République Algérienne Démocratique et Populaire

Ministre de l'Enseignement Supérieur et la Recherche Scientifique

Université Mohamed Khider de Biskra

Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie

Département de l'informatique.

THESE

Pour l'obtention de diplôme de :

Doctorat en Informatique de l'Université de Biskra

Spécialité: Intelligence Artificielle et Génie Logiciel

Méthode Automatique pour l'Interprétation Efficace des Images Histologiques du Cancer du Sein

---

Présentée par: Adel ABDELLI

Soutenue le ..... devant le jury composé de :

Pr. Okba KAZAR	Université de Biskra	Président
Pr. Rachida SAOULI	Université de Biskra	Directrice de thèse
Pr. Laid KAHLOUL	Université de Biskra	Examinateur
Pr. Mohamed BEN MOAHMED	Université de Constantine 2	Examinateur
Dr. Khalifa DJEMAL	Université d'Évry Val d'Essonne, France	Examinateur

Année universitaire: 2021/2022

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

MOHAMED KHIDER UNIVERSITY OF BISKRA

Faculty of Exact Sciences, Natural and Life Sciences

Computer Science Department

THESIS

To obtain the degree of:

PhD in Computer Science from the University of Biskra

Specialty: Artificial Intelligence and Software Engineering

Automatic Method for the Efficient Interpretation of Histological Breast Cancer Images

---

Presented by: Adel ABDELLI

Defended on ..... in front of the jury composed of:

Pr. Okba KAZAR	University of Biskra	President
Pr. Rachida SAOULI	University of Biskra	Supervisor
Pr. Laid KAHLOUL	University of Biskra	Examiner
Pr. Mohamed BEN MOAHMED	University of Constantine 2	Examiner
Dr. Khalifa DJEMAL	Université d'Évry Val d'Essonne, France	Examiner

Academic year: 2021/2022

Acknowledgement

This dissertation would not have been possible without the encouragement, assistance, and friendship of many people. It is my pleasure to express my gratitude to those who have encouraged and mentored me during the period of my thesis.

Foremost, I would like to present my gratitude to the members of the jury: Pr. Okba KAZAR, Pr. Laid KAHLOUL, Pr. Mohamed BENMOHAMED, and Dr. DJEMAL Khalifa for doing me the honor of being reporters and evaluating the work presented in this manuscript.

I'd like to convey my heartfelt thanks to my supervisor SAOULI Rachida, for her unwavering support of my Ph.D. studies and research, as well as her comments, conversations, and feedback. Her advice has been invaluable to me during my Ph.D. studies.

Also, thanks to my friends and lab colleagues at the University of Biskra and at the intelligent computing laboratory (LINFI) for the stimulating discussions we had during our days together and for their support, help and friendship during my thesis period.

Finally, I'd want to thank my family, specifically my parents, brothers, and sisters, for their unwavering support throughout my life. I would not have accomplished my thesis if it hadn't been for their continuous love and encouragement.

Thank you so much for everything.

ADEL

Abstract

Breast cancer is the most common type of cancer in women. Pathologists must often analyze histological imaging slides across the whole slide tissues at various magnifications to determine tumor malignancy and grade. The interpretation of these images is one of the time-consuming and labor-intensive tasks required to make an appropriate diagnosis. Also, the textures and diversity of digital histopathology images are complicated. As a result, accurately interpreting histopathological images necessitates correct classification and identification of the tissue components in these images. Consequently, Computer-Aided Diagnosis (CAD) systems are in great demand. The thesis's goal is to offer autonomous deep learning algorithms for classifying histopathological images of breast tumors.

First, we are mainly interested in classifying the histopathological images of breast tumors into benign or malignant, where we suggest a new classification layer based on Multiple Instance Learning (MIL). A flatten, or global pooling layer is used before the fully connected layers in a conventional Convolutional Neural Network (CNN). However, in our suggested layer, we consider each last feature map in the network to be an instance that the output layer will classify. The image's (bag) class will then be determined using an aggregation method. This mapping enables the model to categorize each feature separately, allowing it to detect micro-objects in complicated tissue images. Furthermore, our method was successful in attaining high accuracy without the need for image pretreatment techniques such as color normalization, stain normalization, or any other techniques.

Second, to classify the breast cancer grade based on the histopathological images, we propose a strategy for detecting breast cancer grades by merging two separate datasets. Our suggested strategy involves adding a new class (grade 0) to the three recognized classes of breast cancer grades, allowing our model to detect both malignancy and grade of breast cancers. For comparing lightweight and heavyweight architectures, our models are trained using two distinct convolutional neural network architectures, the ResNet50 and the MobileNet.

Finally, for both works we did on breast cancer classifications, our models outperform several previous works on breast cancer malignancy classification and grades classification.

Keywords. Breast cancer, histopathological images, multiple instance learning, convolutional neural network, breast cancer grade.

Résumé

Le cancer du sein est le type de cancer le plus répandu chez les femmes. Les pathologistes doivent souvent analyser des lames d'imagerie histologique sur l'ensemble des tissus de la lame à différents grossissements pour déterminer la malignité et le grade de la tumeur. L'interprétation de ces images est l'une des tâches chronophages et laborieuses nécessaires pour établir un diagnostic approprié. De plus, les textures et la diversité des images d'histopathologie numérique sont compliquées. En conséquence, l'interprétation précise des images histopathologiques nécessite une classification et une identification correctes des composants tissulaires dans ces images. Par conséquent, les systèmes de diagnostic assisté par ordinateur sont très demandés. L'objectif de la thèse est de proposer des algorithmes autonomes d'apprentissage profond pour la classification d'images histopathologiques de tumeurs mammaires.

Premièrement, nous nous intéressons principalement à la classification des images histopathologiques des tumeurs mammaires en bénignes ou malignes où nous proposons une nouvelle couche de classification basée sur l'apprentissage à instances multiples (AIM). Une couche Flatten ou de Global Pooling est utilisée avant les couches entièrement connectées dans un réseau de neurones à convolution conventionnel (CNN). Cependant, dans notre couche suggérée, nous considérons chaque dernière feature map du réseau comme une instance que la couche de sortie classera. La classe (sac; Bag en Anglais) de l'image sera alors déterminée à l'aide d'une méthode d'agrégation. Cette méthode permet au modèle de catégoriser chaque Feature Map séparément, ce qui lui permet de détecter des micro-objets dans des images de tissus complexes. De plus, notre méthode a réussi à atteindre une grande précision sans avoir besoin de techniques de prétraitement d'image telles que la normalisation des couleurs, la normalisation des taches (Stain Normalization) ou toute autre technique.

Deuxièmement, pour classer le grade du cancer du sein sur la base des images histopathologiques, nous proposons une stratégie de détection du grade du cancer du sein en fusionnant deux bases de données distincts. Notre stratégie suggérée consiste à ajouter une nouvelle classe (grade 0) aux trois classes reconnues de grades de cancer du sein, permettant à notre modèle de détecter à la fois la malignité et le grade des cancers du sein. Pour comparer les architectures légères et lourdes, nos modèles sont entraînés à l'aide de deux architectures de réseau neuronal convolutionnel distinctes, le ResNet50 et le MobileNet.

Enfin, pour les deux travaux que nous avons effectués sur les classifications du cancer du sein, nos modèles surpassent plusieurs travaux antérieurs sur la classification des malignités du cancer du sein et la classification des grades.

Mots clés. Cancer du sein, images histopathologiques, apprentissage à instances multiples, réseau de neurones convolutifs, grade du cancer du sein.

ملخص

سرطان الثدي هو أكثر أنواع السرطانات شيوعًا بين النساء. يجب على علماء الأمراض في كثير من الأحيان تحليل شرائح التصوير النسيجي عبر أنسجة الشريحة بأكملها بتكبيرات مختلفة لتحديد الورم الخبيث ودرجته. يعد تفسير هذه الصور أحد المهام التي تستغرق وقتًا طويلاً وتتطلب جهدًا كثيفًا لإجراء التشخيص المناسب. كما أن التركيبات وتنوع صور التشريح المرضي الرقمي معقدة. نتيجة لذلك، يتطلب التفسير الدقيق للصور النسيجية المرضية التصنيف الصحيح وتحديد مكونات الأنسجة في هذه الصور. نتيجة لذلك، هناك طلب كبير على أنظمة التشخيص بمساعدة الكمبيوتر. الهدف من هاته الأطروحة هو تطوير خوارزميات التعلم العميق لتصنيف بدقة الصور النسيجية المرضية لأورام الثدي بشكل آلي.

أولاً، نحن مهتمون بشكل أساسي بتصنيف الصور النسيجية المرضية لأورام الثدي إلى حميدة أو خبيثة حيث نقترح طبقة تصنيف جديدة تعتمد على التعلم متعدد الحالات. يتم استخدام طبقة تجميع مسطحة أو شاملة قبل الطبقات المتصلة بالكامل في شبكة عصبية تلافيفية تقليدية. أما، في الطبقة التي اقترحناها، نعتبر كل خريطة الخصائص الأخيرة في الشبكة هيئة او عنصرا ستصنفه طبقة المخرجات على حدة. سيتم بعد ذلك تحديد فئة الصورة (الحقيبة) باستخدام طريقة التجميع. يتيح هذا التعيين للنموذج تصنيف كل ميزة على حدة، مما يسمح له باكتشاف العناصر الدقيقة في صور الأنسجة المعقدة. علاوة على ذلك ، نجحت طريقتنا في الوصول إلى دقة عالية دون الحاجة إلى تقنيات المعالجة المسبقة للصور مثل تطبيع الألوان أو تطبيع البقعة أو أي تقنيات أخرى.

ثانيًا ، لتصنيف درجة سرطان الثدي بناءً على الصور النسيجية المرضية ، نقترح استراتيجية للكشف عن درجة سرطان الثدي عن طريق دمج مجموعتي بيانات منفصلتين. تتضمن استراتيجيتنا المقترحة إضافة فئة جديدة (الدرجة 0) إلى الفئات الثلاث المعترف بها من درجات سرطان الثدي، مما يسمح لنموذجنا باكتشاف كل من الأورام الخبيثة ودرجة سرطان الثدي. لمقارنة البنى خفيفة الوزن وثقيلة الوزن، يتم تدريب نماذجنا باستخدام بنيتين مختلفتين للشبكات العصبية التلافيفية، وهما ResNet50 وMobileNet.
أخيرًا ، في كلا العملين اللذين قمنا بهما على تصنيفات سرطان الثدي، تفوقت نماذجنا في الأداء على العديد من الأعمال السابقة في تصنيف الأورام الخبيثة لسرطان الثدي وتصنيف الدرجات.

كلمات مفتاحية. سرطان الثدي، صور الأنسجة المرضية، التعلم متعدد الأمثلة، الشبكة العصبية التلافيفية، درجة سرطان الثدي.

Contents

List of Figures

List of Tables

Abbreviations

ANNs	Artificial Neural Networks
BC	Breast Cancer
CAD	Computer-Aided diagnosis system
CE	Cross-Entropy loss function
CNNs	Convolutional Neural Networks
DCIS	Ductal Carcinoma In Situ
DL	Deep Learning
FC	Fully Connected
GPU	Graphics Processing Unit
H&E	Hematoxylin and Eosin
IBC	Inflammatory breast cancer
IDC	Invasive Ductal Carcinoma
ILC	Invasive Lobular Carcinoma
LCIS	Lobular Carcinoma In Situ
MIL	Multiple Instance Learning
MILC	Multiple Instance Learning Classifier
ML	Machine Learning
MLP	Multi-Layer Perceptron
ReLU	Rectified Linear Units
ResNet	Residual Network
SVM	Support Vector Machine
Tanh	Hyperbolic tangent activation function
TPU	Tensor Processing Unit
WHO	World Health Organization
WSI	Whole Slide Imaging

Chapter 1

Introduction

1 Thesis context

Cancer is a disease where living cells divide and grow uncontrollably in an organ, and it can spread to many other parts of the body. Cancer can start nearly in all parts of the human body, which is made of trillion cells [84]. Cancer is the leading cause of death in many countries, wherein in 2018 alone, 9.6 million cancer deaths were counted. Among females, the most commonly diagnosed cancer is breast cancer, and it's the leading cause of cancer death among them. In 2018, 2.1 million newly diagnosed female breast cancer cases, accounting for almost 1 in 4 cancer cases among women[13]. To diagnose cancer, doctors use histopathological images, which are generated from a biopsy procedure where a small amount of tissue is removed from a suspected tumor to be examined under the microscope.

Histopathology whole slide imaging (WSI) is made from formalin-fixed paraffin-embedded (FFPE) tissue that contains both tumor and normal tissue. These slides are then stained with agents such as hematoxylin and eosin (H&E) and immunohistochemical stains, which allow the pathologist to identify important features. Pathologists routinely use histopathology slides to determine the type of cancer, stage of cancer, cancer grade, and presence of infiltrating immune cells. Because each cancer type necessitates a different treatment regimen, such as surgery, radiotherapy, chemotherapy, or targeted therapy, an accurate cancer diagnosis is critical for selecting effective treatment options[12].

The introduction of whole slide imaging techniques allows a pathologist to view histopathology slides digitally rather than under a microscope. A digital whole slide image (WSI) has a very high resolution (10 000 to > 100 000 pixels in each dimension) as shown in figure1.1. Viewing all details in a WSI takes time for a pathologist. As a result, some critical information may be overlooked. Furthermore, manual examinations are reliant on pathologists' experience and are subject to intra- and inter-observer variability[12]. As a result, developing a better approach to making slide-level predictions directly is critical in this area's study.

2 Thesis motivation and objectives

Breast cancer is the most commonly diagnosed cancer in women and the main cause of cancer death, and it represents 11.6% of all cancers in the globe[13]. Mammography of the breast is the first step in determining whether there are any suspected cancers or abnormalities. A biopsy is subsequently performed to get histology images, which allow pathologists to determine whether or not the tumor is cancerous based on its location as determined by the mammograms[79]. Doctors can see the aberrant cells and their rate of

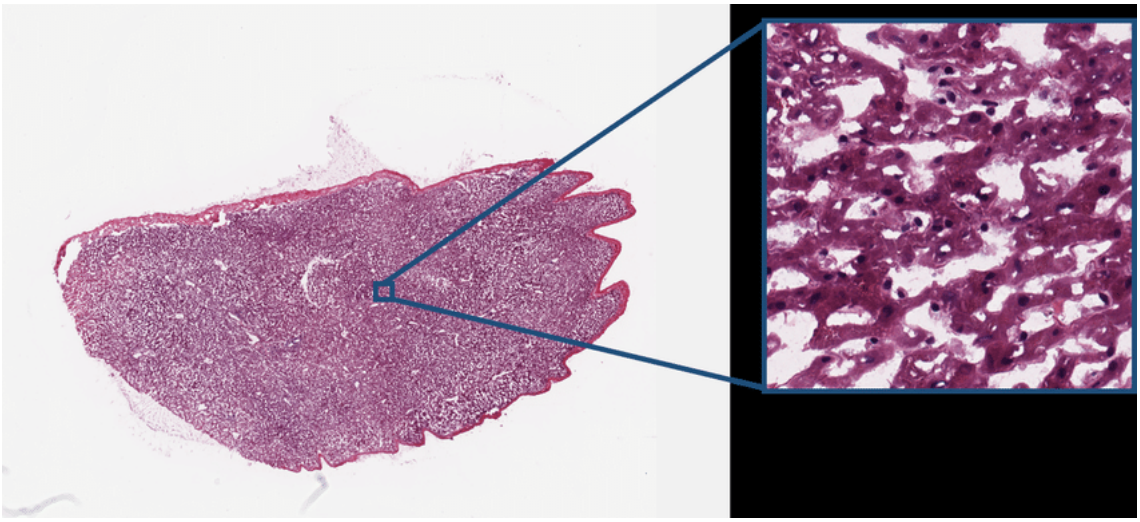

Figure 1.1: A H&E stained whole slide image

growth using histological imaging. Indeed, histological interpretations are used to guide breast cancer treatment plans[36]. These histology pictures have been magnified at different optical magnifications, which have shown great variety and a complex texture in the tumor tissue.

