Support Vector Machine (SVM) to differentiate tumor nests from stroma, and a marker-controlled watershed algorithm for nuclei segmentation [47].

One popular shallow method for classification in the literature is Random Forest (RF). For example, Basavanhally et al. use a different patch sizes procedure for the Whole Slide Image (WSI) from which they extract morphological, textural, or graph-based features. They analyze which features are more informative for different sizes of patches and associate them with them. Afterward, they use RF classifier on aggregated multiple field of view patches to distinguish between tumor grades in breast cancer from histopathology tissue images. They classify low versus high, low versus intermediate, and intermediate versus high grade and obtain an area under curve values of 0.93, 0.72, and 0.74, respectively [48]. One more paper employing RFs, as well as SVMs and some

other shallow methods like K-nearest neighbors and Logistic regression, is reporting high accuracy for the detection of cancer in WSI: AUC = 0.97-0.98 for tumor detection within the whole image area, AUC = 0.84-0.91 for tumor vs. normal tissue classification. The features extracted from the WSIs include information about texture, spatial structure, and distribution of nuclei. The authors of this work (Valkonen et al.) claim that they provide explainability to the results and the way in which the algorithm makes decisions based on the given features. The method was evaluated in breast cancer metastasis detection from lymph node samples [7].

The most important shortcoming of these works is their low efficiency together with high complexity due to the huge amount of features. They require specific expert knowledge which is not easy to obtain for such a study, and even harder without a personal bias. Moreover, in general, they are not able to scale well.

4.2.2 Deep Learning methods

DL methods are very powerful when it comes to classification, automatic feature extraction, or retrieving information from the larage datasets. They do not require pathology domain knowledge and they are able to generalize well.

One of the first works to employ DL in histopathological image classification was done by Cruz-Roa et al. in 2014. They employed a custom 3-layer CNN to classify patches obtained from WSIs. The authors report an F-measure, and balanced accuracy of 71.80% and 84.23%, respectively, when classifying patches in two classes - one containing Invasive Ductal Carcinoma (IDC), and one being healthy [8]. The same dataset is also used in the work by Reza et al. from 2018, who, similarly to Cruz-Roa et al. used a custom 3-layer CNN, but in addition, tried to balance the classes in the dataset with different techniques like Synthetic Minor-

ity Over-sampling Technique (SMOTE). They showed an improvement in the performance, with respect to earlier works, obtaining an F1 score of 84.78% and a balanced accuracy of 85.48% [49]. Later works started to rely more on findings from computer science, and other specialized domains where distinctive NN architectures showed impressive performances. The dataset presented by Cruz-Roa et al. is further exploited in the work by Romero et al. from 2019, which used multilevel batch normalization with Inception network [50], and obtained a balanced accuracy of 89% and an F1 score of 90% [51]. One more work utilizing the same dataset is presented by Celik et al. from 2019, who obtained an F1 score 94.11% and balanced accuracy value of 90.96% using Residual Neural Networks (ResNets) [52], and an F1 score of 92.38% and a balanced accuracy value of 91.57% utilizing Densely connected Neural Networks (DenseNets) [53] on their best trained models [54].

Another work, done by Bejnordi et al. in 2017. proposed a system, using VGG networks [55], for classification of breast biopsy WSI. They used two VGG-like networks, one for classifying WSI into epithelium, fat, and stroma, and the second one to distinguish between cancerous and healthy parts of stromal regions. Finally, they extracted two sets of features from both network outputs and used RF in order to perform the classification of WSI into healthy or cancerous class, obtaining an area under the curve (ROC) of 92% [56].

At this point, we are going to mention two big challenges in histopathology image analysis which are moving the benchmarks in the field, as many teams are competing over the same data.

The first one is BASH [9]: Grand challenge on breast cancer histology images which was conducted with the 15th International Conference on Image Analysis in 2018. It aimed at the classification and localization of different classes in microscopy and WSIs. Different research groups showed that CNNs are the most successful in the challenge. The most

outstanding work was done by Aresta et al. who obtained an accuracy of 87% using ensembles of DenseNets [53] and ResNets [52] on four class-classification of microscopy images. Ensembles are sets of classifiers that have their results combined to provide predictions for the selected problem. Separately, classifiers usually do not perform very well on the task, but, if the errors are uncorrelated, their combination can produce overall satisfying results. Ensemble results can be obtained in several ways, but the most common ones are majority voting or taking the median or average of the produced predictions by separate classifiers.

The second challenge is the CAMELYON Challenge [10] whose goal is to develop algorithms to detect cancer metastasis in lymph node WSI. The winner of this challenge in 2016 was the group of Wang et al. They obtained an AUC of 0.925 for classification of WSI in metastatic breast cancer or healthy one [57]. They used four well known CNN architectures for this task: AlexNet [58], VGG [55], GoogleNet [59], and FaceNet [60].

All of these papers and challenges are showing that DL methods are very powerful tools for obtaining competitive results on the classification of anomalies like cancer in histopathological images without the assistance of experts in pathology. They showed that they can be a relevant and reliable part of CAD systems, and reduce the manual labor done by the doctors. Moreover, computer vision and DL algorithms are the center of attention in today's computer science world thus, new tools and methods are proposed on an everyday basis and they can be used for solving histopathological automation tasks.

4.3 Data availability and evaluation challenges

Even though the number of papers in this research domain is rapidly increasing, we should underline some of the limitations on data availability, and the evaluation process of the presented algorithms.

First of all, we should note that different works are using a differ-

ent type of images in datasets. In particular, there are two types of datasets containing microscopy images and WSI, respectively. The former are usually smaller in size, which is in the order of 1k pixels in width and height. The labels are usually image-based, and not region-based, which means that all the patches from one image are annotated with the same ground truth, even though they might belong to a different class. Datasets that contain this type of image are BreakHis [61], and BASH challenge dataset for classification. The latter type containing WSIs, are the result of digitization of entire slides without loss of biopsy tissue. They provide better resolution for the pathology assessment. The size of these images is in giga-pixels and that is why the automatic analysis of these images requires a lot of computational resources, meaning that they are very memory consuming and computationally expensive. Here we usually deal only with regions of interest, that are divided into smaller patches, and the rest of the image is discarded. A dataset that contains WSI is, for example, the one used in the CAMELYON challenge and in the work done by Cruz-Roa et al. mentioned in the previous Section.

Additionally, most of the datasets contain only hundreds of images, which brings bias to the obtained results. That is not enough to propose generalized solutions, and all the results should be taken cautiously. Multiple papers analyzed in the previous Section, propose and validate solutions based on private datasets which are partly, or even not at all publicly available, limiting the reproducibility of the results. Moreover, many authors omit specific details of the implementation. When it comes to DL they usually specify the type of architecture used, but they rarely go into details on specific hyperparameters, like batch size or learning rates, or adaptations of last layers for classification used for the implementation. They never go into details about data augmentation and the ranges used for the transformations done on the images. Additionally, when it comes to the training of the models, they do not show the graphs

of the training accuracy history. None of the works reports the results on cross-validation nor they explain how they did the split of data in training and validation datasets. Finally, and maybe most importantly, they almost never specify which computational resources they had, and for how long they trained the NN, and this is very important for this type of research especially with WSIs. In the WSI case computations can never be done locally, and additional servers with multiple GPUs are required. That is why all the results from the related works should be considered carefully when trying to compare the obtained results and assess the reliability of the methodology.

These are all the limitations and shortcomings of the related works which need to be resolved to reach better and reliable results before they can be considered as real benchmarks in the field.

4.4 Summary

This section gave a brief overview of the trends in the field of breast cancer detection, the results, and the main drawbacks of the literature.

Since the imaging exams are the first step in breast cancer detection in medicine, we decided to dedicate the first part of the related works to the results in this field. As we saw, many works are successfully able to segment and classify masses as benign or malign which is crucial for the rest of the procedure. In fact, mammography is usually the first exam in breast cancer assessment and based on its results the patient is selected for the biopsy or other procedures.

From here we move to histology images analysis mentioning the solutions that used shallow as well as DL algorithms for classification. It is very important to understand which is better for the task and which of the two approaches is going to thrive in the future because of research in other biomedical and computer science domains. Therefore, even though results of the works using shallow methods are producing comparable re-

sults, the space for research in DL domain is definitely larger and more prosperous because of the increase of available data. Furthermore, DL methods are going towards models that are able to generalize better in different tasks and provide the framework for the fusion of different histopathological tasks in the future.

Finally, we addressed the problem of repeatability and validation which are important for the possible deployment of these systems in the medical practice.

Methodology and Implementation 5

This Chapter explains the chosen methodology, its general concepts, as well as the implementation details which are used for the development of the BRAVE AI pipeline proposed in the thesis. Firstly, in Section 5.1 we give a complete overview of the pipeline followed, while the rest of the sections are going into details with different parts of it. Section 5.2 gives information about the dataset and provides its analysis which is influencing other implementation details in further work. Section 5.3 shows how to address the problems which can occur during training deep NNs. Section 5.4 provides information about particular deep NN architectures used in the work, and their general overview along with the modifications and training details. Finally, in Section 5.5 we explain the evaluation metrics used to analyze the performance of the pipeline.

5.1 BRAVE AI Pipeline Overwiev

This Section provides a general description of the methodology employed in the thesis, with all the steps performed to obtain the final results. The main goal of this section is to provide a general understanding of the framework before going into details regarding the constituting steps of the methodology and implementation.

The pipeline employed is a Machine Learning (ML) one, and it can be divided into three distinctive parts as shown in Figure 5.1. The first one is the dataset analysis and preparation. We are working with a dataset consisting of patches coming from Whole Slide Images (WSIs), which need to be loaded and stored for further analysis. Afterward,

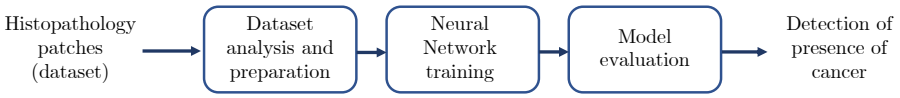


Figure 5.1: BRAVE AI pipeline proposal.

we remove the patches which do not contain useful information about the presence of Invasive Ductal Carcinoma (IDC) on the image, in fact, they are providing no useful information about the content of the WSI. Finally in this part, we provide basic statistics about the dataset, about the number of patches per WSI, and the percentage of those containing the cancerous tissue. This is giving insights into the way the dataset is balanced and its dimensions, and how to handle this in the following parts of the pipeline. Next, we move forward to methods for reducing the overfitting on the training dataset.

After the dataset analysis and preprocessing, we proceed with the part of the pipeline related to choosing and modifying the Neural Network (NN) architecture which is well suited for the imposed problem: decision making in histopathology of breast cancer. After careful reasoning on Deep Learning (DL) models, we are proposing two architectures: Residual Neural Network (ResNet), and Densely connected Neural Network (DenseNet) which are among the most popular architectures for image classification. They are very deep structures with additional connections for the propagation of information that are not suffering from training problems typical for deep architectures, and, at the same time they are more accurate and easier to train. Their details are provided in the following sections. The output of the NN is the probability that the image patch is IDC positive, i.e. that it contains cancerous tissue.

Finally, the last part is the model evaluation. We carefully reason on which evaluation metrics should be used in medical problems, and

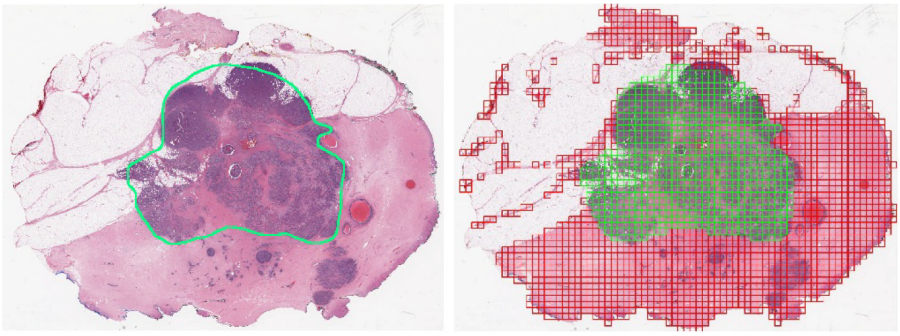
explain the differences and common misconceptions about them. This is highly challenging because of the nature and characteristics of the dataset. Therefore, giving an estimation on how good the model is, and how sure about the results we are, is crucial for medicine, since the treatment and handling of the disease can be highly affected by this.

5.2 Dataset

This section provides information about the dataset used in the thesis. Having a dataset which is reflecting well the problem in the real world is crucial for ML models because they are using the data to learn and memorize the information which is used for making future predictions. Therefore, the dataset must contain information about the events or objects of interest, reflect its true nature, and as many of its variants as possible. Even though having an infinite dataset seems like the perfect solution, it is not because of the computational power and memory resources that are available for training the model. In Section 5.2.1 we provide basic information, while Section 5.2.2 explains the preprocessing, data cleaning, and statistics about the dataset which are influencing further decisions in the work.

5.2.1 Dataset General Information

The work of the thesis relies on the dataset used by Cruz-Roa et al. [8]. The dataset originally comes from the University of Pennsylvania and The Cancer Institute of New Jersey, and it consists of WSIs of IDC, from 162 women. All slides were digitized via a whole-slide scanner at 40x magnification with the resolution of 0.25 µm/pixel. Ground truth annotations of the cancerous regions in the image were provided by a pathologist. They used images with 2x magnification in order to decrease the time needed to provide annotations, with the cost in precision. This



(a) (b)

Figure 5.2: Figure (a) shows ground truth annotation. On Figure (b) is shown result of patching of WSI. Red patches correspond to positive examples of IDC and green patches correspond to negative examples.

means that some stromal and non-invasive tissue near the IDC region is included in the region that is labeled as positive, even though it is not. In the original paper by Cruz et al., they used non-overlapping patches of 100x100 pixels size, while in this work we used 50x50 pixels images. The dataset consists of 277,524 RGB patches (198,738 IDC negative and 78,786 IDC positive). Patches that contain only fat tissue or background are discarded from the dataset. In Figure 5.2a can be seen a WSI with pathologist's annotation, and in Figure 5.2b can be seen the result of patching after discarding irrelevant patches.

This dataset has been used in the state of the art in several works [8], [51], [54], [49]. Furthermore, this dataset has a sufficient amount of WSIs which are patched and labeled. This is very important because of the size of WSIs that is originally in gigapixels (around 10^{10} pixels) which makes them extremely computationally expensive to work with and even their storage represents a problem since they can occupy a couple of terabytes of hard disk space. Moreover, this dataset is increased in size and now consists of 279 WSIs and can be found open-sourced on Kaggle [1].

[1] https://www.kaggle.com/paultimothymooney/breast-histopathology-images

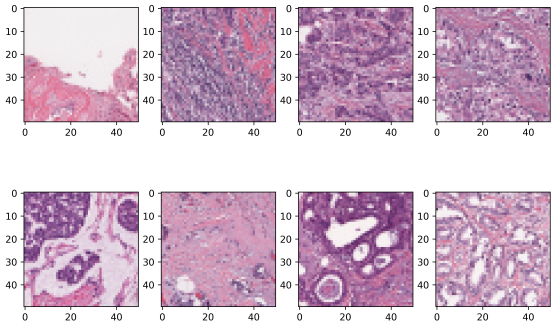


Figure 5.3: Examples of IDC positive patches.

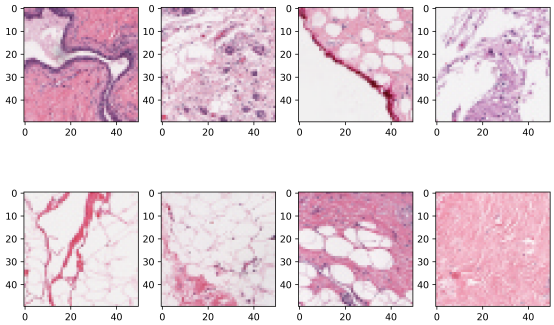


Figure 5.4: Examples of IDC negative patches.

5.2.2 Dataset Analysis and Preprocessing

Since the approach, we are taking in this work is exclusively a DL one, the exhaustive preprocessing of the dataset is not required, nor extracting any kind of features. The whole images are passed as inputs to the NN and the network is extracting features and important information by itself through its layers of neurons. The original dataset has 277,524 images of 50x50 pixels from which 78,786 were IDC positive and 198,738 were IDC negative. This distribution makes the dataset unbalanced and this has consequences for the training of the NN. The examples of patches belonging to the class of IDC positive are shown in Figure 5.3, and the ones belonging to class of IDC negative are shown on in Figure

```
Data: dataset
Result: dataset without non-informative patches
for t=1,...,number of patches do
    image = dataset(t);
    if image shape is different than (50,50,3) then
        remove image from dataset;
    end
    for r=1,...,number of channels do
        add to count1 pixels in channel r with value <
         threshold_min;
        add to count2 pixels in channel r with value >
         threshold_max;
    end
    if count1 or count2 > threshold_pix then
        remove image from dataset;
    end
end
```

Algorithm 1: Pseudocode for removing the non-informative image patches from the dataset

5.4. Every image name consists of the patient ID, the coordinates where the patch is located in the WSI, and the class to which the patch belongs. An example of file name is $9255\_idx5\_x401\_y851\_class0$ where $9255idx5$ is the unique patient ID, $x401$ and $y851$ are coordinates and $class0$ is putting image patch in class of IDC negative samples.

Since the dataset is annotated coarsely as we explained in Section 5.2, we removed from the dataset images which are not exactly 50x50 pixels, and images that are mostly black or white. Algorithm 1 shows how the patches were removed. Thresholds for pixel values were chosen as values that are 10% higher or smaller than the minimal and maximal value of the pixel range ($[0, 255]$). We then count the number of pixels that are out of the restricted range for pixel values. The count threshold needed to remove the image from the dataset was set to 80% of the number of pixels in the image for all of the three channels together ($6,000$ pixels). The number of patches after the removal of the non-informative ones is

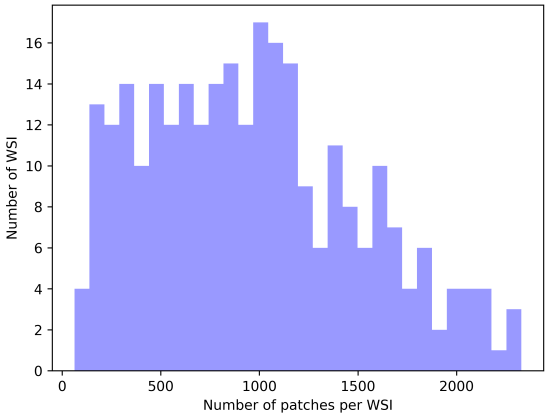


Figure 5.5: Number of patches from all classes per WSI.

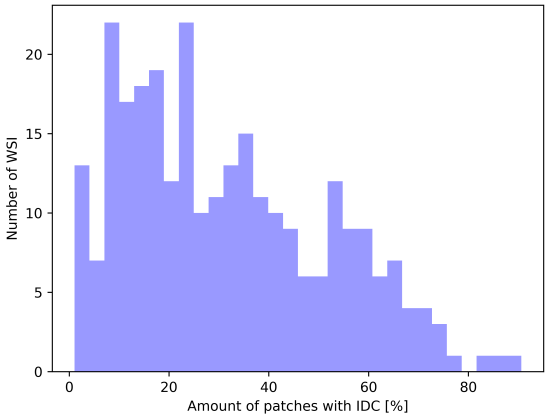


Figure 5.6: Percent of patches with IDC positive class per WSI.

270, 544, which means that we have removed 6, 980 patches (2.52% of all the patches). After the removal, we have 192, 004 of healthy patches and 78, 540 belonging to cancerous regions. From Figure 5.5 we can see that most of the images have from 0 to 2, 000 patches. From Figure 5.6 we observe that the WSIs contain mostly non IDC tissue which is also reflected in the class imbalance of the dataset.

5.3 Reducing Overfitting

As we saw in the previous Sections, the dataset that we use does not have a large amount of WSI and deep NNs are hard to train from scratch with a small amount of data. In these cases, they are prone to overfitting since the capacity of the NN is many times bigger than the variance carried by the dataset and thus it performs poorly on unseen data. But we can at least partially solve this problem using two ML approaches:

- Transfer Learning. In these cases, we approach the problem by using an already trained NN on a bigger dataset like ImageNet [62] to extract the features and then use them with another shallow classifier. This basically means that the network is using already learned filters, and we are relying on the fact that simpler filters are equally important in both datasets. This process is also called fine-tuning. We are explaining how we do this in Chapter 6.

- Data Augmentation. This technique means that the images are transformed using affine transformations like rotating, flipping, shearing, etc. In this way, we are adding noise to the dataset and the NN will learn how to generalize better on images and reduce overfitting. Another way to do data augmentation is to patch the images. This is especially the case with WSIs that can have thousands of pixels and different parts of the image can look very different and even belong to different classes.

At this point, we are going to provide information about the augmentation that we have done on the dataset. We apply horizontal and vertical flips, and color jittering with values of 0.05 for brightness, hue, and saturation. Finally, we also do normalization to mean and standard deviation for the three channels as follows: $[0.485, 0.456, 0.406]$, $[0.229, 0.224, 0.225]$ because of the requirements for the pretrained models. As we are working with patches, and not looking at the WSI of

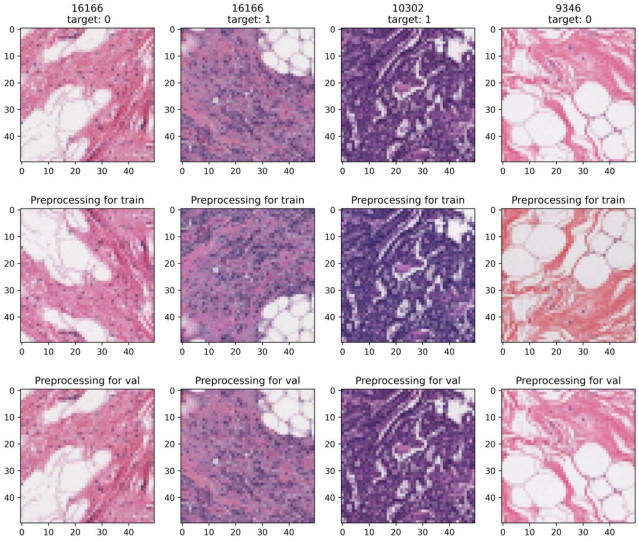


Figure 5.7: Examples of augmentation of different patches.

tissue, we are not losing spatial connections among patches. Additionally, all of the mentioned transformations are creating realistic examples of histopathology patches because cells and tissues can also be stretched and vary in size. The examples of augmentations can be seen in Figure 5.7. The augmentation is done only for the training part of the dataset, and not for the validation considering that the model needs to be validated on the same dataset every time to make decisions about the model and its parameters.