Since histology pictures have such wide ranges of color and texture, creating a computer-aided system (CAD) to classify breast cancer is a tough task. Images of histology vary depending on the kind of cancer and the type of cells from which they were taken[81], as shown in the figure 1.2. This complicates the computer classification of histological images, which have varying magnifications depending on the microscope zoom level from which they are taken[81], as seen in the figure 1.3.

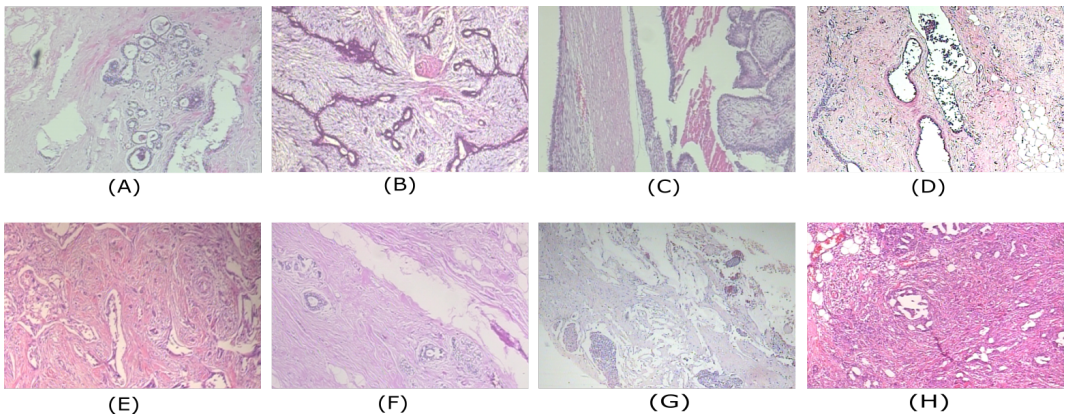

Figure 1.2: Samples a-h are histopathology images of adenosis, fibroadenoma, phyllodes tumor, tubular adenoma, ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma from the BreakHis[74] dataset in magnification factor X40.

We were inspired to write this thesis by deep learning's success in computer vision

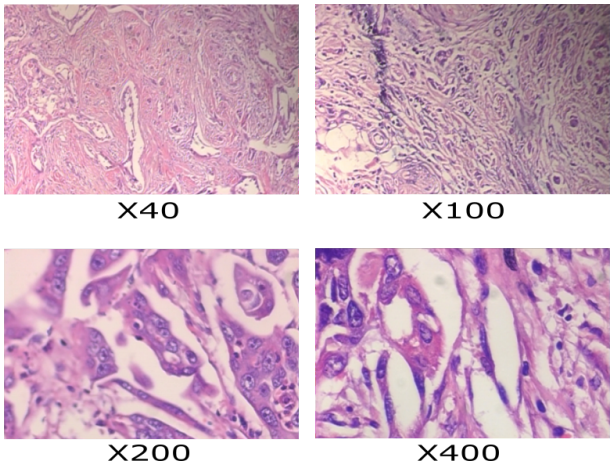

Figure 1.3: Different magnifications of breast cancer histopathology images of ductal carcinoma from the BreakHis[74] dataset.

and breast cancer classification and segmentation. Following a breakthrough in the field of computer vision in 2012, when a team developed a deep learning model called AlexNet [48], this model outperformed state-of-the-art methods and achieved the best results in the field of object recognition. Many deep learning-based research papers [16, 85, 55] in the field of breast cancer histopathological classification have been proposed since 2013.

The current state-of-the-art in deep learning image classification is based on Convolutional Neural Networks (CNNs) [50]. In typical CNN architectures [48, 70], a feature extractor with a bank of convolution filters is commonly found (i.e., trainable parameters). The layers are then pooled to make the images less sensitive and invariant to small translations, as well as to reduce the dimensionality of the feature maps using the Pooling function. A classifier is a final stage in CNN architectures, classifying each pixel (or voxel) into one of many classes.

Deep learning and CNN architectures are successful because their operations (e.g., convolution, pooling,...etc.) can be parallelized to take advantage of massively parallel architectures such as GPU and TPU implementations, also as a result of the large public datasets and the advancement of learning methods. Furthermore, the majority of the gains in machine learning and deep learning come from extracting great and relevant features, with CNNs being the best features extraction algorithm in computer vision.

Finally, despite the success of deep learning and CNNs in computer vision and brain tumor classification, they still have limitations, particularly in terms of how to design dedicated and efficient architectures in terms of computational cost and classification performance. Another well-known problem in CNNs is the combinatorial explosion of choices caused by a large number of hyperparameters (patches, feature maps, kernels, strides, activation functions, connectivity between layers, etc.), where these choices limit and affect the robustness of CNNs. Furthermore, the selection of the input shape is critical for achiev-

ing high classification performance results. The problem with input shape arises from the fact that traditional CNNs do not consider the image's global context.

Furthermore, the most common issue with image classification methods is unbalanced data, which occurs when a class or label of interest has a small amount of data in comparison to other classes. Because of such issues, Artificial Neural Networks (ANNs), including CNNs, tend to favor the more frequent label. Thus, training a CNN model with such data will result in low sensitivity predictions, where the most important part in medical applications is to make the model more sensitive toward the lesion-class, i.e., tumoral regions. Another significant issue is performance degradation caused by a variety of factors such as complex images, vanishing gradients, and overfitting. The goal of this thesis is to build an accurate deep learning model for classifying histopathological images of breast cancer by overcoming the different constraints of the literature works, like the complexity of histopathological images, the influence of hyperparameters on classification performance, the unbalanced data, and the lack of data due to patient privacy.

3 Thesis contribution

In this thesis, we aim to propose automatic, efficient, and accurate deep learning models for breast cancer classification using histopathological images. The proposed models are used to classify the breast tumors into malignant or benign classes and to classify the malignant tumors into grades. Moreover, these models could assist clinicians and pathologists in providing a third opinion diagnostics. For achieving this goal, our contributions are divided into twofold:

- A new layer we call MILC (Multiple Instance Learning Classifier) replaces the conventional Flatten layer and allows us to classify each feature separately before aggregating the results to get the image label. All the magnification factors for each class (benign, malignant) were integrated so that our suggested model can identify the malignancy of an image, regardless of the magnification factor. In addition, we propose an evaluation of our model based on the BreakHis [74] dataset for each magnification factor (X40, X100, X200, X400) to compare it with the current state of the art. The imbalanced data might be solved by applying a weight to each class based on the number of images in that class.

- We concentrated on the issue of limited dataset resources since deep learning requires a large amount of data. A common restriction in medical image classification is a lack of data resources, which leads to models that are less generalizable and less accurate. In order to surpass this limitation, we propose a new strategy based

on the addition of a new class called grade zero, which contains benign histological images from the well-known BreakHis [74] dataset combined with the IDC [22] dataset, which contains the grades classes (grade 1, grade 2, grade 3) of our entire dataset. We used the ResNet50 model, which is the best fit for the histological images characterized by the huge and complex textures. This model performs a feature extraction with skip connection that helps in preventing the vanishing gradients problem [38] which leads to the classification overfitting. Also, we used the MobileNet, which has a low number of trainable parameters, to compare it with the results of ResNet50 architectures.

4 Thesis overview

In this thesis, we are interested in the classification of breast tumors using CNNs architectures trained on histopathological images to assist the pathologists and clinicians in the therapy planning process, in particular, detecting malignant cancer and grading the tumors. The thesis is organized as follows:

chapter 2: in this chapter, we present a full detailed description of breast cancer and the anatomy of the female breast and where breast cancer can start. In addition, we demonstrated the breast cancer types (the ductal and lobular carcinoma, invasive and in situ, triple-negative breast cancer, and inflammatory breast cancer) and the different grades of breast cancer (grade 1, grade 2, and grade 3). Moreover, in this section, we present the breast cancer screening imaging used by clinicians in the diagnostics of breast cancer, where we present the mammograms, ultrasound, magnetic resonance imaging, and histopathology images.

chapter 3: is devoted to a quick overview of deep learning, with a focus on convolutional neural networks and multiple instance learning. So, in this chapter, we presented the different blocks that constitute convolutional neural networks, where we illustrated the convolution operation, which is the main function of CNN. Also, the non-linear activation functions, pooling, and the fully-connected layers are explained in this section alongside the two main CNN's algorithms, the forward and the backward algorithm. For the multiple instance learning method, we presented its different approaches (instance-based approach, embedding-based approach, and the bag-based approach) with different multiple instance learning pooling functions(max, mean, noisy-or, generalized mean, integrated segmentation and recognition, log-sum-exp, and the noisy-and).

chapter 4: in this chapter, we explained our contribution on breast cancer histopathological images classification, where we created a new strategy by combining the multiple instance learning method and the convolutional neural networks. First, in this chapter,

we demonstrated the different works in the field that used both the conventional and deep learning methods. Second, we illustrated the used loss function in our model, the dataset, and the techniques (data augmentation, transfer learning, data balancing) used to enhance the robustness of the model. Finally, we presented our results for the different models that we trained, and we compared our results to the different works of the state-of-the-art.

chapter 5: this chapter illustrates our second contribution on classifying breast cancer based on histopathological images. The chapter presents the different related works and methods used to classify the grade of cancer, and then it demonstrates the used methodology and dataset also the different techniques for enhancing the results (data augmentation, transfer learning, data balancing). The next half of the chapter presents the evaluation metrics and the obtained results compared to the different works of the state-of-the-art.

Finally, our conclusions, summary of the contributions and directions for future research are presented in **chapter 6**.

Chapter 2

Breast Cancer: Medical Overview

1 Introduction

In this chapter, we present a complete overview of breast cancer and how and where it starts. Also, we will explain all the known breast cancer types, the ductal and lobular carcinoma, invasive and in situ, triple-negative breast cancer, and inflammatory breast cancer. Finally, we show the different breast cancer imaging and screening techniques used by clinicians and pathologists to determine the malignancy, the type, and the grade of cancer, where we present the mammograms, ultrasound, magnetic resonance imaging, and histopathology images.

2 Breast cancer types

According to doctors, breast cancer develops when some breast cells grow abnormally. These cells divide faster than healthy cells and clump together, forming a lump or mass. In addition, cells in the breast may spread (metastasize) to the lymph nodes or other parts of the body. There are numerous types of breast cancer and numerous ways to describe them. It's easy to become perplexed. The type of breast cancer is determined by the specific cells in the breast that become cancerous. There are two main and most diagnosed types of breast cancer, Ductal Carcinoma (DC) and Lobular Carcinoma (LB). The ductal carcinoma begins in the cells lining the milk ducts, which transport breast milk to the nipple, and the lobular carcinoma begins in the breast milk-producing glands (lobules) [62].

2.1 Ductal Carcinoma In Situ (DCIS)

DCIS (ductal carcinoma in situ) is a type of non-invasive breast cancer. Ductal cancer begins inside the milk ducts, and in situ cancer means "in its original location." DCIS is referred to as "non-invasive" because it has not spread beyond the milk duct into the normal breast tissue ((see figure 2.1)). DCIS is not fatal, but it does increase the risk of developing invasive breast cancer later in life [62]. According to the American Cancer Society, approximately 60,000 cases of DCIS are diagnosed in the United States each year, accounting for approximately one out of every five new cases of breast cancer [72].

2.2 Invasive Ductal Carcinoma (IDC)

The most common type of breast cancer is invasive ductal carcinoma (IDC), also known as infiltrating ductal carcinoma. Each year, more than 180,000 women in the United States

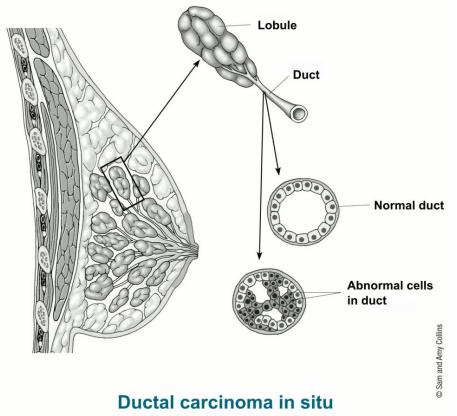

Figure 2.1: Ductal Carcinoma In Situ (DCIS). image source: breastcancer.org

are diagnosed with invasive breast cancer, according to the American Cancer Society. Approximately 80% of them have invasive ductal carcinoma [72]. The term "invasive" refers to cancer that has "invaded" or spread to the surrounding breast tissues. The term "ductal cancer" refers to cancer that begins in the milk ducts, which are the "pipes" that transport milk from the milk-producing lobules to the nipple (see figure 2.2). Overall, "invasive ductal carcinoma" refers to cancer that has broken through the milk duct wall and has begun to invade the breast tissues. Invasive ductal carcinoma can spread to the lymph nodes and possibly to other parts of the body over time[62]. There are many sub-types of IDC like Tubular, Medullary, Mucinous, Papillary, and Cribriform carcinoma.

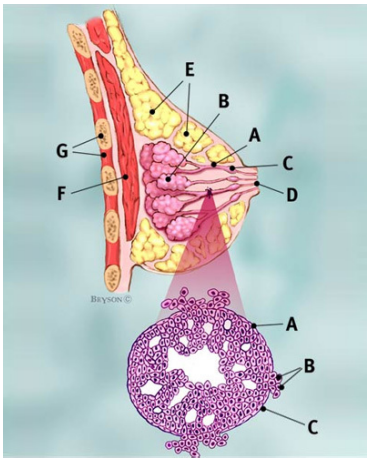

Figure 2.2: Invasive Ductal Carcinoma (IDC). A :Ducts B: Lobules C: Dilated section of duct to hold milk D: Nipple E: Fat F: Pectoralis major muscle G: Chest wall/rib cage / **Enlargement:** A: Normal duct cell B: Ductal cancer cells breaking through the basement membrane. C: Basement membrane. image source: breastcancer.org

2.2.1 Tubular Carcinoma

Breast tubular carcinoma is a subtype of invasive ductal carcinoma. Tubular carcinomas are typically small (less than 1 cm in diameter) and composed of tube-shaped structures known as "tubules." These tumors are typically low-grade, which means that their cells resemble normal, healthy cells and grow slowly[62]. Tubular carcinomas are used to account for about 1-4 percent of all breast cancers[72]. However, now that screening mammography is widely used, tubular carcinomas are being diagnosed more frequently, often before a doctor can feel a lump.