Moreover, it should be noted that the traditional pipeline for augmentations assumes doing the augmentations before continuing with the classifier, whereas our framework does augmentations on the fly. So, instead of showing the exact same item at every epoch, there is a variant that has been changed in a different way each time. So after three epochs, the model has seen three random variants of each item in a dataset.

5.4 Deep Neural Networks

It is known that simple NN with only one wide layer can learn to represent any function given enough training data. But the issue arises with increasing the number of neurons in the same layer because the algorithm becomes prone to overfitting and its generalization error is going to increase. In fact, we may say that such networks are very good at memorization, and not at generalization. This means that the network will perform very well on already seen samples, but with the new ones will completely fail. Therefore, the networks needed to become deeper, i.e. to have more hidden layers.

In theory, as you make a NN deeper, it should only do better and better on the training set. Empirically, as the number of layers increases, the training error will tend to decrease, but after a while, it will start to increase. This is the problem of vanishing or exploding gradients that can arise during backpropagation [63]. This is due to the fact that each of the weights in the NN receives an update proportional to the partial derivative of the error function with respect to the current weight in each iteration of training. If the gradient at the output of the network is small its propagation backward is going to make the error smaller and smaller as we are reaching the layers closer to the input, and in extreme cases, the updates of those layers are going to be negligible. This means that the weights in the NN are not changing therefore, the NN is not learning. The opposite thing may happen if the gradients are too large: they become larger and larger until the weight update goes to infinity. In both cases, the updates become meaningless and the NN is not able to learn the patterns.

Moreover, this is not the only problem that very deep networks may have. If there are many layers the training may be very computationally expensive and need a lot of memory, or just take too long to train. Nevertheless, deep NNs are a very powerful tool and that is why the

Artificial Intelligence (AI) community finds deep NNs very interesting as a research domain and produces many new architectures as solutions to the problems mentioned above. In the next section we are going to explain the ones used in this work.

5.4.1 Residual Neural Network - ResNet

Inspired by the pyramidal cells from the cerebral cortex, comes the idea of ResNets, which can facilitate solving some of the mentioned problems and make the NN deeper. The architecture of ResNets introduces the concept called Residual Learning. This network uses a method called skip connections or shortcuts, which skips training from a few layers and connects to the neurons in some of the subsequent layers. This work was presented in 2015 by He et al. [52] and it was a major breakthrough in training very deep NNs.

Figure 5.8 represents a small part of a simple feedforward NN where l represents the layer, and $a^{[l]}$ the activation in that layer.

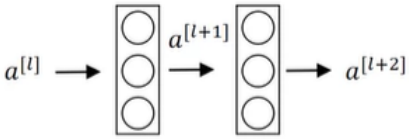


Figure 5.8: Simple feedforward NN block.

Starting from activation $a^{[l]}$ in the layer l we firstly apply linear operations as written in matrix Equation 5.1 where W represents the weights of the connections and b are the biases. The result of the linear operation z is now the input to the non-linear function $g()$ which is the activation function in the layer $[l + 1]$. The analogue operations are applied in the following layer. The mathematical equations ruling this process are:

$$z^{[l+1]} = W^{[l+1]}a^{[l]} + b^{[l+1]} \tag{5.1}$$

$$a^{[l+1]} = g\left(z^{[l+1]}\right) \tag{5.2}$$

$$z^{[l+2]} = W^{[l+2]}a^{[l+1]} + b^{[l+2]} \tag{5.3}$$

$$a^{[l+2]} = g\left(z^{[l+2]}\right) \tag{5.4}$$

On the other hand, in ResNet, except this main flow of information, we add new connections that are skipping some of the layers. The authors of [52] also call them identity mappings. This can be seen in Figure 5.9. These connections are not adding computational complexity, nor new hyperparameters to tune and the backpropagation training algorithm for the NN stays the same.

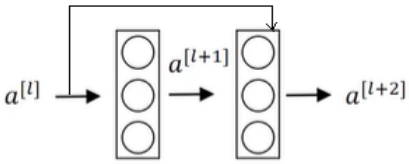


Figure 5.9: Residual block.

Mathematically, there is only one change in the flow of data in Equations 5.1 - 5.4 and it comes before applying non-linear function in Equation 5.4:

$$a^{[l+2]} = g\left(z^{[l+2]} + a^{[l]}\right) \tag{5.5}$$

It should be noted that, in the mathematical equations we provided, the shortcut skips only one layer whereas in chosen architecture we can skip as many layers as we want. If we look at the corner case where

the weights tend to go to 0, and with it $z^{[l+2]}$ goes to 0 the result of the non-linear function is going to be at least equal to the activation of the layer l which is $a^{[l]}$. This means that it is very easy for ResNet to learn identity mappings. Therefore, by adding more residual blocks, the NN is going to perform at least the same as without them, but in many more cases, the deeper layers will be able to learn useful mappings from the data.

5.4.2 ResNet18 implementation

We started the research with a smaller and simpler ResNet model with 18 layers. The activation functions and the max/average pooling layers are not counted in the 18 layers. The architecture starts with a 7x7 kernel size and 64 filters. Here begins the first residual block. Then we have 2 convolutions with kernel size 3x3 and 64 feature maps. This is where the residual block ends and the feature maps are added together to the next block. The same residual block is repeated once again. Then the network continues with 3 more stages, and each of them having two residual layers with 2 convolutions. All filters have a 3x3 kernel size, but the number of filters is increasing twice in every following stage. The network is ending with a fully-connected layer and the softmax activation. We are using already validated models implemented in the PyTorch. In our implementation, we are removing the last fully connected layer and adding 3 fully connected layers with 512, 256, and 128 neurons respectively ending with the softmax for classification. After the Rectified Linear Unit (ReLu) activation in each layer, we are adding batch normalization and dropout of 0.2.

5.4.3 ResNet50 implementation

In the second experiment, we have increased the capacity of the ResNet and implemented ResNet50, a residual network that has 50 layers di-

vided into several stages. The activation functions and the max/average pooling layers are not counted in the 50 layers. It should be noted that this implementation has shortcut connections that are not skipping two layers like ResNet18, but three layers since we are adding 1x1 convolutions. In the first layer, we perform the initial convolution using 7×7 kernel sizes and 64 filters, which is the same as in ResNet18. Next, we proceed with a stage that has three residual blocks with the number of filters being 64, 64, and 256 which is repeated 3 times adding 9 more layers. Every next stage of the network is doubling the number of feature maps in every residual block of the stage. Residual block is repeated 4, 6, and 3 times in the next three stages (2, 3, 4) respectively. Initially, the ResNet50 as the last, 50th layer, has a fully connected layer with the softmax for classification. We are using already validated models implemented in the PyTorch. In our implementation, as in ResNet18, we are removing the last fully connected layer and adding 3 fully connected layers with 512, 256, and 128, respectively, ending with the softmax for classification. After the ReLu activation in each layer, we are adding batch normalization and a dropout of 0.2.

5.4.4 Densely Connected Neural Network - DenseNet

One of the more recent solutions to making the NNs deeper are DenseNets introduced by Huang et al. in 2017 [53]. They are exploiting the idea that deep NNs are more efficient to train if they contain shorter connections among the layers close to the input and those close to the output. In order to achieve this, each layer in the dense block connects to every other layer in a feed-forward fashion. Figure 5.10 shows one example of 5-layer dense block.

Traditional Convolutional Neural Network (CNN) with L layers have L connections while DenseNet has $L(L+1)/2$ direct connections in one DenseNet block. Moreover, in contrast to ResNets where we sum acti-

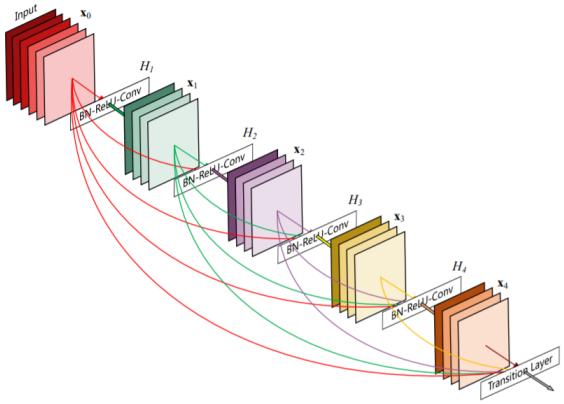


Figure 5.10: A 5-layer dense block.

vation of the previous layer before the non-linear function, in DenseNets the activations from previous layer are concatenated before being passed into the following layer as written in Equation 5.6. This increases the variation of the inputs and passes forward all the knowledge from the previous layers thus it is increasing efficiency. Mathematically, the equations we had for traditional NN 5.1 - 5.4 need to be changed with the following ones:

$$A^{[l]} = [a^{[0]}a^{[1]}...a^{[l]}] \tag{5.6}$$

$$z^{[l+1]} = W^{[l+1]}A^{[l]} + b^{[l+1]} \tag{5.7}$$

$$a^{[l+1]} = g\left(z^{[l+1]}\right) \tag{5.8}$$

The advantages of DenseNet are: alleviation of the vanishing-gradient problem, improving feature propagation and increasing feature reuse, and reducing the number of parameters that need to be learned. Since each layer receives feature maps from all of the preceding ones with features with different complexity levels, so the knowledge is passed more

efficiently and the network can be smaller and more compact. Furthermore, dense connections have the effect of regularization, which reduces overfitting when training with smaller datasets. Additionally, since each layer has direct access to the gradients from the loss and the original input signal, it can be interpreted as deep supervision. Deeply supervised CNNs have classifiers attached to intermediate hidden layers which are forcing them to learn more discriminative features [64], while in DenseNet we have only one loss function which provides direct supervision to many intermediate layers not just to the last one.

5.4.5 DenseNet121 implementation

In order to further increase the capacity of the model, we have implemented a DenseNet with 121 layers. The architecture starts with 5 convolutonal layers and it is followed by 4 dense blocks. Between each of the dense blocks there is a transition block which is used to downsample the feature maps using the pooling operation. The first dense block consists of 1x1 and 3x3 convolutins, resulting in 32 feature maps, which are repeated six times. Every next dense block has convolutions of the same size repeated 12, 24, and in the last layer 16 times. At the end there is one fully connected layer with the softmax for classification. We are using already validated models implemented in PyTorch. In our implementation, we are removing the last fully connected layer and adding 3 new fully connected layers with 512, 256 and 128 neurons, respectively, ending with the softmax for classification. After the ReLu activation in each layer we are adding batch normalization and a dropout of 0.2.

5.4.6 Training

Training a NN means finding the appropriate weights of the NN connections in the feedback loop thanks to the algorithm called backpropagation. The loop starts with presenting the model with the inputs which

```
for each epoch do
    for each training iteration do
        take batch of images and corresponding labels from
         training dataset as input;
        forward propagate the images and obtain the output;
        calculate weighted cross entropy loss;
        calculate and save training accuracy and loss values;
        compute the gradient penalties;
        propagate the gradients backward and update the weights;
        update learning rate ;
        if number of training iterations passed > threshold then
            for each batch in validation dataset do
                forward propagate the batch images and obtain the
                 output;
                calculate and save cross entropy loss;
            end
            calculate accuracy on whole validation dataset;
            calculate average loss on whole validation dataset;
        end
    end
end
```

Algorithm 2: Pseudocode for training of the NN and obtaining accuracy and loss on training and validation datasets.

are propagated in feedforward fashion through the NN in order to get the prediction. In our case it is the probability of the output class, the one of healthy tissue and the other with IDC. After a batch of images is propagated we can calculate the error function, commonly known as loss of the NN.

Since the dataset used in the thesis is imbalanced, as shown in Section 5.2.2 we need to address this problem in the training in order not to overfit on one of the classes. That is why we implemented the weighted cross-entropy as the loss function as follows:

$$J_w = -\frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M} t_k \times y_m^k \times \log\left(h_\omega\left(x_m, k\right)\right) \tag{5.9}$$

where M is the number of training examples, K is the number of classes, t_k is the weight for class k, y_m^k is the target label for training example m for the class k, x_m is the input for training example m, and h_ω is the model with NN weights ω. The class weights are calculated as the negative logarithm of fraction of specific class elements in all elements in the dataset. Therefore, the class which has higher occurence will have lower weight and vice versa. Minimizing the loss function is the final goal of the training, and the way we update the weights and learning rate is defined by the optimizer we use. Training is performed in a loop epoch by epoch as shown in Algorithm 2.

5.5 Evaluation

Evaluation of the model requires the assessment of its ability to correctly predict the output value for any given input. How the model generalizes on unseen data is a highly important part of a ML project. Therefore, the estimation of the model performance should be as realistic as possible. In the Sections 5.5.1 and 5.5.2, we explain two methods used to evaluate the proposed model namely holdout method, and cross-validation. Moreover, a model can produce satisfying results when evaluated using one metric but may perform poorly when evaluated with another. Therefore, it is very important to choose the metrics wisely depending on a given problem. It should be also noted that different metrics are used to evaluate classification and regression models, or any other which goal is segmentation or natural language processing. In the Sections 5.5.3 and 5.5.4 we go into details for the metrics used in classification.

5.5.1 Holdout method

The holdout method is a way to estimate how well the model performs on unseen data using only the dataset that we have for developing the

model itself. In this method, depending on how much data we have, the dataset is randomly split into two or three subsets, namely training, development (validation), and test set. The fraction is usually 80 : 20 for two and 80 : 10 : 10 for three datasets.

The training dataset is used to fit the model, i.e. update weights and biases until the loss function is minimized. Therefore, the model learns from this data. A validation dataset is used to assess the model performance on unseen data, providing its unbiased estimation. It must not contain the data from the training dataset. Even though the model itself is not learning from this data, we use it to choose the model hyperparameters like the number of layers or neurons in each layer. Therefore validation dataset is affecting the model indirectly. Finally, the test set provides the result of the model evaluation. It is used only once at the end of the training of the model and it should be a well-made reflection of the real-world data. Moreover, if having a dataset like this is not possible it is better to divide the original one only into training and validation datasets. The holdout method gives, as the final performance of the model, the assessment obtained on the validation dataset.

If a model fits the training set much better than it fits the test set, overfitting is probably the cause. It means that it has memorized examples from the training dataset, without learning how to generalize on the unseen data. If it is the opposite, the model underfits, which means that the model does not have enough capacity, or that the validation/test dataset is easier to learn than the training dataset.

The holdout method is very simple, and quick to execute, but its main disandvantage is high variability depending on the dataset train/-validation/test split and possible bias for the estimation.

5.5.2 K- fold cross-validation

Another, more reliable method to estimate the model performance is k-fold cross-validation. The dataset is, as in holdout method, divided in training and validation subset but this divison is done multiple times in order to reduce bias of the estimation. K-fold cross-validation involves partitioning the original dataset into k subsets of equal size which are called folds. The model is then trained on dataset formed from $k-1$ folds that are combined together, and validated of the one that is left. This is repeated k times, so that each of the k folds is used exactly once as validation dataset. The performance estimation is then averaged over all k trials to obtain the final unbiased performance assessment. The number of folds k is usually chosen to be at least 5. It should be noted that cross-validation is not giving the information of which model exactly is the best, but how well that class of models is able to fit that data. Moreover, it is computationaly very expensive since we have to train the model k times.

5.5.3 Confusion matrix and standard metrics

Almost all commonly used measures of a model's quality come from a table known as confusion matrix or matrix of errors. Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class (or vice versa). A model is perfect if this matrix is diagonal, which means that it is not making any errors on the given data. Non-diagonal elements are errors. In case of binary classification the dimension of the matrix is $2x2$ and the classes are classified as positive and negative one. Following that the generic confusion matrix is shown in Table 5.1, and every cell from the table is defining a specific concept.

True positives (TP) are positive instances that are correctly classified by the model as positives. True negatives (TN) are negative instances

Table 5.1: Generic confusion matrix.

		Predicted Value	
		Positive	Negative
Actual Value	Positive	TP	FN
	Negative	FP	TN

correctly classified as negatives. False negatives (FN) are instances from the positive class missclassified by the model as negative. False positives (FP) are negative instances that are predicted to be positive by the classifier. In our case, positives are the images that contain IDC, and negatives are the images with the healthy tissues in it. Based on these four terms we can define all of the metrics for the evaluation of a binary classifier.

The first metric that is coming from the confusion matrix is the accuracy. It represents the fraction of correctly classified instances in the whole dataset, and it is defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.10}$$

Accuracy is the most used metric in classification, but it should be noted that it is not highly informative when classes are not balanced. For example, if one class has 99% of all the samples in the dataset, and the other class only 1% the accuracy of 99% can be achieved wiht a classifier that puts all instances in the first class. Furthermore, there are metrics that are not using all of the information from the confusion matrix, but are relevant for the problem like precision, sensitivity (recall) and specificity. Precision is the fraction of true positive instances among the positive instances (Eq. 5.11). On the other hand, sensitivity is the fraction of true positives among all instances that belong to the positive class of the dataset (Eq. 5.12), and specificity is the fraction of true negatives among all instances that belong to the negative class of the dataset (Eq. 5.13). Sensitivity and specificity are very popular in

medical fields as ours. If we assume that having a condition means belonging to a positive class, a test that is highly sensitive will flag almost everyone who has the disease and will not generate many false-negative results, whereas highly specific model will flag correctly almost all true negatives and will generate some false-positive results. These two are connected and increasing one is often at the cost of the other. Furthermore, when in the field of medicine, very high sensitivity of classifier is desired because we want to minimize false negatives. In other words, we do not want to treat the person as healthy if the person is not because it can lead to progressing of the cancers or tumors into higher stadiums and eventually lead to death.

$$Prec = \frac{TP}{TP + FP} \tag{5.11}$$

$$Se = \frac{TP}{TP + FN} \tag{5.12}$$

$$Sp = \frac{TN}{TN + FP} \tag{5.13}$$

It should be noted here that using only one of the metrics is usually not relevant because it can lead to missleading results. For example, if we classify all instances as positive ones sensitivity will be 1. On the other hand, if we do the same for true negatives specificity will be 1. In order to get more relevant results we sometimes use metrics that are combining them, like F1 and balanced accuracy which are defined in Equations 5.14 and 5.15, respectively. Balanced accuracy is good for imbalanced classes.

$$F1 = 2\frac{Prec \times Se}{Prec + Se} \tag{5.14}$$

$$Acc_b = \frac{Se + Sp}{2} \tag{5.15}$$

It should be noted here what is the difference between the loss function and the metric in general. Loss functions are showing a measure of the model performance, and they are used to train a ML model. They should be differentiable with respect to the parameters of the model. On the other hand, metrics are used to monitor and measure the performance of a model both during training and testing, and they are not used to update the model's weights.

5.5.4 Reciever Operating Characteristic and Precision-Recall curve

The Section 5.5.3 explained how the confusion matrix summarizes all of the possible conditions of a binary classification task. What should be noted here is that, usually, the output of the model is the probability that a sample belongs to a class. In order to put a sample in one or the other class we have to choose a cut-off threshold. The results change when the threshold is changed, thus the performance with varying threshold should be estimated. Threshold-invariant metrics, such as area under the curve (AUC), are capable of measuring the overall model performance despite any chosen threshold. There are two commonly used curve which we are going to explain here - reciever operating characteristic (ROC) and precision-recall (PR) curve.

ROC curve reports $Sensitivity$ on y-axis vs. $(1 - Specificity)$, also known as false positive rate (FPR), on x-axis calculated for different thresholds of classification. A ROC graph depicts relative trade-off between benefits which are TP, and costs which are FP. There are several points in the ROC space that are important to note. The lower left point (0, 0) means that the classifier is never giving a positive class as a result, thus it has neither FP errors nor TP instances. The opposite classifier is represented by the upper right point (1, 1) which is always predicting that sample belongs to positive class. The point (0, 1) represents perfect

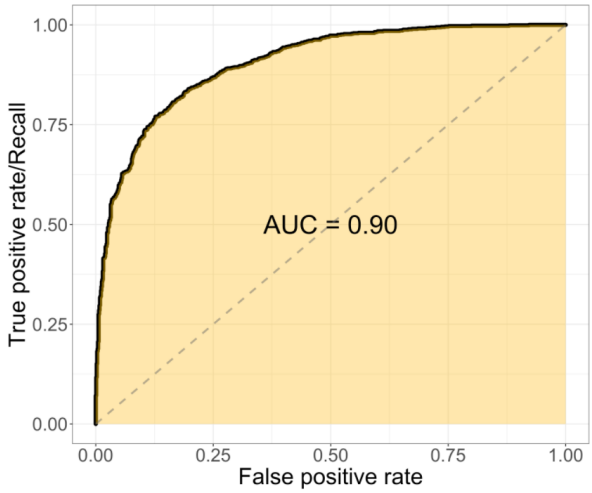


Figure 5.11: ROC curve.

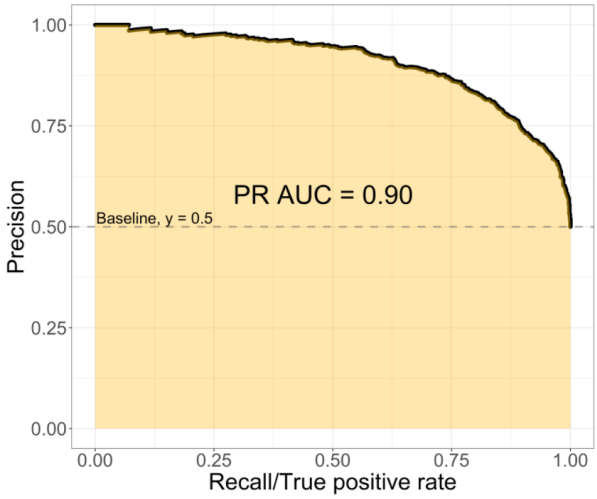


Figure 5.12: PR curve.

classification. The classifier is better as the points are located more to the upper left part of the ROC. If the classifier is random the ROC is a diagonal line [65]. The example of ROC curve is shown in Figure 5.11.