2.2.2 Medullary Carcinoma

Medullary carcinoma of the breast is a rare subtype of invasive ductal carcinoma that accounts for about 3-5 percent of all breast cancer cases. The tumor is called "medullary" carcinoma because it is a soft, fleshy mass that resembles the medulla, a part of the brain. Medullary carcinoma can develop at any age, but it most commonly affects women in their late forties and early fifties. Medullary carcinoma cells typically have a high-grade appearance but a low-grade behavior. In other words, they resemble aggressive, highly abnormal cancer cells but do not behave like them. Medullary carcinoma grows slowly and rarely spreads to the lymph nodes outside the breast. As a result, it is usually easier to treat than other types of BC[62].

2.2.3 Mucinous Carcinoma

Mucinous breast carcinoma, also known as colloid carcinoma, is a rare invasive ductal carcinoma. The tumor in this type of cancer is composed of abnormal cells that "float" in pools of mucin, a key component of the slimy, slippery substance known as mucus. Mucus normally lines the majority of the inner surface of our bodies, including the digestive tract, lungs, liver, and other vital organs. Mucus is produced by many types of cancer cells, including the majority of breast cancer cells. Mucin, on the other hand, becomes a part of the tumor and surrounds the breast cancer cells in mucinous carcinoma. Under a microscope, the cancer cells appear to be dispersed throughout pools of mucus [62]. According to research, only about 2-3% of invasive breast cancers are "pure" mucinous carcinomas, which means this is the only type of cancer present within the tumor. A mucinous component appears to be present in about 5% of invasive breast cancers, along with other types of cancer cells. Men are improbable to develop mucinous carcinoma [72].

2.2.4 Papillary Carcinoma

Breast invasive papillary carcinomas are uncommon, accounting for less than 1-2 percent of all invasive breast cancers. Most of these tumors are found in older women who have already undergone menopause. An invasive papillary carcinoma has a distinct border and is composed of small, finger-like projections. On a scale of 1 to 3, grade 2 describes cancer cells that look and behave somewhat like normal, healthy breast cells, while grade 3 describes extremely abnormal, fast-growing cancer cells. DCIS (ductal carcinoma in situ) is present in the majority of cases of invasive papillary carcinoma [62].

2.2.5 Cribriform Carcinoma

Cancer cells invade the stroma (connective tissues of the breast) in nestlike formations between the ducts and lobules in invasive cribriform carcinoma. There are different holes between the cancer cells within the tumor, giving it the appearance of Swiss cheese. Invasive cribriform carcinoma is usually of low grade, which means that its cells resemble normal, healthy breast cells in appearance and behavior. In about 5-6 percent of invasive breast cancers, some portion of the tumor is cribriform. Typically, some cribriform ductal carcinoma in situ (DCIS) is also present [62].

2.3 Lobular Carcinoma In Situ (LCIS)

LCIS is an area (or areas) of abnormal cell growth that increases a person's risk of developing invasive breast cancer later in life. Lobular refers to the fact that the abnormal cells begin to grow in the lobules, which are milk-producing glands located at the end of the breast ducts. The term "in situ" refers to an abnormal growth that remains within the lobule and does not spread to surrounding tissues. People with LCIS usually have more than one lobule affected [62]. LCIS is most commonly diagnosed between the ages of 40 and 50 before menopause. Less than 10% of women diagnosed with LCIS have already reached menopause. LCIS is extremely rare in men [72].

2.4 Invasive Lobular Carcinoma (ILC)

This type of breast cancer, also known as infiltrating lobular carcinoma, is second only to invasive ductal carcinoma as the most common form of breast cancer in the United States; according to the National Cancer Institute (NCI), this type of cancer began in the milk-producing lobules [62]. According to the American Cancer Society (ACS), more than 180000 women in the US are diagnosed each year with invasive breast cancer. Nearly

10% of all invasive breast cancers are lobular carcinomas, and invasive ductal carcinomas make up around 80% of all breast cancers [72].

2.5 Inflammatory breast cancer (IBC)

Inflammatory breast cancer (IBC) is a rare and deadly type of breast cancer. According to the American Cancer Society, inflammatory breast cancers account for about 1% of all breast cancer cases in the United States [72]. Instead of a distinct lump, inflammatory breast cancer usually begins with reddening and swelling of the breast. IBC grows and spreads rapidly, with symptoms worsening in days or even hours. Therefore, it is critical to recognize signs and seek treatment as soon as possible [63].

In the United States, the average age at diagnosis for inflammatory breast cancer is 57 for white women and 52 for black women. These ages are approximately five years younger than the average age of diagnosis for other types of breast cancer. Inflammatory breast cancer is more common in Black women, according to the American Cancer Society [72].

2.6 Male Breast Cancer

Male breast cancer is a rare disease. Men account for less than 1% of all breast cancer cases. In 2021, approximately 2,650 men were expected to be diagnosed with breast cancer, with an estimated 530 men dying from the disease. The lifetime risk of being diagnosed with breast cancer for men is approximately 1 in 833 [4].

3 Breast cancer grades

When cancer cells are removed from the breast and tested in the lab, they are graded. The grade is determined by how much the cancer cells resemble normal cells. The grade is used to help predict your outcome (prognosis) and decide which treatments may be most effective. A low-grade number (grade 1) usually indicates that cancer grows slowly and is less likely to spread. A high-grade number (grade 3) indicates cancer that is rapidly growing and likely to spread. Cancer with an intermediate grade number (grade 2) grows faster than a grade 1 cancer but slower than a grade 3 cancer [11].

There are various "scoring systems" for determining the grade of breast cancer. The Nottingham Histologic Score system (also known as "the Elston-Ellis modification of the Scarff-Bloom-Richardson grading system") is one of these. There are three factors that pathologists consider when using this scoring system (see figure 2.3 [26]:

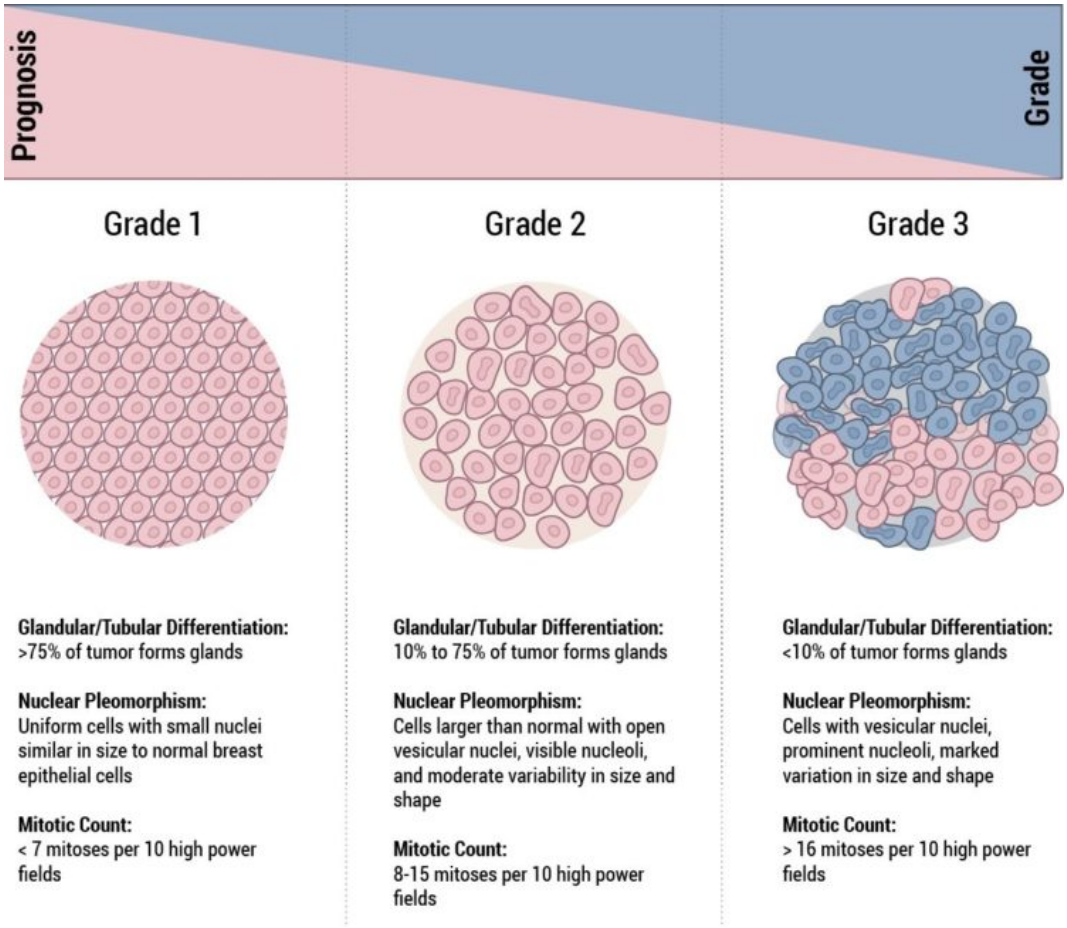

Figure 2.3: Breast cancer grades scoring (Nottingham Histologic Score system). image source: Johns Hopkins pathology depatement.

- The amount of gland formation (cell "differentiation," or how well tumor cells try to replicate normal glands).
- The nuclear characteristics (the degree of "pleomorphism" or how "ugly" the tumor cells look).
- The activity of mitotic cells (how much the tumor cells are dividing or proliferating).

Each of these features is scored on a scale of 1-3, and the scores are added to give a final total score ranging from 3 to 9. The final total score is used to calculate the grade as follows[26]:

- Tumors in grade I have a total score of 3-5.
- Tumors in grade II have a total score of 6-7.
- Tumors in grade III have a total score of 8-9.

4 Breast cancer screening and imaging

Screening exams detect disease before symptoms appear. The goal of screening is to see the disease at its most treatable and early stage. A screening program must meet a number of criteria, including reducing the number of deaths from the given disease, in order to be widely accepted and recommended by medical practitioners. There are a lot of screening techniques like mammography (Mammograms images), ultrasound, and breast cancer magnetic resonance imaging. After the screening, if a tumor is localized, a biopsy procedure is needed to generate histopathological images and to detect a malignant tumor (if it exists) stage and its grade.

4.1 Mammograms

Using a low dose of X-ray radiation, mammographers can view the inside of the breasts (see fig. 2.4). Early detection and diagnosis of breast disorders can be made possible with a mammography exam, which is commonly known as a mammogram. Doctors use X-rays to help them diagnose and treat patients. To obtain photographs of the inside of your body, a little amount of ionizing radiation is used during the procedure. Radiography is the most often used and oldest form of medical imaging. Recent improvements in mammography include digitalization, computer-aided detection (CAD), and breast tomosynthesis [43].

It's called full-field digital mammography (FFDM) since the X-ray film has been replaced with electronic devices capable of producing mammographic images of the patient's breast. It is possible to achieve better photographs while using less radiation with devices like this since they are similar to those in digital cameras. The radiologist will be able to review and store these images on a computer for future reference. In terms of the patient's experience, digital mammography is comparable to a standard film mammogram [43].

An abnormality in the density, mass, or calcification deposition in digital mammograms can be detected using computer-aided detection (CAD) systems. In these images, the CAD system identifies these spots, which prompts the radiologist to pay particular attention to them [43].

As a form of sophisticated breast imaging, breast tomosynthesis, also known as 3-D mammography and digital breast tomosynthesis (DBT), uses several images to create a three-dimensional image set. Similar to computed tomography (CT) imaging, in which a series of thin "slices" of the body are stitched together to generate a 3-D reconstruction, 3-D breast imaging is similar in this regard [43].

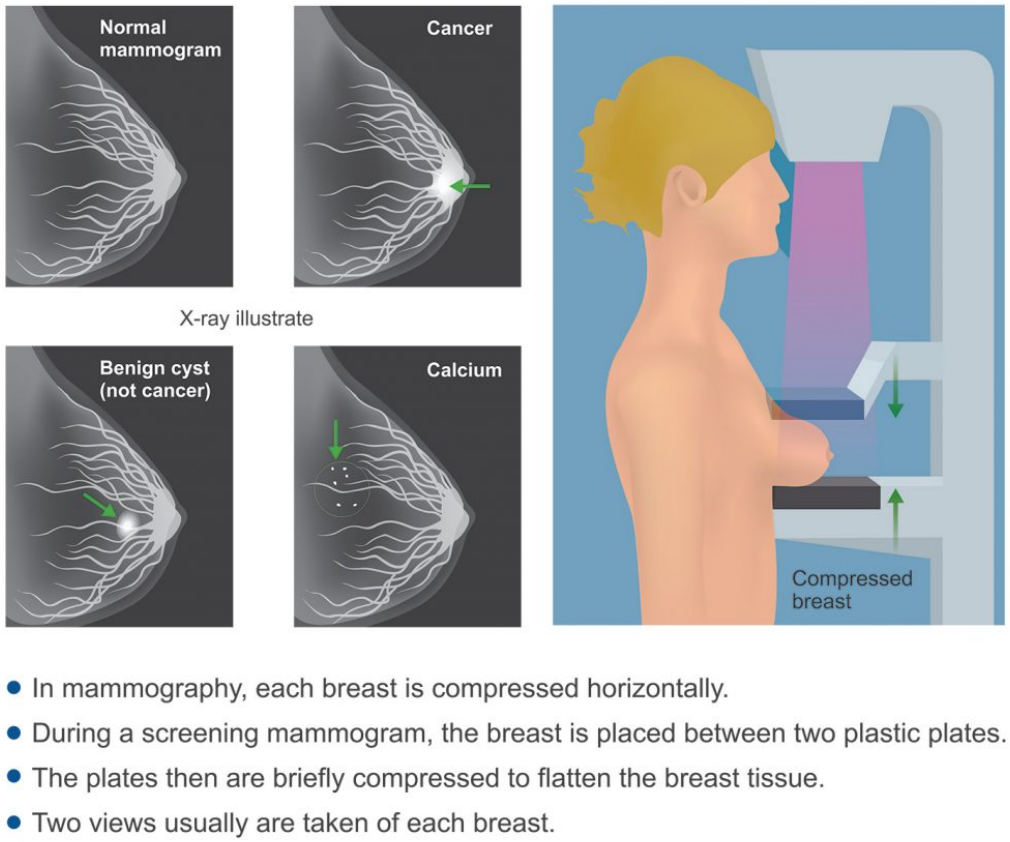

Figure 2.4: Breast mammography. image source: Green Imaging.