Nevertheless, ROC curves can be misleading when it comes to working with imbalanced classes with a severe skew in the dataset where the

minority class is more important. This is exactly the case in our dataset and in medicine in general. Usually, disease is a rare condition, but it is very important to be detected. In these cases, common alternative is the PR curve, which shows the trade-off between precision and recall, as shown in Figure 5.12. One thing that should be noted is that not every point in the PR space is achievable. That is, for a given dataset it is possible to construct a confusion matrix that corresponds to any sensitivity and FPR pair, but it is not possible to do this for every sensitivity and precision pair. It is shown that the size of the unachievable region is a function of the class skew and it is equal to the fraction of positive samples with respect to all classes [66]. When classes are balanced it is equal to 0.5.

Experimental Evaluation 6

In this Chapter we present the results obtained to validate the BRAVE AI pipeline in this thesis. Section 6.1 provides information of the experimental setup in terms of programming language, frameworks, and devices employed for the given analysis which are both the main resource and the biggest constraint of this research. Section 6.2 presents the obtained results by different Neural Network (NN) architectures. Section 6.3 summarizes the model performances and compare them with the state of the art. Finally, in Section 6.4 we present the visualization of the results on one WSI as it would have been presented in a possible clinical environment.

6.1 Experimental setup and resources

The whole pipeline and the experiments have been developed in Python™, which is a highly employed language for the development of machine learning projects and applications. Characteristics that make Python the best fit for ML-based projects are, in the first place, simplicity and consistency, then access to good libraries and frameworks for ML, flexibility, platform independence, and a wide community.

The biggest part of the work, meaning the neural networks and their training, are developed in PyTorch which is a well-known open-source machine learning framework [67]. Its main advantages are accelerated tensor computations by exploiting graphical processing units (GPU) and simplicity when it comes to building and training NN architectures. Additionally, PyTorch offers dynamic computational graphs, which can

be changed during runtime. For example, this is not enabled with its main competitor TensorFlow which uses static computational graphs [68]. This is highly useful when there is no estimation on how much memory will be required for creating a NN model. It can be compared to static and dynamic memory - dynamic memory is taken and released on the fly and leads to better efficiency.

In particular we used Python version 3.6.8 and the version of PyTorch 1.7.1 with CUDA 11.1. CUDA (Compute Unified Device Architecture) is a parallel computing platform and programming model developed by NVIDIA for CUDA enabled GPUs. In GPU-accelerated applications, the sequential part of the workload runs on the CPU – which is optimized for single-threaded performance – while the compute intensive portion of the application runs on thousands of GPU cores in parallel [69]. The GPU used in this work is a GeForce GTX 1660 Ti.

6.2 Models

In the work of the thesis we explored how different models perform on the dataset, starting from simpler ones and progressively increasing the capacity. We did transfer learning starting from NN trained on ImageNet dataset, which is the standard benchmark in any classification problems. In all of the experiments we performed 10-fold cross-validation. Cross-validation dataset splits are done on patients since having patches from the same patient both in training and validation dataset is giving biased estiomation of the model performance. This happens due to the correlation of the patches from the same patient, and it could mean that the model is seening in validation dataset almost the same data as in training which is resulting in better performance estimation. It should be noted that cross-validation estimation of the performance is usually giving numbers which are more pessimistic than the best model that can be trained on that data.

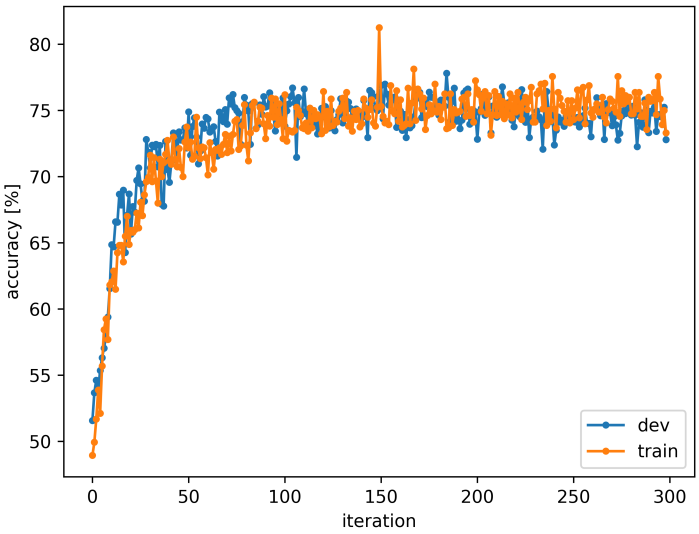


Figure 6.1: Training and validation accuracy history of a ResNet18 model.

All NN models explored in the thesis, utilize the following parameters for training:

- batch size equal to 32;
- stochastic gradient descent optimization;
- cyclic learning rate with minimum value 10^{-6}, maximum value 10^{-3}, and half-cycle length one third of steps in the epoch;
- batch normalization momentum of 0.3.

6.2.1 ResNet18 results

The first model developed in this work relied on ResNet18. In this experiment we used transfer learning with all the layers frozen except the last fully connected layers that we added to the network. The NN was

trained for 2 epochs. The result on 10-fold cross-validation is $(75.06 \pm 2.92)\%$. We have chosen one of the best performing models from 10 folds, and the training history is shown in Figure 6.1.

We calculated all the evaluation metrics for this model and visualized the confusion matrix on the validation data set. They are shown in Tables 6.2 and 6.1, respectively. Analyzing the results obtained in Table 6.2 we can see that area under the Precision-Recall (PR) curve (PR_AUC) is 58.77% which is low, therefore we decided to plot the PR curve and the distribution of the output probabilities that a patch is containing cancerous tissue, and they are shown in Figures 6.3 and 6.2 respectively. We can observe that the PR curve is quickly going down to the lowest value when recall is equal to 1.0 and that is the indication that the model is not performing well on the given dataset. Similarly, we can observe in Figure 6.2 that the output probability distribution is indicating that the model is not sure in its predictions and that many patches are around 0.5 which means that the model is not reliable.

Table 6.1: Confusion matrix of ResNet18 model.

		Predicted value	
		cancer	healthy
Actual Value	cancer	6738	2030
	healthy	5534	16258

Table 6.2: Evaluation performance of ResNet18 model in [%].

Model	Acc	Se	Sp	$F1$	$Prec$	Acc_b	PR_AUC
ResNet18-1	75.25	76.85	74.61	76.04	54.91	75.73	58.77

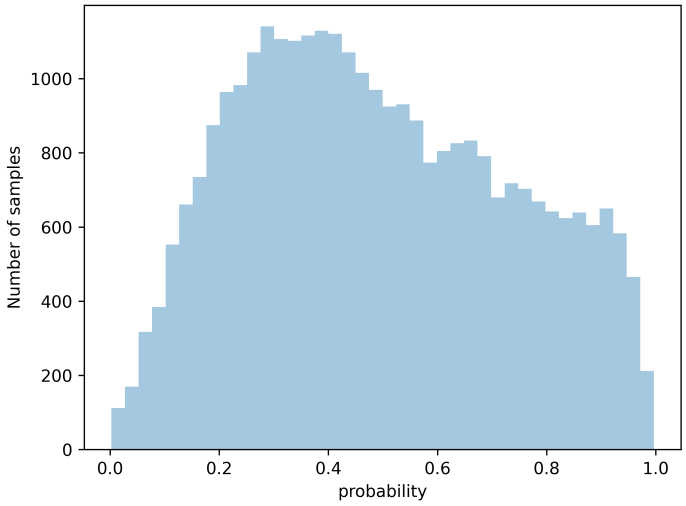


Figure 6.2: Output probabilities distribution for ResNet18 model.

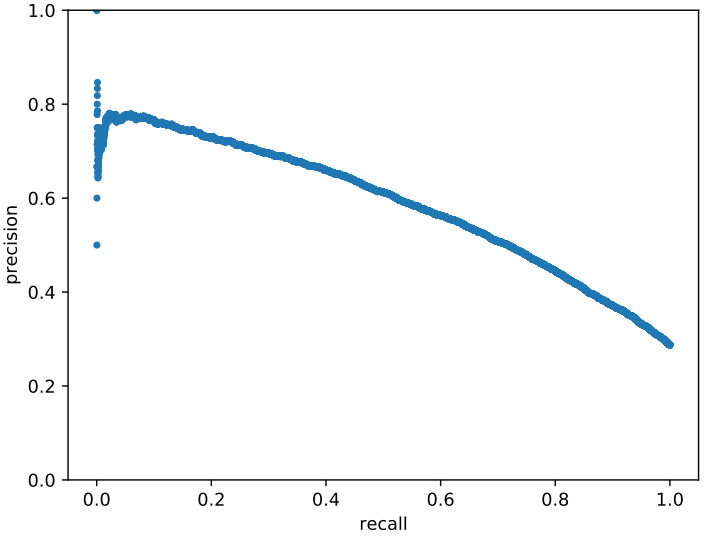


Figure 6.3: Precision-recall curve for ResNet18 model.

6.2.2 ResNet50 results

For developing the ResNet50 model we used again transfer learning, and we froze the first half of the network. The result on 10-fold cross-validation is an accuracy of $(85.62 \pm 1.58)\%$ which means that the model is providing stable solution for the given problem. We have chosen the best model from the cross-validation folds, and the training history of the model is shown in Figure 6.4. We can observe that the model's training and validation accuracy are very close over time, which means that the model is a good fit for the given problem. The model is not overfitting since it is not performing better on the training set than on validation, therefore it is able to generalize well on unseen data. Moreover, we can see a slight upward trend of the accuracy on training dataset which means that the model's accuracy would have been higher if the model has

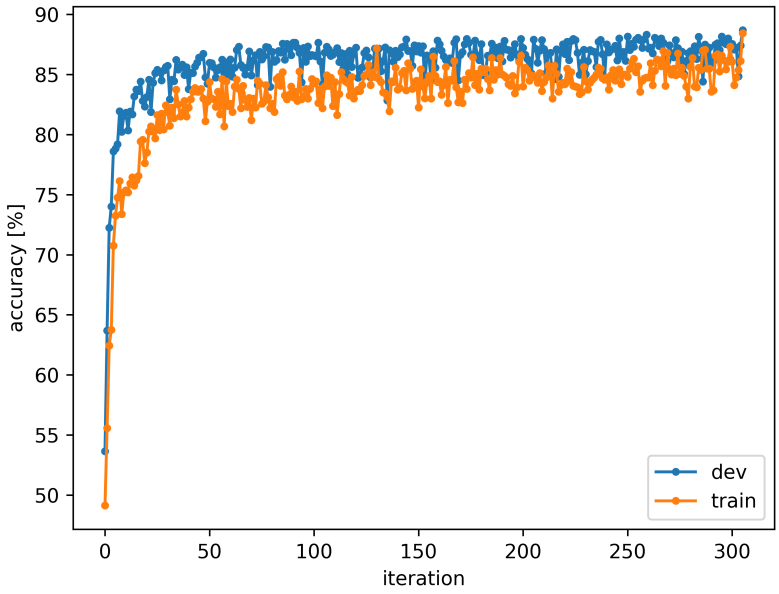


Figure 6.4: Training and validation accuracy history of ResNet50.

been trained for more than two epochs. Moreover, training the model for many epochs require extensive computational resources.