4.2 Ultrasound

Doctors use ultrasound imaging to diagnose and treat a variety of medical disorders. It's completely risk-free and doesn't cause any discomfort. It employs sound waves to create images of the human body's inside. An ultrasound technique known as Sonography is referred to as ultrasound imaging. It uses a transducer and a gel that is applied directly to the skin to achieve its results. Through the gel, high-frequency sound waves enter the body from the probe. The probe picks up the sounds that are reflected back at it. A computer uses these sound waves to create an image. Radiation (X-rays) is not used in ultrasound examinations. Figure 2.5 shows how real-time ultrasound imaging can provide insight into the interior organ structure and movement. In the ultrasounds imaging, the blood can be seen flowing via blood vessels [43].

4.3 Breast MRI

Magnetic resonance imaging (MRI) of the breast employs a strong magnetic field, radio waves, and a computer to generate detailed images of the structures within the breast (see

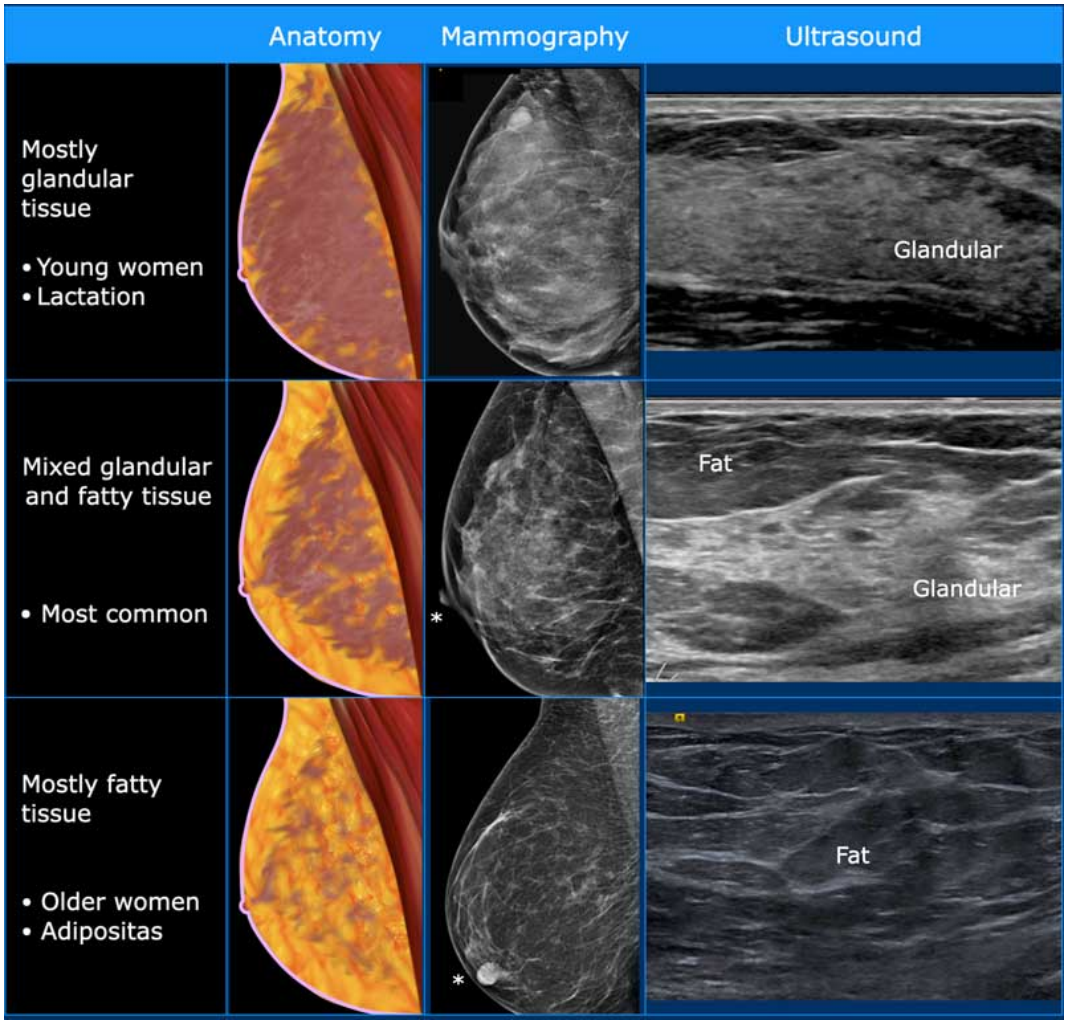

Figure 2.5: Breast Mammography vs Ultrasound. image source: The Radiology Assistant.

figure 2.6. It is primarily used as a supplement to mammography or ultrasound for breast screening. It could be used to screen women at high risk for breast cancer, assess the extent of cancer after a diagnosis, or investigate abnormalities detected on mammography. Radiation is not used in MRI (x-rays) [43].

5 Histopathology images

Histopathology is the study of disease symptoms through microscopic examination of a biopsy or surgical specimen that has been processed and fixed onto glass slides. The sections are stained with one or more stains to allow the different components of the tissue to be seen under a microscope [37]. Pathologists examine the tissue under a microscope slide at different magnification levels to identify morphological characteristics that indicate the presence of diseases such as cancer. However, such a prognosis is subjective. Then, a

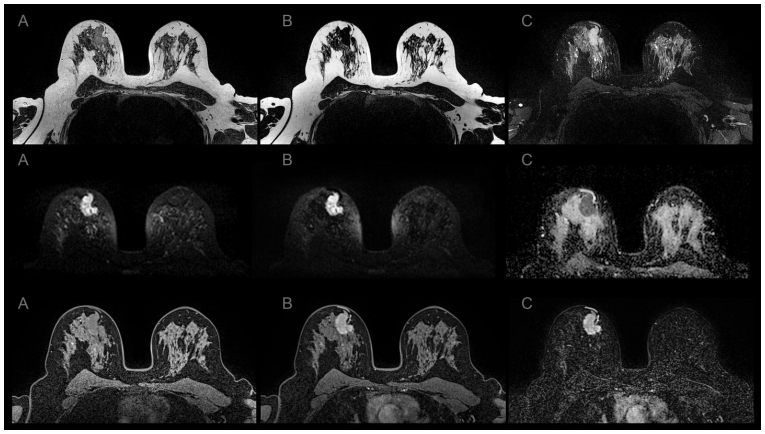

Figure 2.6: Breast MRI. image source: Siemens-Healthineers.

quantitative evaluation of these Whole Slides Images (WSIs) is required for an objective diagnosis. Many computer-assisted disease systems have been developed to aid histological diagnosis and reduce subjectivity in the last ten years. By extracting features from histological images, all these systems attempt to mimic pathologists [42].

5.1 Acquisition

After the breast tumor is removed or biopsied in the operating room, the specimen is submitted to the pathology laboratory for further examination. Tissue preparation begins with formalin fixation and paraffin embedding, which is the initial step in the process. Next, a microtome is used to cut sections of paraffin blocks with a thickness of 3-5 m and mount them on glass slides (a high-precision cutting instrument). The nuclei and cytoplasm of interest in the tissue, which are ordinarily visible on the mounted sections, are obscured. So they have to be stained with dyes that bring out their features. Figures 1.3, 1.2 show the most popular staining techniques, which is the Hematoxylin and Eosin(H&E). All patients' diagnostic and prognostic procedures begin with H&E staining, even though this method has been in use for more than a century. DNA is stained with hematoxylin, which binds to DNA, whereas proteins are stained with eosin, which binds to proteins (cytoplasm, stroma, and so on)[62].

Using a combination of microscope-based hardware and software, whole-slide images can be captured easily. In addition, whole-slide images can be generated using advanced research microscopes, with the necessary auxiliary hardware, such as motorized stages, cameras, etc. These microscopes can be multipurpose laboratory devices that take high-resolution, 2D or 3D compatible whole slide images [27].

5.2 Preprocessing

Histopathological images are pretty large and complex in structure. As a result, they pose a challenge to machine learning algorithms. Because these high-resolution images contain so much information about image texture, they provide the most accurate diagnosis for almost all cancers[69]. These images are analyzed using advanced image analysis methods. The primary goals of these methods are to assist the expert in decision-making, to provide consensus among experts, to gain time for the expert, and to identify image patterns that specialists find difficult to notice [82]. However, high-resolution image analysis takes a long time. Simultaneously, the complexity of the background and distracting factors can slow down the processing speed and reduce success. Image preprocessing algorithms aid in avoiding this undesirable situation.

There are a lot of techniques used for preprocessing the histopathological images before using them in computer-aided diagnosis (CAD), one of the most used techniques is stain normalization, which is an important processing task for CAD systems in modern digital pathology. This task reduces the color and intensity variations present in stained images from different laboratories. This technique is used to transfer the color distribution of the source image to that of the target image, where will all the input images have the same distribution. In this way, the CAD system achieves better accuracies, but at the same time, it has computational challenges that this normalization step must overcome, particularly for real-time applications, where the memory and run-time bottlenecks associated with high-resolution image processing. Furthermore, stain normalization can be affected by the quality of the input images, for example, when they contain stain spots or dirt[5].

Color matching Reinhard[61] and stain separation Macenko[51], and Vahadane[80] have been employed, where Reinhard et al. presented a strategy based on color transfer between a standard image and a color-variable image by utilizing the mean and variance of both images. The source image is changed to the target image using this procedure. In this case, the color distribution of the source image is transformed to that of the target image using a linear transform in perceptual color space[61]. Macenko et al. proposed a method for determining specific stain vectors for each image based on the colors contained in the image. There is a particular stain vector in this approach that corresponds to each of the two stains in the image, and it is a completely automated method that is suited for evaluating several slides quickly due to having very few parameters and no optimizations necessary[51]. Vahadane et al. proposed to divide the image into sparse and non-negative stain density maps. The stain density maps are then merged based on the stain color of a pathologist's selected target image. As a result, only the color is changed, but the structure remains unchanged[80].

Today, Generative Adversarial Networks (GANs) [34] are the cutting-edge technology in stain normalization. A generator changes an image from domain A to domain B. A discriminator network attempts to recognize genuine domain B images from forgeries, hence assisting the generator in improving. Zhu et al. proposed a CycleGAN[86] , which is a GAN-based architecture where in this network, there are two generators: one that converts from domain A to domain B and another that converts from domain B to domain A. The purpose of these two models is to be able to recreate an original image in the following directions: A -> B -> A or B -> A -> B. Discriminators are also used by CycleGANs to forecast real versus produced images for each domain.

6 Conclusion

In this chapter, we presented an overview of breast cancer, where we demonstrated the different breast cancer types that can affect women (also men in rare cases). There are two main types of cancer, ductal carcinoma and lobular carcinoma, and both can be invasive or in situ. Also, we have presented the different grades of BC, wherein most scoring systems they three grades (low, medium, and high grade). Moreover, we showed the different screening techniques and imaging used in breast cancer (mammography, ultrasound, and magnetic resonance imaging). Finally, we discussed the histopathological images, which are the current trend in which large amounts of visual data are made available for automatic analysis. It enables the visualization and interpretation of pathologic cell and tissue samples using high-resolution images and computer tools.

The histopathological images are known for their complexity and variations, where sometimes, even experience pathologists find the interpretation of these images very hard and time-consuming. This opens the door to developing image analysis methods of artificial intelligence that assist pathologists and support their image descriptions and classification (e.g., staging, grading).

In the following chapter, we will discuss Convolutional Neural Networks (CNNs) and how we can use them to classify breast cancer using histopathological images. Furthermore, we present the fundamental building blocks of the CNNs algorithm, as well as the operations that power it, such as convolution, non-linear activation function, pooling, flattening, and fully-connected layers. Furthermore, we present two essential algorithms: the forward and backward algorithms.

Chapter 3

Background Theory

1 Introduction

Deep learning is a subfield of Machine Learning that enables computers to learn from experience and data, and machine learning is a subfield of Artificial Intelligence (AI). AI is a collection of fields investigating how machines can mimic human cognitive functions such as learning and problem-solving. Machine learning is essentially concerned with how to program a computer without explicitly programming it. Deep Learning, in fact, dates back to the 1940s with the work of (McCulloch and Pitts, 1943[54]; Hebb, 1949[39]; Rosenblatt, 1958[64]), which consisted of a set of algorithms inspired by brain function and the idea of stacking many layers of neural networks to transform raw data from one representation space to another with more compact and discriminative features.

CNNs (Fukushima, 1980[28]; LeCun et al., 1989 [49]) have demonstrated in recent years that show they are an effective technique for detecting and extracting features in a wide range of computer vision tasks. LeNet5, developed by (LeCun et al., 1998[50]), was one of the first known CNNs architectures, appearing after many previous versions of CNNs since the work of handwritten zip code recognition in 1989 [49].

In this chapter, we present the fundamental building blocks of Convolutional Neural Networks, such as convolution, non-linear activation function, pooling, fully-connected layers, as well as forward and backward algorithms. Also, we present the Multiple Instance Learning (MIL) technique, where we explain its different approaches and the different MIL pooling functions.

2 Convolutional Neural Networks

CNN networks are typically applied to images or, more broadly, matrices because images are spatially correlated, and the image's features are distributed across the entire image. As a result, CNN networks have taken advantage of these properties and added a new layer known as convolution. This convolution layer is applied to the images that are fed into it (raw pixels). The addition of a convolution layer before the classifier allowed for the extraction of relevant features at multiple locations. Kernels in each CNNs layer play an important role as a feature detector by stacking many convolution layers (a pooling layer can be added instead of convolution). So, by selecting a specific kernel, we can detect the desired feature. Furthermore, the kernels' parameters represent the weights that must be updated in order to obtain the best configuration; in other words, the kernels are tuned during the training process to force them to respond strongly to specific features. The Backpropagation [66] algorithm is frequently used to update weights. Whereas lower kernels' parameters at lower-layer CNNs networks are trained to detect features like edges

and corners, higher kernels' parameters at higher-layer CNNs networks are trained to detect more complex or high-level features like shapes and object parts. CNN's algorithm is built on four standard operations: convolution, pooling, non-linear activation function, and multi-layer perceptron (MLP) [65], as well as two algorithms for computing the loss function and new parameters: forward and backward algorithms.

2.1 CNN's layers and operations

A simple CNN is a series of layers, and each layer of a CNN uses a differentiable function to transform one volume of activations into another. CNN architectures are built using three main types of layers: Convolutional Layer, Pooling Layer, and Fully-Connected Layer (see figure 3.1).