Table 6.3: Confusion matrix of ResNet50 model.

		Predicted value	
		cancer	healthy
Actual Value	cancer	6798	946
	healthy	2207	15105

Table 6.4: Evaluation performance of ResNet50 model in [%].

Model	Acc	Se	Sp	$F1$	$Prec$	Acc_b	PR_AUC
ResNet50	87.41	87.78	87.25	87.59	75.49	87.51	82.88

Confusion matrix with absolute numbers for image patches from validation dataset is shown in Table 6.3. All the metrics related to this model are around the same value which indicates stability of the model performance for different classes as shown in Table 6.4. Distribution of the resulting probabilities for the given classes is shown in Figure 6.5. It can be observed that the distribution still does not show them as completely separable classes, but we can observe that two peaks have formed close to the borders of the range. PR curve is shown in Figure 6.6, and it reports the highest values for small recall, and then it falls down till the value defined by the class balance in the validation dataset.

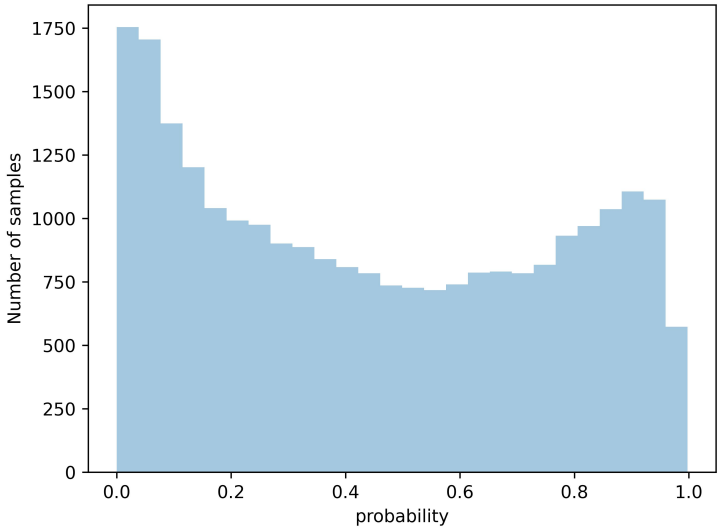


Figure 6.5: Output probabilities distribution for ResNet50 model.

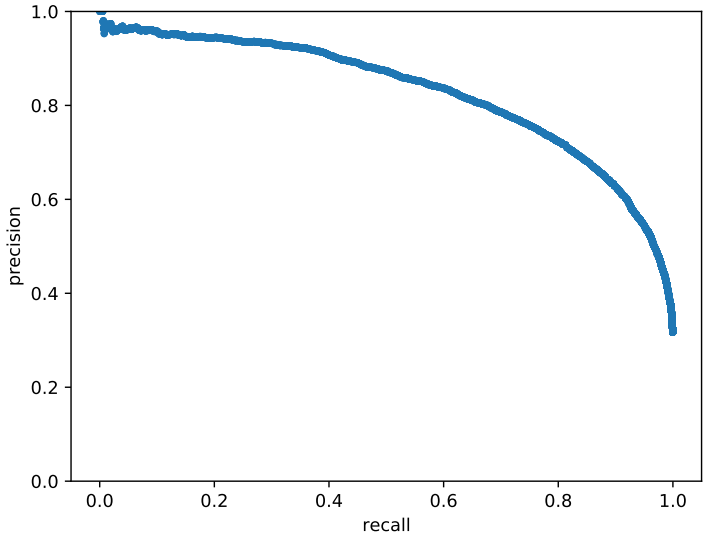


Figure 6.6: Precision-recall curve for ResNet50 model.

6.2.3 DenseNet121 results

We used, like in the previous experiments, transfer learning with pre-trained network on ImageNet dataset and we froze the first 2/3 of the network since the network is very deep and we wanted to avoid overfitting, but at the same time have enough capacity to capture the variation of the data. The result on 10-fold cross-validation has an accuracy of $(87.44 \pm 1.45)\%$ which means that the model is providing stable solution for the given problem. The training history of the best performing model from cross-validation is shown in Figure 6.7. Again, the model's training and validation accuracy are very close over time, which means that the model is a good fit. Furthermore, we can see a slight upward trend of the accuracy on training dataset also in this model, which means

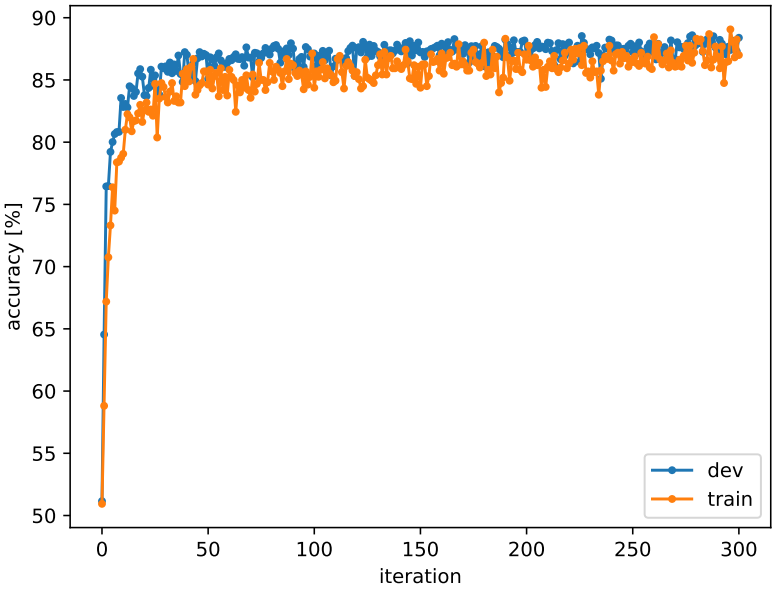


Figure 6.7: Training and validation accuracy history of the chosen model of architecture DenseNet121.

that the model can be trained for longer time and produce even better performance.

Confusion matrix with absolute numbers of the image patches from validation dataset is shown in Table 6.5. All evaluation metrics are shown in Table 6.6. We can observe that all of them have similar values which indicates that the model is balanced, and not performing better or worse for different classes. Distribution of resulting probabilities for given classes is shown in Figure 6.8, where we can see that probabilities now have a distribution that resembles bi-modal distribution with peaks further apart. PR curve is shown in Figure 6.9, and reports the highest values for small recall as expected, and then it slowly falls down till the value defined by the class balance in the validation dataset.

Table 6.5: Confusion Matrix of DenseNet121 model.

		Predicted value	
		cancer	healthy
Actual Value	cancer	9106	795
	healthy	2941	16470

Table 6.6: Evaluation performance of Densenet121 model in [%].

Model	Acc	Se	Sp	$F1$	$Prec$	Acc_b	PR_AUC
DenseNet121	87.25	91.97	84.84	89.55	75.59	88.41	90.12

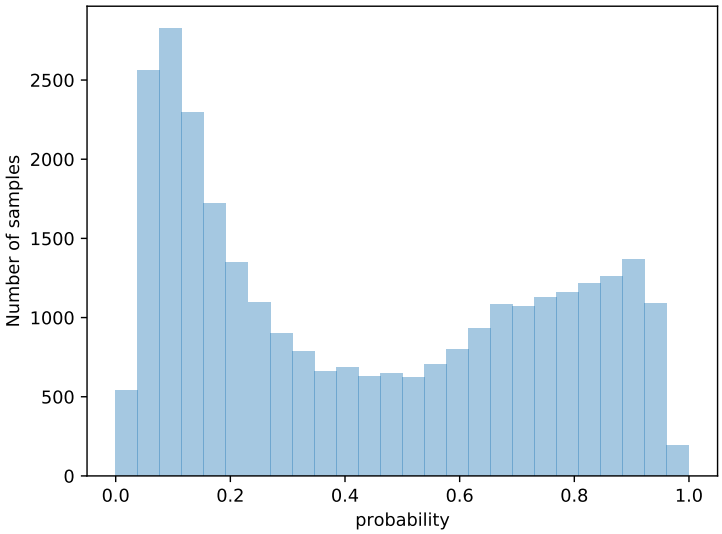


Figure 6.8: Output probabilities distribution.

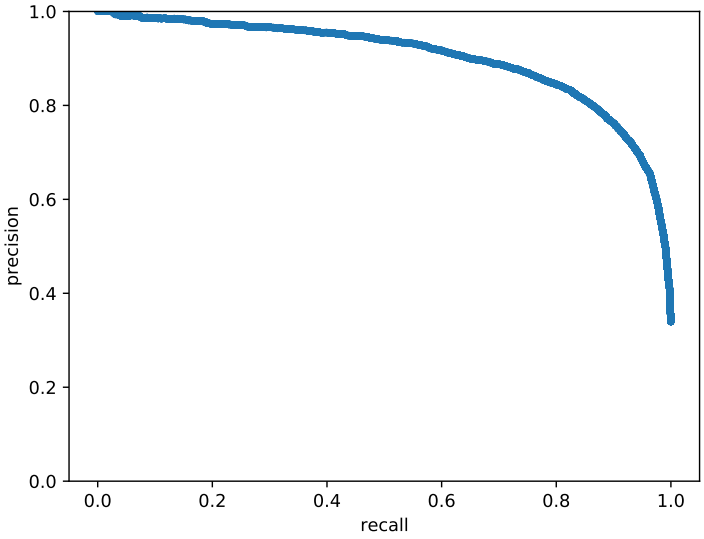


Figure 6.9: Precision-recall curve for DenseNet121 model.

6.3 Model comparison and comparison with the State of the Art

Firstly, we are going to compare the models developed in the thesis. Evaluation metrics for all of the models are shown in Table 6.7 together with the metrics for the models proposed in the state of the art. Analysing the results shown in details in the previous sections, we can conclude that the model using the ResNet18 architecture is not able to capture the information contained in the data, and perform reliable classification. We can observe that all of its evaluation metrics are lower than the other models, and the state of the art, and it is not a good model for solving this problem. The model obtained using the ResNet50 architecture has a better performance, but the best performing model of those proposed in the thesis is the one relying on the architecture of DenseNet121 with a balanced accuracy of 88.41%, an F1 score of 89.55%, and a sensitivity of 91.97%.