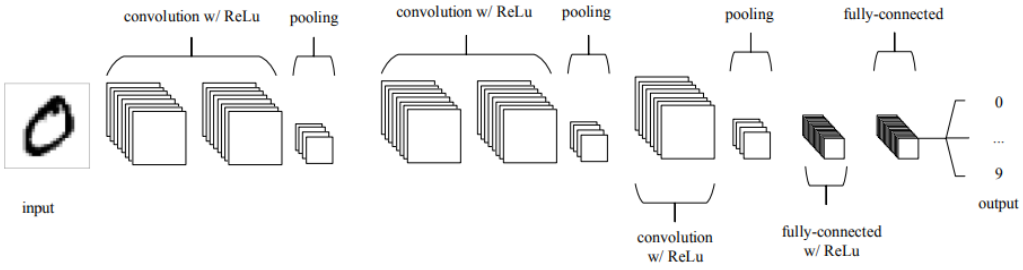

Figure 3.1: A simple CNN architecture. Image source: [56]

2.1.1 Convolution layer

An image is a grid or matrices of numbers to a computer, so to perform a Convolution operation on the input image, a set of filters (also known as Kernels) is applied to raw pixels. Furthermore, while applying filters to images is well known in conventional methods such as the Sobel filter, Gaussian filter, high-pass filter, and so on, the newly added property of CNNs in comparison to conventional methods is that CNNs use many filters with trainable parameters in successive layers. As a result, the latter property enables CNNs to outperform traditional methods by extracting hierarchical features across the entire image using the same filters (shared weights) and only some input units contributing to the output unit (sparse connectivity). Convolution also has the advantage of being applicable to a wide range of multidimensional data, and computer vision tasks [33].

Furthermore, the filters used in many stackable convolutional layers (other layers can be added alternatively with convolution) function as a feature detector. So, by selecting a specific filter, we can detect the desired feature. The filter parameters represent the weights that must be updated in order to obtain the best configuration, or, in other

words, the filter tuned during the training process to respond strongly to a specific feature. Backpropagation[66] algorithm is frequently used for weight updating. The goal of backpropagation is to minimize the loss function (also known as an objective function) by adjusting the parameters (weights and biases) towards the global minimum. Whereas lower-level filter parameters are trained to detect features like edges and corners, higher-level filter parameters are trained to detect more complex or high-level features like shapes and object parts. The result of a convolution operation is known as a feature map, and it is calculated as follows [33]:

$$Y_i = b_i + \sum_j W_{ij} * X_j \tag{3.1}$$

Where Y denotes a feature map (the result of a convolution operation), b_i denotes a trainable parameter bias, and W_{ij} denotes trainable parameter weights. The kernel matrices parameters in CNN architectures are W_{ij}. X is the input, and * is the convolution operator. In contrast to traditional feed-forward neural networks, which have dense connectivity, the convolution layer reduces the number of dot-product operations by using the property of sparse interactions, in which all input units are connected to all hidden layer units. Convolution works by convolving the filter with the input, where the filter begins with the leftmost part and slides to the right via a step called Stride after computing the elementwise production between the filter elements and the input elements. The results are then added together, along with a bias term, to form an output unit (see figure 3.2). This process is repeated until the filter has slid all of the input positions and computed all of the output units. Each filter is initialized with different weights to force it to look for a distinct feature on the input. Multiple filters are used in standard CNN architecture for each layer, and after calculating the feature map, they are grouped together to form the output [33].

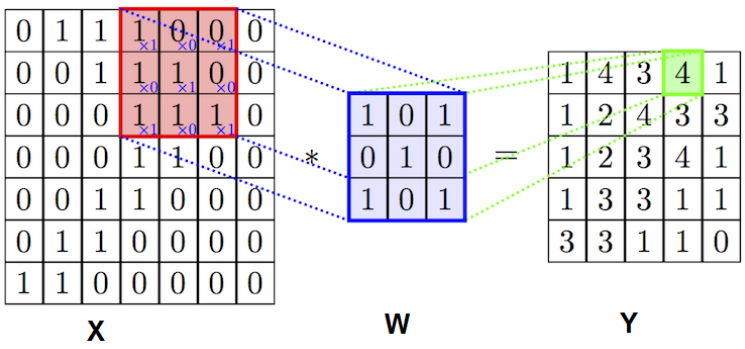

Figure 3.2: An example of a Convolution Operation on a single-channel image (Equation 3.1). The result of convolving the input X with a filter W is the feature maps Y.

In addition to the size of the Input Image, three parameters control the size and number

of feature maps [33]:

Depth: the number of filters used for the Convolution operation affects the depth of the output (Convolved feature). In the standard CNNs architecture, we use many filters in each layer, resulting in many feature maps, which are then stacked as 2D matrices.

Stride: the amount by which a kernel is moved across the input image with each step.

Zero-padding: this is the operation of adding zeros around the border (outside the matrix) to control the size of the feature map (the output); this operation is also known as wide convolution.

More formally, a convolution output O is described by kernel K, I Input volume size, Padding P and Stride S, and it's calculated as follows (also, see a concrete example in figure 3.3):

$$O = \frac{(I - K + 2P)}{S} + 1 \tag{3.2}$$

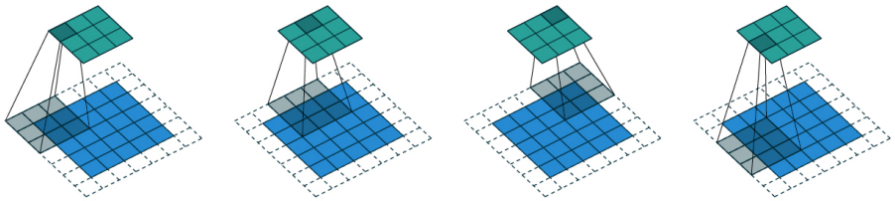

Figure 3.3: An example of the convolution layer, an Input image I= 5 × 5 (blue), Kernel K= 3 × 3 (gray), Padding P=1 ,Strides S = 2, the produced output O = 3 × 3 (green). Image source: [25]

2.1.2 Activation functions

Just after the convolution operation, we use another important operation known as the nonlinear activation function to separate and transform the feature maps from a space representation to another representation with more discriminative feature maps. There are numerous activation functions in the literature (See figure 3.4): Sigmoid (see equation 3.3) has a function curve in the range [0,1], Hyperbolic tangent (Tanh) (see equation 3.4) has a function curve in the range [-1,1], and Rectified Linear Units (ReLU) (see equation 3.5) has a function curve in the range $[0, \infty[$.

$$F(w \times x + b) = \text{Sigmoid}(w \times x + b) = \frac{1}{1 + e^{-(w \times x + b)}} \tag{3.3}$$

$$F(w \times x + b) = \text{Tanh}(w \times x + b) = \frac{e^{(w \times x + b)} - e^{-(w \times x + b)}}{e^{(w \times x + b)} + e^{-(w \times x + b)}} \tag{3.4}$$

$$F(w \times x + b) = \text{Relu}(w \times x + b) = \begin{cases} w \times x + b & , if(w \times x + b) > 0 \\ 0 & , \text{otherwise} \end{cases}. \quad (3.5)$$

Where $(w \times x + b) \in R$ is the neuron's output, w and b are trainable parameters weight and bias, respectively. F is the activation function that converts the neurons' output from linear space to non-linear space.

The Tanh, Sigmoid, and ReLu functions are naturally used in neural network theory as the activation function of a neural unit [17], and they have recently been adopted by almost all deep learning networks. When used with gradient-based algorithms to adjust neural network parameters, such as stochastic gradient descent and its variants, Sigmoid and Tanh activation functions cause the problem of vanishing or exploding gradient [31].

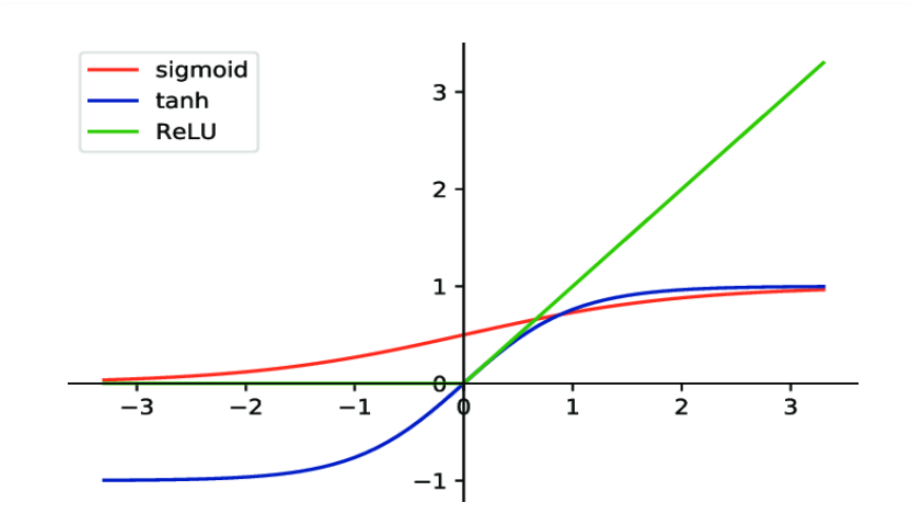

Figure 3.4: An illustration of Sigmoid, Tanh, and Relu function curves. Image source: [31]

Today, the ReLU function is the most commonly used activation function in neural networks [32]. ReLU has a significant advantage over other activation functions in that it does not activate all neurons at the same time. The image for the ReLU function above shows that it converts all negative inputs to zero and does not activate the neuron. Because only a few neurons are activated at a time, it is very computationally efficient. It does not reach saturation in the positive region. In practice, the ReLU activation function converges six times faster than the Tanh and sigmoid activation functions [48].

2.1.3 Pooling layer

The pooling operation entails sliding a two-dimensional filter over each channel of the feature map and summarizing the features that fall within the filter's coverage region. The

feature maps' dimensions are reduced by pooling layers. As a result, it reduces the number of parameters to learn as well as the amount of computation done in the network. It also summarizes the features found in a region of a feature map generated by a convolution layer. As a result, subsequent operations are carried out on summarised features rather than precisely positioned features generated by the convolution layer. As a result, the model is more resistant to changes in the position of the features in the input image [33]. There are many types of Pooling: Max, Average, ...etc. The pooling operation output size is calculated as follows:

$$O = \frac{(I - W)}{S} + 1 \tag{3.6}$$

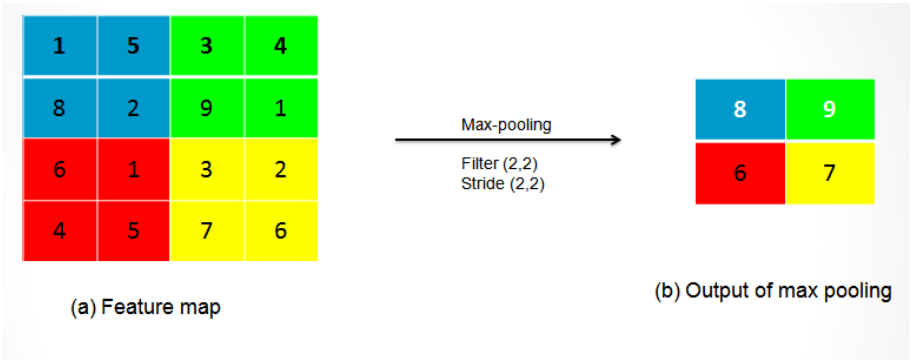

Figure 3.5: An example of max pooling

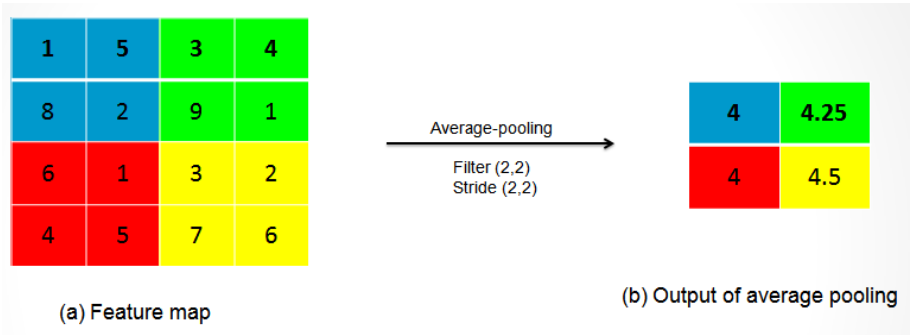

Figure 3.6: An example of average pooling

In the examples (see figures 3.5, 3.6), we shift the 2 x 2 Pooling window by two cells (called Stride) across the rectified feature map, then we take the maximum/average value in each region, and then shift the window again to the right. We repeat this process until the Pooling window has shifted all of the feature maps' positions.

2.1.4 Fully connected layers

In all the convolutional neural networks, the output of the final Pooling or Convolution layer will be flattened and fed into the Fully Connected (FC) layer as input. FCs are a collection of non-linear functions that are dependent on one another. Each individual function is made up of a neuron (or a perceptron). The neuron is fully connected layers and applies a linear transformation to the input vector via a weights matrix. The product is then subjected to a non-linear transformation via a non-linear activation function f. The FC layers are simply a Multi-Layer Perceptron (MLP) with three types of layers (see 3.7): input, hidden, and output. The input layer (feature maps) is a flattened vector of features that can be either integer or float. The hidden layers extract more representative features by transforming them from a non-linear space representation to a space representation with more compact and discriminative features. An embedded classifier is attached at the end of the output layer to classify the extracted features from the input into one of the predefined classes[33]; in our case, for the issue of breast cancer classification, there are two classes: benign or malignant, and there are three classes in breast cancer grading (grade 1, grade 2, grade 3).

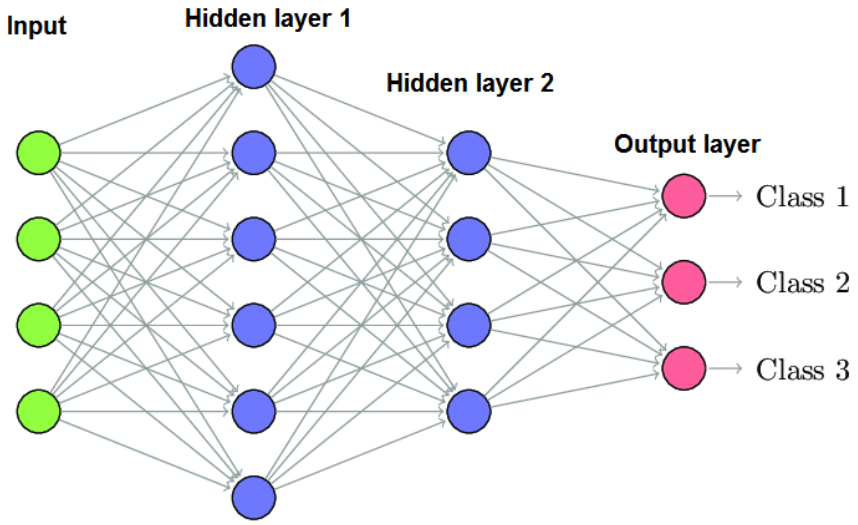

Figure 3.7: An example of a fully connected layers or Multi-layer perceptron (MLP)

Softmax (see equation 3.8), Sigmoid (see equation 3.9), or other functions can be used as classifiers. The classifier's embedded functions should have an important property that squashes the outputs to be between 0 and 1 for classification purposes. After computing the output layer's probability vector, we assign the winning or most likely true class C to the one with the highest probability [33]. The output of a j^{th} neuron in an MLP network

is computed as follows:

$$s = \sigma(Wx + b) \tag{3.7}$$

Where $s \in R$ represents the neuron's output, x represents the input, $W \in R^{f \times h}$ represents the weight matrix, and $b \in R^h$ represents the bias vector. The activation is represented by σ. Typically, the ReLu function and its variants, such as leaky ReLu, are used as an activation function of hidden layers in deep learning networks.

$$\text{Softmax}(x)_i = p(y = i \mid x; \theta) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} \quad i \in [C_0, \ldots, C_N] \tag{3.8}$$

$$\text{Sigmoid}(x) = p(y = i \mid x; \theta) = \frac{1}{1 + e^{-x}} \quad i \in [C_0, C_1] \tag{3.9}$$

Where $x \in R$ is the input to the classifier, i represents one of the classes C_i in the dataset, N represents the total number of classes, and θ is the weights of the neural networks. $(y \mid x; \theta)$ is the condition of the probability $p(y \mid x; \theta)$ that the prediction y is i, given x and θ.