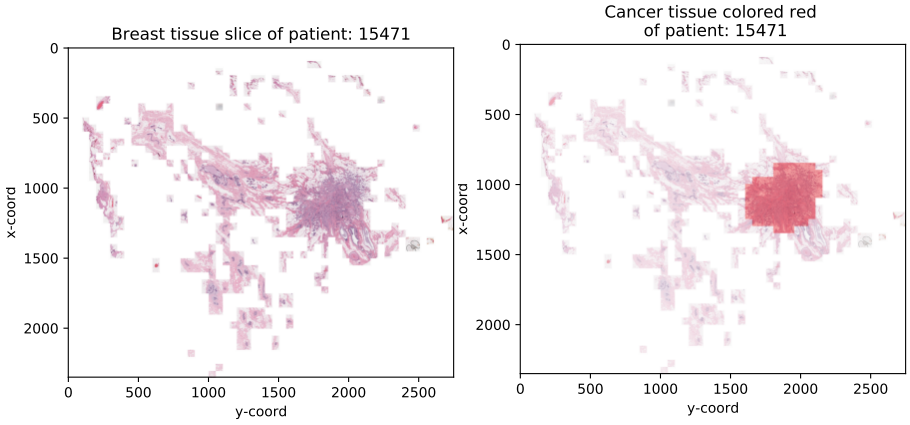
Comparing our best performing model with the other works we can conclude that our model achieves comparable performance with the ones proposed in the state of the art which used the same dataset. In fact, we can see that the only work obtaining metrics that are above 90% is the one done by Celik et al. [54]. It should be noted that all of

Table 6.7: Comparison of results with state of the art. The unit of numbers in the table is in [%].

Source	Model	Epoch	*Acc*	*Se*	*Sp*	*F*1	*Prec*	Acc_b	*PR_AUC*
BRAVE AI	ResNet18	2	75.25	76.85	74.61	76.04	54.91	75.73	58.77
BRAVE AI	ResNet50	2	87.41	87.78	87.25	87.59	75.49	87.51	82.88
BRAVE AI	**DenseNet121**	**2**	**87.25**	**91.97**	**84.84**	**89.55**	**75.59**	**88.41**	**90.12**
Reza et al.[49]	3 layer CNN	20	85.48	80.85	90.12	84.78	-	85.48	-
Cruz-Roa et al.[8]	3 layer CNN	25	-	79.60	88.86	71.80	65.40	84.23	-
Celik et al.[54]	**ResNet50**	**30**	**91.96**	**93.64**	**88.28**	**94.11**	**94.58**	**90.96**	**-**
Celik et al.[54]	**DenseNet161**	**30**	**91.20**	**89.59**	**93.56**	**92.38**	**95.34**	**91.57**	**-**
Romero et al.[51]	Inception	55	-	-	-	89.70	-	89.0	-

these works trained their models for at least twenty or more epochs, while we showed that comparable results can be obtained with only two epochs. Moreover, none of the works in the state of the art used cross-validation to prove the stability of their proposed architectures and training, but used only holdout validation. Furthermore, as we stated before, the training history of our ResNet50 and DenseNet121 shows a slight upward trend with respect to the accuracy, which indicates that they can reach even better performances if trained for longer periods of time, which was not feasible with our computational resources.

6.4 Output image visualization



(a) (b)

Figure 6.10: Figure (a) shows reconstructed slide from patches. On Figure (b) is shown the ground truth - patches labeled as cancer.

In Figure 6.10a we present how the extracted Whole Slide Image (WSI) patches in the image look like, and where exactly they are positioned in the image. As explained in Section 5.2 the dataset contains only the patches that have the tissue on them, and the patches containing background are removed from the dataset. The Figure 6.10b shows

the ground truth of the image, where the patches containing Invasive Ductal Carcinoma (IDC) are labeled with red colour masks. The Figure 6.11 presents the output of the BRAVE AI framework, which is a probability map of the presence of IDC on the tissue of the slice. The red parts show that the classifier gives high probability that the patch is cancerous, whereas blue regions represent parts without cancer.

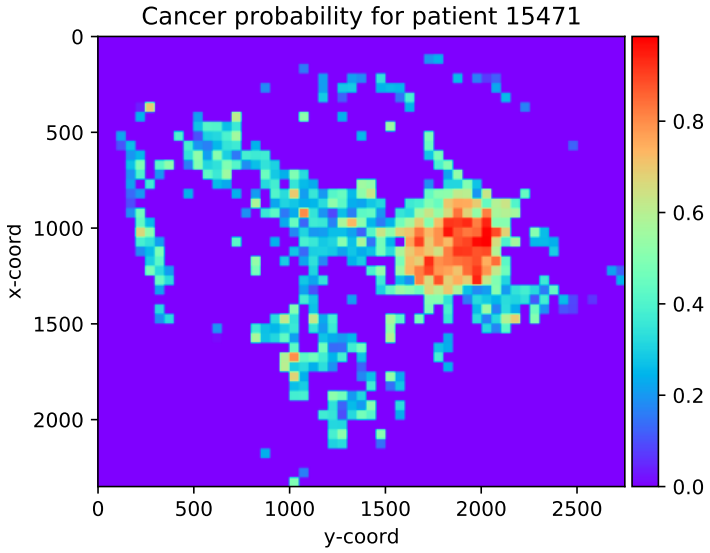


Figure 6.11: Visualization of results of automatic classification of patches from the slice. Red color of the image represents high probability of cancer, while blue part of the spectrum represents patches with very low probability of cancer.

Conclusion and Future Works 7

This Chapter summarizes all the work done in the thesis, providing a general overview of what has been done, as well as conclusions and plans for the future. In Section 7.1 we provide the main contributions of the work done in the thesis. Section 7.2 reflects on limitations of the work, while Section 7.3 provides ideas for the future and possible ways of overcoming the limitations mentioned before and improving the results.

7.1 Contribution

In this Section, we summarize the contribution of the thesis and the proposed solution.

Firstly, we provide the design and the implementation of an automated, end-to-end pipeline for breast cancer assessment. This means that we can produce the cancer probability maps from tissue slides without the need for additional tunning, thresholding, or counting of cells for each patient, which enables standardization of breast cancer assessment. Moreover, we propose the BRAVE AI framework that can reduce the workload for the pathologist and enable higher throughput from the pathology departments which pace of work is the bottleneck of the modern hospitals' dynamic today.

Secondly, we provide the results of the exploration of Deep Learning (DL) space of solutions for histopathological image classification. We start with smaller Convolutional Neural Networks (CNNs) showing that they are not able to capture all the variability of the data, thus going deeper is required for finding an optimal solution. We move forward with

deeper architectures, and we demonstrate that the models' performance is increasing. Therefore, we provide initial solutions, carefully narrowing down the exploration space, which can be a good starting point for future research.

Finally, we provide a detailed explanation of the methodology and implementation of the models with all the hyperparameters used, allowing the reproducibility of proposed solutions. Moreover, we provide exhaustive results with different metrics for the evaluation of the obtained models, also on cross-validation, which are the proofs of the stability of the proposed solution on the unseen data. We show that our best performing model, using DenseNet121 architecture, reach a balanced accuracy of 88.41%, an F1 score of 89.55%, and a sensitivity of 91.97%, achieving performance comparable to the state of the art.

7.2 Limitations

Despite all the presented work, there are some limitations that are affecting the result of the work and limiting its impact, which could not be solved during the work on the thesis in order to improve the results.

Firstly, the dataset that we are using contains only 279 WSI and the same number of patients. Therefore, even though the cross-validation is done subject-wise in order to avoid overfitting on the subjects, the dataset is small for the field of DL. On the other hand, some of the available datasets that are larger, like CAMELYION, contain a couple of hundreds of unprocessed WSI which require terabytes of memory to be stored, and high computational power to be processed, and used for further analysis. This is the problem that many research groups are encountering in biomedical engineering, due to the lack of computing infrastructure.

On the other hand, the ground truth provided for the used dataset is not perfect, meaning that the pathologist did the annotations on smaller

images, which affects the border patches of the selected regions. This is decreasing the performance estimation for two reasons: firstly, if the patch that is not containing cancer is labeled erroneously as cancerous in the ground truth, and correctly classified by the model as negative, it is lowering the accuracy estimate in the evaluation since the ground truth and the prediction are discordant; secondly, since the model is learning from the presented data, incorrect ground truth can alter the model performance and bring confusion to the model while training. This limitation can be improved with a quality check done by another pathologist on the given annotations, carefully classifying the tissue patches into the two classes.

Furthermore, the whole pipeline is tested on only one dataset, which is not enough to claim that the model is able to generalize well on the unseen data from the real world and that it is good enough to be implemented in a clinical workflow. In order to provide a benchmark for the detection of breast cancer, the whole workflow should be validated on different datasets from several points of acquisition in various hospitals. The main reason is the fact that acquisition machinery can be different and produce different artifacts that could affect the result, even though we are addressing this limitation through the introduction of augmentation of the dataset during the training of the Neural Network (NN).

7.3 Future Work

Taking into account all said, there are several ways in which the system could be improved. First of all, we should process a larger number of patients with the supervision done by an expert in pathology, for the segmentation of the image. In this way, we can have better annotations, and better segmentation provided for the Whole Slide Image (WSI), which would have increased the performance of the model. Secondly, with the baseline results provided for different models, we can choose

only one of them and increase the strength of augmentations while letting the model train for a larger number of epochs. This would have lowered the accuracy at the beginning but could result in a more robust model which can generalize better on different datasets. Furthermore, we can explore the idea of an ensemble of classifiers, which are known to increase performance, if the errors, that the models are making individually, are not correlated.

Lastly, as a far future, we could start thinking about how to include a system like this in the clinical workflow as an assisting device to the pathologist which would require automatic loading of the WSI from the scanner and automatic patching and storing of the image before the framework we proposed in the thesis.

Concluding, we have proposed the prototype of an automated pipeline that aims to be credible support to the pathologist in the identification of breast cancer in histopathological images. It provides cancer maps of histology slide on the output which suggests to the expert the regions with a high probability of cancer to which the attention should be focused. In this way, it can reduce the workload, and the time needed for decision making which is the bottleneck of today's diagnosis process in the laboratories.

List of abbreviations

AI Artificial Intelligence. 20, 23, 24, 51
ANN Artificial Neural Network. 26

CAD Computer Aided Diagnosis. 32
CNN Convolutional Neural Network. 3, 23, 26, 28–30, 32, 34–36, 54, 56, 81

DenseNet Densely connected Neural Network. 3, 35, 36, 42, 54–56
DL Deep Learning. 2–4, 21, 23–26, 32, 34, 36–39, 42, 45, 81, 82
DNA Deoxyribonucleic Acid. 9, 18
DP Digital Pathology. 19, 20

FDA Federal Drug Association. 20

H&E Hematoxylin and Eosin. ix, 18

IDC Invasive Ductal Carcinoma. i, 2, 4, 34, 42–47, 57, 61, 80

ML Machine Learning. 2, 7, 20, 21, 23–25, 31, 33, 41, 43, 48, 58, 63
MRI Magnetic Resonance Imaging. 16, 32

NN Neural Network. i, 2, 25–30, 32, 35, 38, 41, 42, 45, 48, 50–58, 67–69, 83

PR Precision-Recall. 70, 73, 76

ReLu Rectified Linear Unit. 26, 53

ResNet Residual Neural Network. 3, 35, 36, 42, 51–54

RF Random Forest. 3, 33, 35

RNA Ribonucleic Acid. 18

SMOTE Synthetic Minority Over-sampling Technique. 34

SVM Support Vector Machine. 33

WHO World Health Organization. 1, 10

WSI Whole Slide Image. i, ix, 2, 4, 20, 25, 33–38, 41–44, 46–48, 79, 83, 84

Bibliography

[1] *WHO report on cancer: setting priorities, investing wisely and providing care for all*. World Health Organization, 2020.

[2] David J. Dabbs. *Breast pathology*. Elsevier Saunders, 2012.