2.1.5 Forward and backward propagation

There are two essential algorithms in the deep neural networks family, including convolutional neural networks: forward propagation and backward propagation. Deep learning architectures can use the first algorithm (Forward Propagation) to compute and predict the probability $p(y = C \mid x)$ that a class "C" is true. On the other hand, the goal of the backward algorithm is to compute the new parameters (weights and biases) associated with each layer and connection between every two successive neurons, with the goal of finding the best weights "W" and biases "b" that give the best prediction "p" over all the inputs.

2.1.5.1 Forward propagation

A training set contains a pair of input observations $X_{1:m}$ and corresponding targets $Y_{1:m}$ of a dataset $(X_{1:m}, Y_{1:m})$, where the class $Y \in [C_1, ..., C_n]$. During the training phase, the forward algorithm's goal is to propagate data X from the input layer to the hidden layers, then to the output layer, and to calculate the error "E" relative to the desired output Y using a differentiable Loss function L; also known as cost, or objective function. For a regression problem, for example, the Mean Squared Error (MSE) function is computed as follows:

$$L_{\text{MSE}}(y, p)_i = \frac{1}{n} \sum_{i=1}^{n} \|y_i - p_i\|^2 \tag{3.10}$$

We use a binary Cross-Entropy (which computes the difference between true and es-

timated distributions) in a binary classification problem with two classes (C = 2). CE is also known as Logistic Loss, Log-Loss, and Multinomial Logistic Loss. The following is how the CE loss function is computed:

$$\begin{aligned} L_{CE}(y,p) &= -\sum_{i=1}^{C=2} y_i \log(p_i) \\ L_{CE}(y,p) &= -(y_1 \log(p_1) + y_2 \log(p_2)) \\ L_{CE}(y,p) &= -(y_1 \log(p_1) + (1-y_1)\log(1-p_1)) \end{aligned} \tag{3.11}$$

Where "p" is the deep learning model prediction and "y" is the ground truth for each class "i" in "C", "p_1" and "y_1" are the prediction and ground truth of C_1, respectively, and "o_2" = $(1 - o_1)$" and "$y_2 = (1 - y_1)$" for "C_2". In a binary classification problem, "p" is the result of an activation function (Softmax or Sigmoid, see equations 3.8 or 3.9).

In the case of a multi-class classification with more than two classes (C > 2):

$$L_{CE}(y,p) = -\sum_{i=1}^{C} y_i \log(p_i) \tag{3.12}$$

In a multi-class classification problem, one-hot encoding is typically used to encode the labels (ground truth) over the "C" classes; this encoding gives all probabilities of the ground truth to one class, and the rest classes become zero (e.g. v = [0,0,1,0]), "v" indicates that in this case, the third class is the correct class, the equation 3.12 can be written as follows:

$$\begin{aligned} L_{CE}(y,p) &= -\log\left(\frac{e^{p_+}}{\sum_j^C e^{x_j}}\right) \\ L_{CE}(y,p) &= -\log\left(e^{p_+}\right) + \log\sum_j^C e^{x_j} \end{aligned} \tag{3.13}$$

Where L_{CE} is cross-entropy loss function of the output "$i \in [1, ..., C]$", "y" is the real value (Ground truth), Where LCE is the cross-entropy loss function of the output "$i \in [1, ..., C]$", "y" is the real value (Ground truth), and "p" is the neural network prediction (p_+ is the positive class prediction). Furthermore, the loss function's role is to calculate the distance between the ground truth and the prediction value across the entire training set. If neural networks have multiple outputs, the total loss function is the sum of all error values from all outputs, and it's calculated as follows:

$$L_{\text{total}} = \sum_{i=1}^{C} E_i = E_1 + E_2 + \ldots + E_C \tag{3.14}$$

Where E_i is the error for the i_{th} output, with "$i \in [1, ..., C]$". The total loss function of the equation 3.10 for a binary classification problem in neural networks can be written as follows:

$$L_{\text{total}} = \sum_{i=1}^{2} E_i = E_1 + E_2 = \frac{1}{2}(y_1 - p_1)^2 + \frac{1}{2}(y_2 - p_2)^2 \tag{3.15}$$

In neural networks, the input data can be propagated in batches "b" ($b \in [1...n]$), in this case, the total loss function is is the sum of losses of all batches, which computed as follows:

$$L_{\text{total},b_1:n} = \sum_{i=1}^{n} L_{bi} = L_{b1} + L_{b2} + ... + L_{bn} \tag{3.16}$$

The forward algorithm is capable of estimating the value of the error in each iteration using a loss function dedicated to each task, as well as producing input inference to obtain an output.

2.1.5.2 Backward propagation

The Backpropagation or backward propagation of errors algorithm calculates the gradient of the error from the output layer to the input layer using chain rule (a technique for computing the derivative of a composite function). The following is how the chain rule is calculated:

$$\begin{gathered} \frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} \\ \frac{dy}{dx} = \text{derivative of y with respect to x} \\ \frac{dy}{du} = \text{derivative of y with respect to u} \\ \frac{du}{dx} = \text{derivative of u with respect to x} \end{gathered} \tag{3.17}$$

Each layer in a Neural Network is a function of the layer before it; thus, to compute the derivative of a parameter "W" we must first compute the derivative of all parameters preceding "W". Updating the weights and biases for many epochs reduces the error between the predicted output and the real value; therefore, to compute the new parameters, we must propagate the gradient of the error using the chain rule method. Lets consider the example in the figure 3.8. To calculate the best parameters of W_{21}, W_{22}, b, first, we must compute the gradient of the loss function, and then we must propagate the gradient from the output layer of the neuron to the input layer.

The partial derivative $\frac{\partial L_{\text{total}}}{\partial W_{21}}$ (also known as the gradient) of the total error L_{total} with

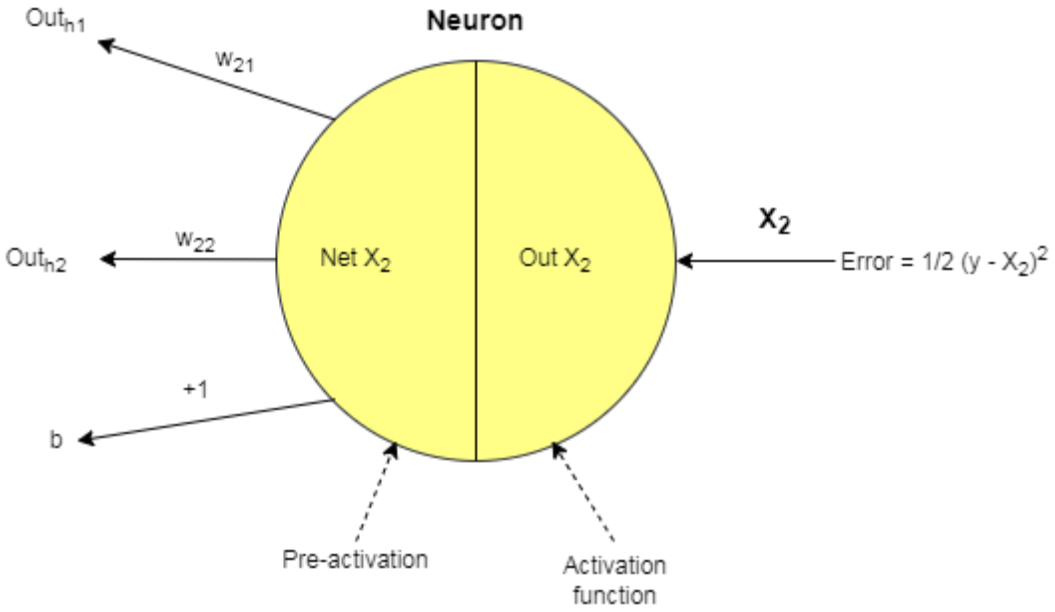

Figure 3.8: An example of propagating the gradient of an error through a one-neuron network. W and b are the weights and bias, Net X is the pre-activation function, Out X is the neuron's output, and Out h is the input.

respect to weight W_{21} is computed as follows:

$$\frac{\partial L_{\text{total}}}{\partial W_{21}} = \frac{\partial L_{\text{total}}}{\partial X_2} * \frac{\partial X_2}{\partial NetX_2} * \frac{\partial NetX_2}{\partial W_{21}} \tag{3.18}$$

Where X_2 is the neuron's output, NetX_2 is the pre-activation function. We have $X_2 = OutX_2$ the predicted output, therefore:

$$NetX_2 = W_{21} \times Out_{h1} + W_{22} \times Out_{h2} + b \tag{3.19}$$

$$\text{Out } X_2 = \text{ Sigmoid } (NetX_2) = \frac{1}{1 + e^{-NetX_2}} \tag{3.20}$$

Backpropagation algorithms use gradient-based algorithms to adjust the parameters (weights and biases), and it works by changing the parameters in the opposite direction of the gradient at the learning rate. Using the example in figure 3.8, and the gradient descent algorithm, the new weight W_{21}^{+} is computed as follows:

$$\begin{gathered} W_{21}^{+} = W_{21} - \eta \left(\frac{\partial L_{\text{total}}}{\partial W_{21}} \right) \\ W_{21}^{+} = W_{21} - \eta \left(\frac{\partial \left(\frac{1}{2} (y_i - p_i)^2 \right)}{\partial W_{21}} \right), p_i = X_2 \\ W_{21}^{+} = W_{21} - \eta \left(\frac{\partial \left(\frac{1}{2} (y - X_2)^2 \right)}{\partial W_{21}} \right) \end{gathered} \tag{3.21}$$

Where $\frac{\partial L_{\text{total}}}{\partial W_{21}}$ is the partial derivative of the total error L_{total} with respect to weight W_{21}, η is the learning rate, X_2 is the output of the neuron, y is the ground truth.

The backward algorithm's main operation is to compute the gradient of the objective function with respect to some parameters, which is obtained using the chain rule technique. Furthermore, the backward algorithm can estimate the next best deep neural network parameters based on the previous parameters.

3 Multiple Instance Learning

Hardware and software advancements have made it possible to efficiently parallelize calculations and train a machine learning model with little effort. When it comes to medical imaging, on the other hand, there is a limited amount of images accessible for training, yet a single image can contain billions of pixels. A single label for a single image is also usually offered, which limits the number of possible variations. It follows as a logical consequence that the topic of how to handle such huge images and learn from poorly labeled training data arises naturally. The search for local patterns and the combination of these patterns into a global choice are two possible solutions. The multiple instance learning method [21] is an alternative to traditional supervised learning in which one label corresponds to a single image. In contrast to traditional supervised learning, where one label corresponds to a collection (a bag) of multiple images (instances), we consider a situation in which one label corresponds to a collection (a bag) of multiple images (instances) [21]. Using a technique similar to that used for processing a minibatch, we can handle a huge image by processing all smaller instances simultaneously.

3.1 Methodology

Dietterich et al. [21] first proposed the multiple instance learning framework, which addressed the problem of predicting molecule drug activity. To perform their functions, most drugs are small molecules that bind to much larger molecules such as enzymes and cell surface receptors. By rotating its bonds, each drug molecule can take on a variety of shapes known as conformations. A drug molecule is labeled "active" if at least one of its conformations can bind to a binding site. In any of its possible conformations, a "inactive" molecule cannot bind to a binding site. In this context, an instance is a single conformation of a molecule, and a bag is all conformations of a specific molecule.

In binary supervised learning, the goal is to find a mapping from an instance $x \in R^D$ to a label $y \in \{0, 1\}$; in MIL, the goal is to find a mapping from a bag of instances $X = x_1, ..., x_K$ to a label $Y \in \{0, 1\}$. It's worthy to note that the number of instances K in a bag isn't always

the same for all bags in X. In MIL, we assume that instances in a bag are unsorted and unrelated to one another. Furthermore, we assume that each instance in a bag has a binary label, i.e., $y_1, ..., y_K, y_k \in \{0, 1\}$ for all $k = 1, ..., K$ even though we do not have access to these instance labels during training [21]. We can now define MIL's main assumption as follows:

$$Y = \begin{cases} 0, & \text{if and only if } \sum_{k=1}^{K} y_k = 0 \\ 1, & \text{otherwise.} \end{cases} \tag{3.22}$$

We use the Bernoulli distribution to model the probability of Y given the bag of instances X because our label Y is a binary random variable:

$$p(Y \mid X) = S(X)^Y (1 - S(X))^{(1-Y)} \tag{3.23}$$

in which $S(X) = p(Y = 1 \mid X)$ denotes the scoring function of a bag X.

3.2 Multiple instance learning approaches

Three approaches dominate the MIL literature: instance-based approaches, embedding-based approaches, and bag-based approaches[41]. We will go over each of them in detail in the sections that follow. We'll show later that there are models that aren't necessarily limited to one of the three approaches.

3.2.1 Instance-based approach

In the instance-based approach, we attempt to infer instance scores directly. As a result, for each instance, we train a deep neural network that is shared across instances to compute a score (a scalar value between 0 and 1).In a subsequent step, a MIL pooling layer combines each instance's score and computes a label for the entire bag of instances [41]. Figure 3.9 depicts a potential architecture for an end-to-end trainable deep neural network using the instance-based approach.

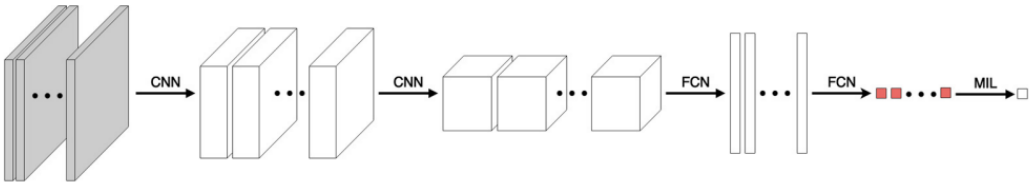

Figure 3.9: Instance-level approach: An instance score is calculated for each instance in a bag using a combination of convolutional and fully connected layers. Finally, a MIL pooling layer is used to deduce the bag label.