[3] Arnau Oliver et al. «A review of automatic mass detection and segmentation in mammographic images». In: *Medical Image Analysis* 14.2 (2010), pp. 87–110. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2009.12.005.

[4] Azam Hamidinekoo et al. «Investigating the Effect of Various Augmentations on the Input Data Fed to a Convolutional Neural Network for the Task of Mammographic Mass Classification». In: June 2017, pp. 398–409. ISBN: 978-3-319-60963-8. DOI: 10.1007/978-3-319-60964-5_35.

[5] J.M. Bueno de Mesquita et al. «The impact of inter-observer variation in pathological assessment of node-negative breast cancer on clinical risk assessment and patient selection for adjuvant systemic treatment». In: *Annals of Oncology* 21.1 (2010), pp. 40–47. ISSN: 0923-7534. DOI: https://doi.org/10.1093/annonc/mdp273.

[6] Marugame A. et al. «Categorization of HE Stained Breast Tissue Samples at Low Magnification by Nuclear Aggregations». In: *World Congress on Medical Physics and Biomedical Engineering, September 7 - 12, 2009* 25.4 (2009). Ed. by Dössel O. and Schlegel W.C. DOI: 10.1007/978-3-642-03882-2_45.

[7] Mira Valkonen et al. «Metastasis detection from whole slide images using local features and random forests». In: *Cytometry. Part A : the journal of the International Society for Analytical Cytology* 91 (Apr. 2017). DOI: 10.1002/cyto.a.23089.

[8] Angel Cruz-Roa et al. «Automatic detection of invasive ductal carcinoma in whole slide images with Convolutional Neural Networks». In: *Progress in Biomedical Optics and Imaging - Proceedings of SPIE* 9041 (Feb. 2014). DOI: 10.1117/12.2043872.

[9] Guilherme Aresta et al. «Bach: Grand challenge on breast cancer histology images». In: *Medical image analysis* 56 (2019), pp. 122–139.

[10] Babak Ehteshami Bejnordi et al. «Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer». In: *JAMA* 318.22 (Dec. 2017), pp. 2199–2210. ISSN: 0098-7484. DOI: 10.1001/jama.2017.14585. eprint: https://jamanetwork.com/journals/jama/articlepdf/2665774/jama_ehteshami_bejnordi_2017_oi_170113.pdf. URL: https://doi.org/10.1001/jama.2017.14585.

[11] Stanley Leonard Robbins et al. «Cell Injury, Cell Death, and Adaptations». In: *Robbins basic pathology*. Elsevier, 2018.

[12] Luiz Carlos. Junqueira and Anthony L. Mescher. *Junqueira's basic histology: text and atlas*. McGraw-Hill Medical Publishing, 2013.

[13] Brian Nation and Guy Orchard. «What is histopathology?» In: *Histopathology*. Oxford University Press., 2018.

[14] Gerald Karp and James G. Patton. *Cell and molecular biology: concepts and experiments*. John Wiley, 2013.

[15] Harvey F. Lodish. «Life Begins with Cells». In: *Molecular cell biology*. W. H. Freeman, 2008.

[16] Barbara Young, Phillip Woodford, and Geraldine O'Dowd. *Wheater's functional histology: a text and colour atlas*. Churchill Livingston/Elsevier, 2014.

[17] *Cancer*. 2018. URL: https://www.who.int/en/news-room/fact-sheets/detail/cancer.

[18] *What Is Cancer?* URL: https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[19] Stanley Leonard Robbins et al. «Female Genital System and Breast». In: *Robbins basic pathology*. Elsevier, 2018.

[20] *Breast cancer: prevention and control*. 2016. URL: https://www.who.int/cancer/detection/breastcancer/en/.

[21] *Breast Tumors*. 2020. URL: https://www.nationalbreastcancer.org/breast-tumors/.

[22] *What Is Breast Cancer?: Breast Cancer Definition*. URL: https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html.

[23] Bcnapinklady. *Types of breast cancer*. URL: https://www.bcna.org.au/understanding-breast-cancer/what-is-breast-cancer/types-of-breast-cancer/.

[24] *Breast Cancer - Diagnosis*. 2020. URL: https://www.cancer.net/cancer-types/breast-cancer/diagnosis.

[25] *Diagnostic Mammogram*. 2020. URL: https://www.nationalbreastcancer.org/diagnostic-mammogram.

[26] *Ultrasound*. 2019. URL: https://www.nationalbreastcancer.org/breast-ultrasound.

[27] *Biopsy*. 2020. URL: https://www.nationalbreastcancer.org/breast-cancer-biopsy.

[28] BAppSc Geoffrey Rolls and Global Marketing Manager James Anderson. *An Introduction to Routine and Special Staining*. 2012. URL: https://www.leicabiosystems.com/knowledge-pathway/an-introduction-to-routine-and-special-staining/.

[29] Pranab Dey. *Basic and Advanced Laboratory Techniques in Histopathology and Cytology*. Springer Singapore, 2018.

[30] Anthony L. Mescher. *Junqueira's basic histology: text and atlas*. McGraw-Hill Education, 2018.

[31] Hani A Alturkistani, Faris M Tashkandi, and Zuhair M Mohammedsaleh. «Histological Stains: A Literature Review and Case Study». In: *Global Journal of Health Science* 8.3 (2015), p. 72. DOI: 10.5539/gjhs.v8n3p72.

[32] S. Kim Suvarna, Christopher Layton, and John D. Bancroft. *Bancroft's theory and practice of histological techniques*. Elsevier, 2019.

[33] Laura Barisoni et al. «Digital pathology and computational image analysis in nephropathology». In: *Nature Reviews Nephrology* 16.11 (2020), pp. 669–685.

[34] J.D. Pallua et al. «The future of pathology is digital». In: *Pathology - Research and Practice* 216.9 (2020), p. 153040. ISSN: 0344-0338. DOI: https://doi.org/10.1016/j.prp.2020.153040. URL: https://www.sciencedirect.com/science/article/pii/S0344033819330596.

[35] *Class 3 Device Recall Leica Microsystems Inc.* URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfres/res.cfm?id=139321.

[36] IBM Cloud Education. *What is Artificial Intelligence (AI)?* 2020. URL: https://www.ibm.com/cloud/learn/what-is-artificial-intelligence.

[37] Kevin P. Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 2012.

[38] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. «Deep learning». In: *Nature* 521.7553 (2015), 436–444. DOI: 10.1038/nature14539.

[39] Subana Shanmuganathan and Sandhya Samarasinghe. *Artificial neural network modelling*. Springer, 2016.

[40] F. Rosenblatt. «The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain». In: *Psychological Review* (1958), pp. 65–386.

[41] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. «Learning representations by back-propagating errors». In: *nature* 323.6088 (1986), pp. 533–536.

[42] Jonghong Kim et al. «Convolutional Neural Network with Biologically Inspired Retinal Structure». In: *Procedia Computer Science* 88 (2016). 7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016, held July 16 to July 19, 2016 in New York City, NY, USA, pp. 145–154. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2016.07.418. URL: https://www.sciencedirect.com/science/article/pii/S187705091631674X.

[43] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[44] Wenda He et al. «A review on automatic mammographic density and parenchymal segmentation». In: *International journal of breast cancer* 2015 (2015).

[45] A. Šerifović-Trbalić et al. «Classification of benign and malignant masses in breast mammograms». In: *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2014, pp. 228–233. DOI: 10.1109/MIPRO.2014.6859566.

[46] Dinesh Pandey et al. «Automatic and fast segmentation of breast region-of-interest (ROI) and density in MRIs». In: *Heliyon* 4.12 (2018), e01042. ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2018.e01042.

[47] J.M. Chen et al. «New breast cancer prognostic factors identified by computer-aided image analysis of HE stained histopathology images». en. In: *Sci Rep* (May 29, 2015). PMID: 26022540; PMCID: PMC4448264. DOI: 10.1038/srep10690..

[48] A. Basavanhally et al. «Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides». en. In: *Epub 2013 Feb 5. PMID: 23392336; PMCID: PMC5778451* (Aug. 8, 2013). DOI: 10.1109/TBME.2013.2245129..

[49] Md Shamim Reza and Jinwen Ma. «Imbalanced Histopathological Breast Cancer Image Classification with Convolutional Neural Network». In: *2018 14th IEEE International Conference on Signal Processing (ICSP)* (2018), pp. 619–624.

[50] C. Szegedy et al. «Going deeper with convolutions». In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.

[51] F. P. Romero, A. Tang, and S. Kadoury. «Multi-Level Batch Normalization in Deep Networks for Invasive Ductal Carcinoma Cell Discrimination in Histopathology Images». In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019, pp. 1092–1095. DOI: 10.1109/ISBI.2019.8759410.

[52] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[53] Gao Huang et al. *Densely Connected Convolutional Networks*. 2018. arXiv: 1608.06993 [cs.CV].

[54] Yusuf Celik et al. «Automated Invasive Ductal Carcinoma Detection Based Using Deep Transfer Learning with Whole-Slide Images». In: *Pattern Recognition Letters* 133 (Mar. 2020). DOI: 10.1016/j.patrec.2020.03.011.

[55] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[56] Babak Ehteshami Bejnordi et al. «Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images». In: vol. 2017. Apr. 2017, pp. 929–932. DOI: 10.1109/ISBI.2017.7950668.

[57] Dayong Wang et al. «Deep Learning for Identifying Metastatic Breast Cancer». In: *ArXiv* abs/1606.05718 (2016).

[58] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. «ImageNet Classification with Deep Convolutional Neural Networks». In: *Commun. ACM* 60.6 (May 2017), 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386.

[59] C. Szegedy et al. «Going deeper with convolutions». In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.

[60] D. Wang, C. Otto, and A. K. Jain. «Face Search at Scale». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pp. 1122–1136. DOI: 10.1109/TPAMI.2016.2582166.

[61] Fabio Spanhol et al. «A Dataset for Breast Cancer Histopathological Image Classification». In: *IEEE transactions on bio-medical engineering* 63 (Nov. 2015). DOI: 10.1109/TBME.2015.2496264.

[62] J. Deng et al. «ImageNet: A Large-Scale Hierarchical Image Database». In: *CVPR09*. 2009.

[63] Xavier Glorot and Y. Bengio. «Understanding the difficulty of training deep feedforward neural networks». In: *Journal of Machine Learning Research - Proceedings Track* 9 (Jan. 2010), pp. 249–256.

[64] Liwei Wang et al. «Training Deeper Convolutional Networks with Deep Supervision». In: *CoRR* abs/1505.02496 (2015). arXiv: 1505.02496. URL: http://arxiv.org/abs/1505.02496.

[65] Tom Fawcett. «An introduction to ROC analysis». In: *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010. URL: https://www.sciencedirect.com/science/article/pii/S016786550500303X.

[66] Kendrick Boyd et al. «Unachievable Region in Precision-Recall Space and Its Effect on Empirical Evaluation». In: *CoRR* abs/1206.4667 (2012). arXiv: 1206.4667. URL: http://arxiv.org/abs/1206.4667.

[67] Adam Paszke et al. «PyTorch: An Imperative Style, High-Performance Deep Learning Library». In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[68] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[69] NVIDIA, Péter Vingelmann, and Frank H.P. Fitzek. *CUDA*. 2020. URL: https://developer.nvidia.com/cuda-toolkit.