The advantage of the instance-based approach is its ability to highlight key instances. Because a practitioner can investigate the highlighted instances, i.e., instances with a high instance score, the approach is highly interpretable. Multiple studies have shown that, when compared to the embedding-based and bag-based approaches, the instance-based approach produces poorer classification performance [83]. Because the instance labels are unknown during training, the deep neural network predicting instance scores may be undertrained, introducing an additional error to the bag label prediction.

3.2.2 Embedding-based approach

The building blocks of the embedding-based approach are the same as those of the instance-based approach. The primary distinction between the two approaches is in the ordering of fully connected layers for classification and the MIL pooling layer. The main goal of the embedding-based approach is to find a compact embedding (latent representation) of a bag. In the following step, we combine the instance embeddings to form a single embedding that represents the entire bag [41]. A MIL pooling layer, similar to the instance-based approach, is used to combine the instance embeddings. In this case, however, the MIL pooling layer must be able to handle a vector input rather than a scalar value. We guarantee that all bag embeddings share the same latent space by using the same deep neural network. Figure 3.10 depicts a possible architecture for an end-to-end trainable deep neural network using an embedding-based approach. MIL models using the embedding-based approach have been shown to outperform instance-based approaches in bag classification [83]. When using the embedding-based approach, however, there is no way to infer instance scores. This makes this approach inapplicable in a wide range of situations where interpretability is critical[83].

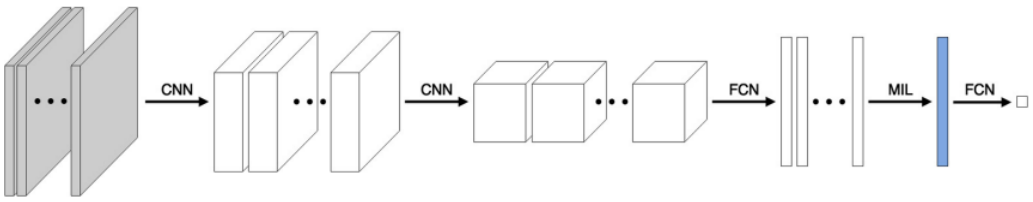

Figure 3.10: Embedded-level approach: Each instance in a bag is first embedded into a low-dimensional space using convolutional and fully connected layers. Second, the instance embeddings are combined into a single bag embedding using a MIL pooling layer. Finally, the bag label is inferred using a series of fully connected layers.

3.2.3 Bag-based approach

The goal of bag-based approaches is to find dissimilarities between bags. They recast a MIL task as a regular supervised problem using metrics such as bag distances, bag kernels, and bag dissimilarities. The bag is treated as a whole in this case, and the implicit assumption is that bag labels can be related to bag distances. The most complex aspect of this method is determining a suitable definition of distance or similarity. A distance or similarity measure is usually only suitable for one task and must be chosen a priori; that is, the measure must be fixed during training. There is no research paper that we are aware of that combines the bag-based approach with deep learning [41].

3.3 Multiple instance learning pooling functions

Finding adequate MIL pooling functions is one of the most difficult difficulties in the MIL problem in general. In the context of artificial deep neural networks, a pooling function is utilized within a pooling layer to get the predicted bag label. Depending on the technique, the pooling layer is in charge of either combing instance scores (instance-based approach) or instance embeddings (instance-based approach), or both (both approaches) (embedded based approach).

Now we'll look at some of the most commonly used MIL pooling functions. Let us consider a bag of instances:

$$X = \{x_1, ..., x_K\} \tag{3.24}$$

Where $x_k \in R^D$ is an image that is represented by its raw pixel values, for example. Using the function f_θ we get a bag of embeddings by parameterizing it with a deep neural network that is shared across instances:

$$H = \{h_1, ..., h_K\}, \text{ where } h_k = f_\theta(x_k) \tag{3.25}$$

The embedding of an instance is either a scalar in the case of the instance-based approach, $h_k \in [0, 1]$, or a vector in the case of the embedding-based approach, $h_k \in R^M$, where $M < D$. We use a pooling function to combine the instance embeddings to a bag embedding $\mathbf{z}$ after we have obtained the embeddings for all instances in a bag. We should note that $\mathbf{z}$ has the same dimensionality as each of the instance embeddings, i.e., dimension 1 for instance-based approaches and dimension $\mathbf{M}$ for embedding-based approaches. The goal of bag embedding is to use a low-dimensional representation to capture the most important information about a bag. A bag embedding, unlike a bag of instance embeddings, can be directly mapped to the corresponding bag label. The following section focuses on the most commonly used MIL pooling functions. There are many other, more specialized

MIL pooling functions not included in our list.

3.3.1 Max

alongside the mean function, it is one of the most commonly used MIL pooling functions. As a result, we can derive the bag label z as follows from a set of instance labels $h \in R^K$:

$$z = \max_{k=1.....K} \{h_k\} \tag{3.26}$$

3.3.2 Mean

is also frequently used in MIL problems and is calculated as follows:

$$z = \frac{1}{K} \sum_{k=1}^{K} h_k \tag{3.27}$$

3.3.3 Noisy-or

is a function inspired by the logic OR gate, and it's defined as follows [53]:

$$z = 1 - \prod_{k=1}^{K} (1 - h_k) \tag{3.28}$$

3.3.4 Generalized mean

is an expansion of the regular mean function in which a scalar r > 0 was introduced, as seen below[60]:

$$z = \left(\frac{1}{|K|} \sum_{k} h_k^r \right)^{\frac{1}{r}} \tag{3.29}$$

3.3.5 The integrated segmentation and recognition (ISR)

it was created by Keeler et al. [44] and is designed to recognize and segment hand-printed numerals, the function is defined as follows:

$$z = \sum_{k} \frac{h_k}{1 - h_k} / \left(1 + \sum_{k} \frac{h_k}{1 - h_k} \right) \tag{3.30}$$

3.3.6 The log-sum-exp

is a function that combines the logarithm, the sum, and the exponential functions. 'r' is a positive constant that differs from zero in this function's hyperparameter. The function is defined as follows[60]:

$$z = r \log \left[\frac{1}{K} \sum_{k-1}^{K} \exp\left(r h_k\right) \right] \tag{3.31}$$

3.3.7 The noisy-and

Kraus et al. [47] first presented this function for classifying and segmenting microscope images, which replicates the behavior of the logic AND function. It uses the Sigmoid as an activation function with two parameters: $'a'$, a fixed constant that regulates the slope of the function, and $'b'$, a trainable parameter that provides an adaptive soft threshold for each class. The functions are as follows:

$$z = \frac{\sigma(a(H-b)) - \sigma(-ab)}{\sigma(a(1-b)) - \sigma(-ab)} \tag{3.32}$$

4 Conclusion

In this chapter, we presented Convolutional Neural Networks (CNN) and Multiple Instance Learning (MIL). Wherein addition to the fundamental building blocks of the CNNs algorithm (Convolution), we explained the operations behind this algorithm, such as non-linear activation function, pooling, and fully-connected layers alongside the forward and backward algorithms which are the backbone of the CNN. On the other side, we detailed the multiple instance learning approach where we explained the overall of the method and the different MIL approaches and MIL pooling functions.

The CNN is known for its ability to extract the features from the complex input data like images, where it's used mainly in images classification and object detection. The CNN is sometimes weak in classifying very complex images like the microscopic or histopathological images where these images aren't object-centered images but have many complex micro-objects, which makes CNN frustrated in classification. Many researchers add different techniques to enhance the CNN in classifying images, e.g., the histopathology images by adding the the color normalization, stain normalization [30], or other layers [19, 77].

Unlike the literature works, we propose a new idea and strategy to enhance the CNN inspired by the James Surowiecki's philosophy of crowds wisdom, which states that the aggregation of information in groups is often better than any single member of the group could have made. In the next chapter, we present our novel contribution which combines

both the CNN and MIL approach, where we added a new layer based on the MIL to the a CNN architecture that helped in achieving good results in the binary classification of these microscopic images.

Chapter 4

Classification of breast cancer malignancy

1 Introduction

Breast cancer is a sickness in which breast cells proliferate uncontrollably. The ducts (tubes that convey milk to the nipple), lobules (glands that produce milk), and connective tissue (fibrous and fatty tissue that binds everything together) are all various types of breast cells where cancer might arise [46]. According to [13], 2.09 million cases of breast cancer were diagnosed in 2018, with 627 thousand fatalities, accounting for 11.6 percent of all cancer deaths. The first step in detecting breast cancer is to use mammography to detect any suspicious tumors or lumps in the breast. Pathologists use mammography images to define the tumor's location, then perform a biopsy to obtain histology images to determine the tumor's malignancy, according to [79]. Histological scans can be used by doctors to identify abnormal cells and evaluate how quickly they grow [36]. Indeed, breast cancer treatment processes are based on histological interpretation, which includes a description of the tumoral tissue, which has a lot of diversity and complex texture that can be seen at different optical magnifications of histopathological images.

The high variability and complex textures of histological images make it difficult to build a Computer-Aided System (CAD) for breast cancer classifications using these images. The textures of the histological images differ depending on the growth of the tumor and the types of cells from where the biopsy is taken [81] as we can see in figure 1.2. In addition, as shown in figure 1.3, histological images have varying magnifications depending on the microscope zoom level from which they were acquired, making computer classification problematic. In this work, we define a new strategy for assessing breast cancer malignancy based on a poorly supervised method known as MIL applied to a convolutional neural network (ResNet50). This technique, known as weakly supervised learning [7] is described by instances (inputs) gathered together into bags (sets), and each bag being labeled as a whole [7]. So, if at least one instance of the bag is positive, the bag is normally presumed to be positive. If all occurrences of the bag are negative, the bag is typically assumed to be negative [53].

This work's main contribution is divided into two categories: (i) we propose a new layer called MILC (Multiple Instance Learning Classifier), in which Multi-Instance layers (blocks) are added to the output layer, where each feature map is mapped to the activation function, and each feature map is given an output. Then, using a generalized mean pooling function, all of these outputs will be combined. (ii): we merged the input images from all magnification factors for each class (benign, malignant) so that our suggested model can detect image malignancy regardless of magnification factor. We also offer an evaluation of our model based on the BreakHis [74] dataset for each magnification factor (X40, X100, X200, X400) to compare it with the state-of-the-art works. This chapter is organized

as follows, related works are discussed in section 2. The methods are more detailed in section 3. Section 4 elaborates on the experiments and results in detail; also, the discussion and comparisons with the recently published methods are described. Finally, section 5 concludes this chapter.

2 Related work

Researchers have begun to use artificial intelligence in breast cancer classifications as a result of the digital imaging revolution in pathology. Many notable studies on malignancy classifications using machine learning and deep learning algorithms have been published. We present previous works that have been published in recent years in this section for both conventional methods and deep learning methods.

2.1 Conventional methods

By conventional methods, we mean the methods where we manually select features where researchers used handcrafted feature extraction methods before the advent of automatic feature extraction algorithms and deep learning algorithms. Different feature descriptors were used in Spanhol et al. [74], including Local Phase Quantization, Local or completed Binary Patterns, Parameter-Free Threshold Adjacency Statistics, Gray Level Co-Occurrence Matrices, and Oriented FAST and Rotated BRIEF (ORB). Quadratic Linear Analysis (QDA), 1-Nearest Neighbor (1-NN), Random Forests of decision trees, and Support Vector Machines (SVM) are the four classifiers associated with these descriptors. Breast cancer malignancy classification, as well as benign and malignant subclasses classifications, have been proposed using a dataset of histological images of breast cancer (BreakHis).

The handcrafted descriptors were also used as feature extraction layers by Samah et al. [67]. The representations were then fed into a K-NN classifier based on the BreakHis dataset. Dora et al. [23] proposed a novel Gauss-Newton approach for breast cancer classification (benign/malignant) that was tested on two datasets: the Wisconsin Diagnosis Breast Cancer (WDBC) database and the Wisconsin Breast Cancer Database (WBCD)[76], with a 100% accuracy reported on the limited WBCD dataset.

In addition, Gupta et al. [35] trained various classifiers (support vector machines, nearest neighbors, decision trees, discriminant analysis, ensemble classifiers) based on color-texture features for breast cancer classification, with the overall accuracy determined by a voting method. To compare the two methods, the authors trained a cross-magnification (combined magnifications) model and a magnification-specific model, with the cross-

magnification model outperforming the magnification-specific model.

2.2 Deep learning methods

Following the change in computer hardware and the advent of deep learning technologies, researchers began applying deep learning algorithms to breast cancer classifications, which became increasingly popular. Cruz et al. [15] employed a convolutional neural network to detect invasive ductal carcinoma in a breast cancer patient. To produce non-overlapping image patches from the 162 entire slide images, the researchers grid sampled them and then entered the patches into their CNN model (see Figure. 4.1). With the assistance of skilled pathologists, these patches were divided into four categories: positive slides, negative slides, background slides, and fatty tissue slides, with the latter two being removed. Additionally, the researchers utilized a random forest classifier on their dataset in order to compare alternative handcrafted descriptors with the deep learning method, which generated the best results with an accuracy of 84.23 % as measured by the balanced accuracy. Abd el Zaher et al. [1] developed a Computer-Aided Diagnosis (CAD) method for the detection of breast cancer that was based on the well-known Wisconsin Breast Cancer Dataset (WBCD). The researchers utilized a deep belief network and a back-propagation neural network on 690 samples with varying train validation fractions, and they reported an accuracy of 99.68%.

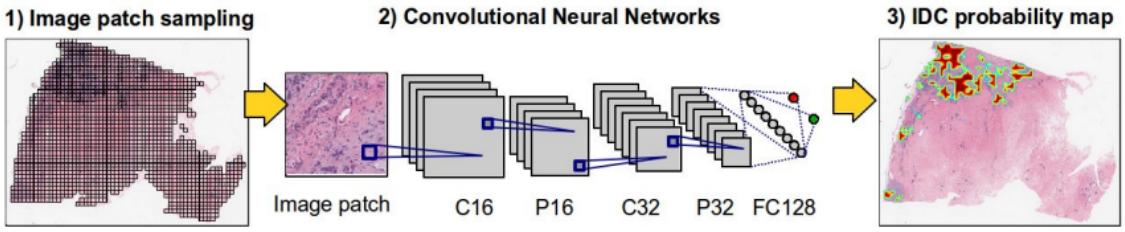

Figure 4.1: Overall framework for automated detection of IDC in WSI using CNN. Image source:[15]

Araujo et al. [6] proposed two different models based on convolutional neural networks, the image-wise model and the patch-wise model, where the first is the result of different voting techniques performed on the outputs of the patch-wise model, and where the second is the result of different voting techniques performed on the outputs of the patch-wise model. There are four categories of output from the models, each of which was trained using 249 histology images from the Bioimaging 2015 challenge [59]. Normal tissue, benign lesion, in situ carcinoma, and invasive carcinoma were the categories used to train the models. Khan et al. [45] proposed a novel deep learning framework for breast cancer detection on 8000 cytology images using transfer learning. The researchers

combined and fed three pre-trained models (ResNet, VGGNet, and GoogLeNet) into a fully connected layer using average pooling, resulting in a highly accurate breast cancer detection system (see Figure 4.2). When it came to both training time and accuracy, the authors demonstrated that their approach outscored CNNs that were taught from scratch.

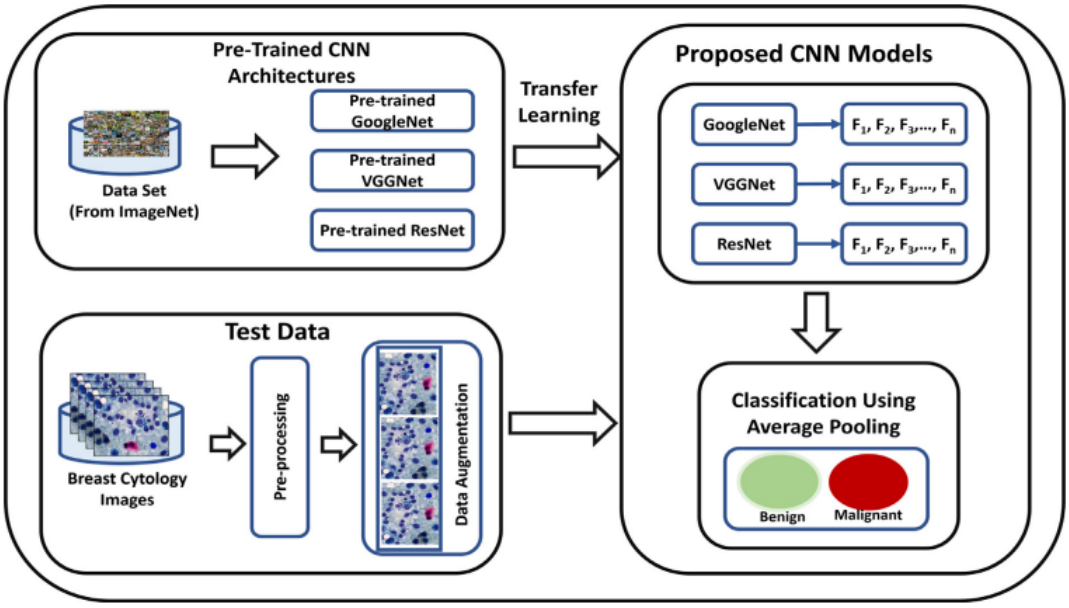

Figure 4.2: Block Diagram of the Proposed Deep Learning Framework in [45]

Following the publication of the BreakHis [74] dataset, a slew of studies have been conducted, including Spanhol et al. [73], who used a Deep Convolutional Activation Feature (DeCAF) for breast cancer detection for each magnification factor separately. The researchers used a pre-trained CaffeNet model trained on ImageNet to show that using DeCAF features, which are repurposed features from another CNN trained on non-histological images, is a good alternative for quickly creating deep learning models for breast cancer classification. The authors of [30] proposed a framework for breast cancer classification into benign and malignant subtypes using MUlti-category classification of breast histopathological image using DEep Residual Networks (MuDeRN) and stain normalization techniques (see Figure. 4.3), the images are classified as benign or malignant and then further classified into subtypes. The framework is trained in two stages on ResNet-152: the first stage classifies image patches into benign and malignant categories, and the second stage classifies benign and malignant images into sub-categories for each magnification factor.

Authors proposed different models on different magnification factors in [8], where they used two different approaches: a machine learning approach based on handcrafted features fed into a support vector machine and a deep learning approach based on convolutional neural networks. The authors of this paper demonstrated that deep learning outperformed

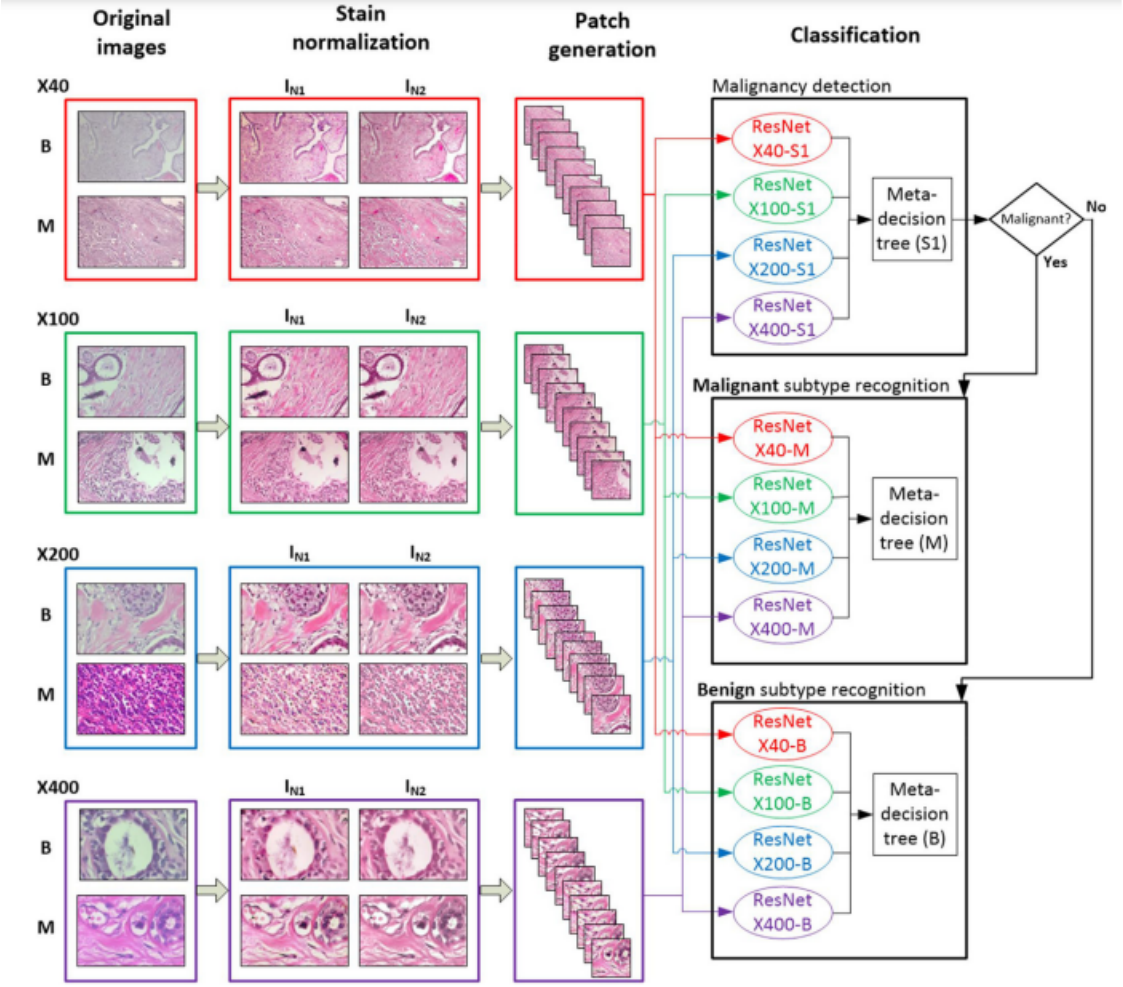

Figure 4.3: The steps of MuDeRN in [30]

machine learning, claiming that CNN models had an accuracy of 96.15 % to 98.33 % for malignant/benign classification. Bayramoglu et al. [9] proposed two CNN architectures for classifying histological images of breast cancer regardless of magnification factor. The first is a multi-task architecture that uses the magnification factor of the images to predict breast cancer. The second architecture consists of a single task: determining whether the images are benign or malignant (see Figure. 4.4).

Das et al. [19] used multiple instance learning techniques based on the BreakHis dataset for breast malignancy classification. The researcher used VGGNet [70] as the CNN architecture and treated the entire slide image (WSI) as a bag. Even though the patches of the WSI contain both malignant and benign regions, this method helped them learn the label of the WSI. In the X40, X100, X200, and X400 magnification factors, the authors reported accuracies of 89.52 %, 89.06 %, 88.84 %, and 87.67 %. Researchers in [77] used multiple instance learning techniques to classify histological images into benign or malignant for each magnification factor based on the BreakHis dataset. The models were

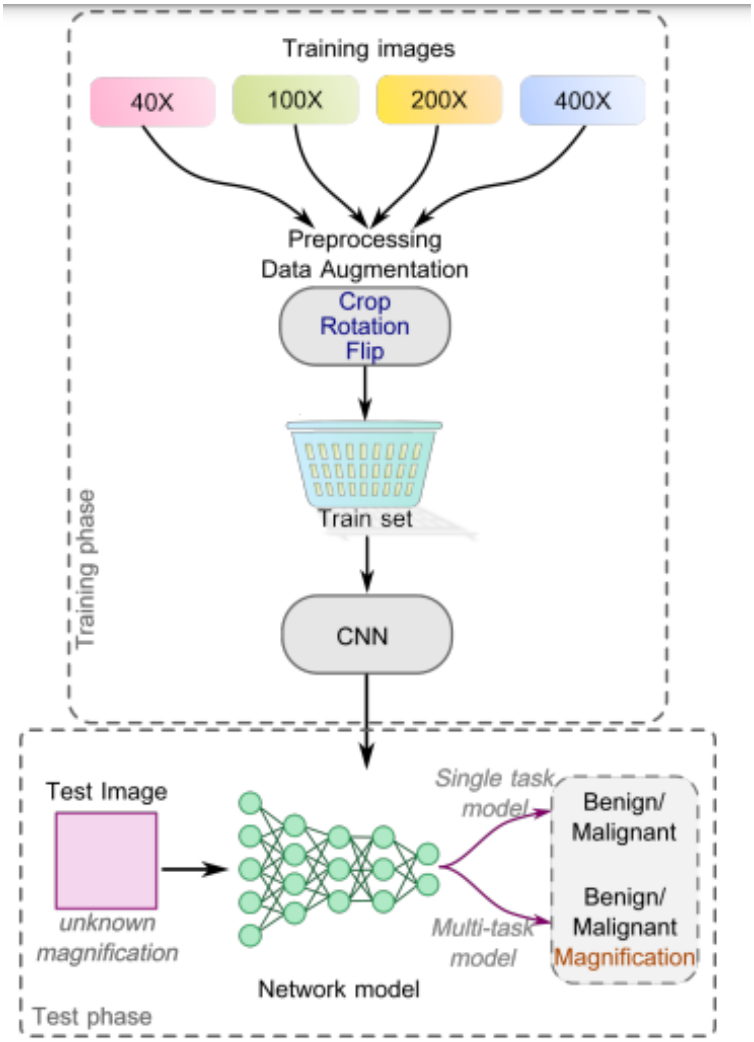

Figure 4.4: Bayramoglu et al. proposed model [9]

trained using two different bag strategies: patient as a bag, which considers all patient images as instances, and image as a bag, which considers each region in an image as an instance (see Figure 4.5). The authors reported best accuracies of 92.1 %, 89.1 %, 87.2 %, and 82.7 % in the X40, X100, X200, and X400 magnification factors, respectively, where they obtained it on the patient using a bag strategy.

In all prior investigations, algorithms that were not the best fit for histopathological and microscopic images were utilized, and methodologies that were developed for categorizing centering objects were applied, which may have resulted in overfitting in microscopic images in some cases, with the exception of the works [77, 19], which employed multiple instances learning techniques but had poor accuracy. This lack encouraged us to build a model for non-centering object images like the histopathological images of breast cancer.

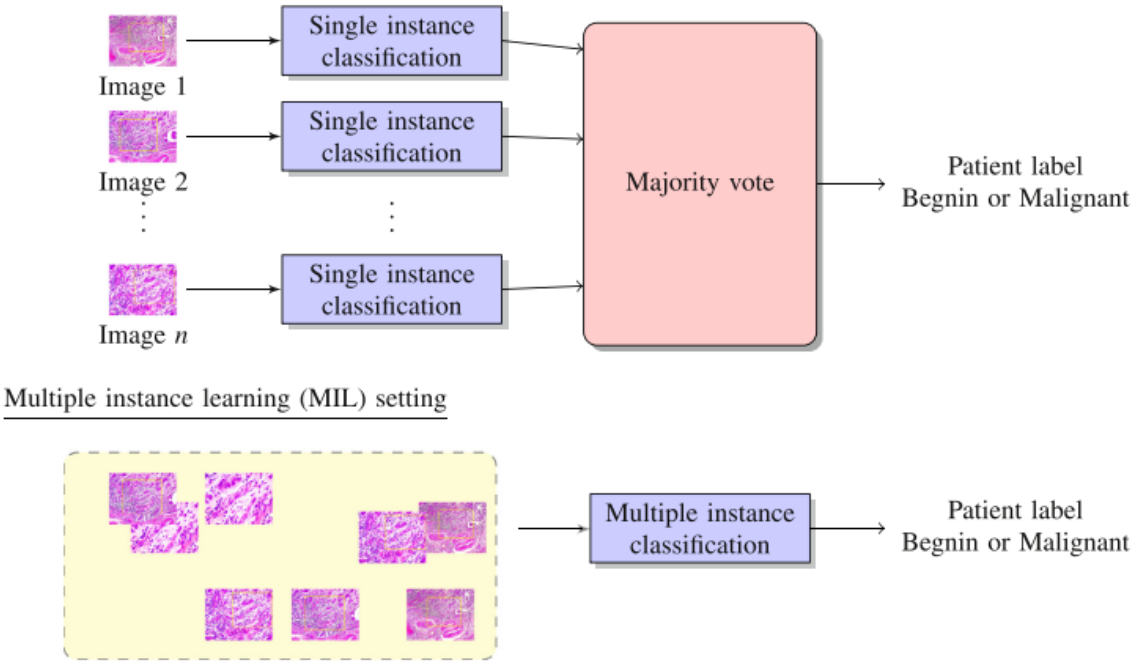

Figure 4.5: Multiple instance learning vs single instance classification proposed models in [77]

3 Residual neural network-based multiple instance learning for breast cancer classification

We discuss our technique for histopathology image categorization in this section, which takes into account the diverse textures of these microscopic images. For classifying breast cancer histopathology images, our fundamental idea is to combine multiple instances learning techniques with convolutional neural networks.

3.1 The proposed architecture

A 1D vector is created out of the feature maps and is then assigned to the output layer, which has an activation function that predicts which class the input picture belongs to in a conventional CNN. Our proposed architecture involves flattening the final feature maps into a 2D matrix and mapping them to the output layer with an activation function (Softmax), which will output different predicted classes for each feature map, and then applying multiple instances learning pooling function to these outputs (instance predicted classes) for aggregation and to get the entire image class (bag predicted classes), as shown in the given figure 4.6.

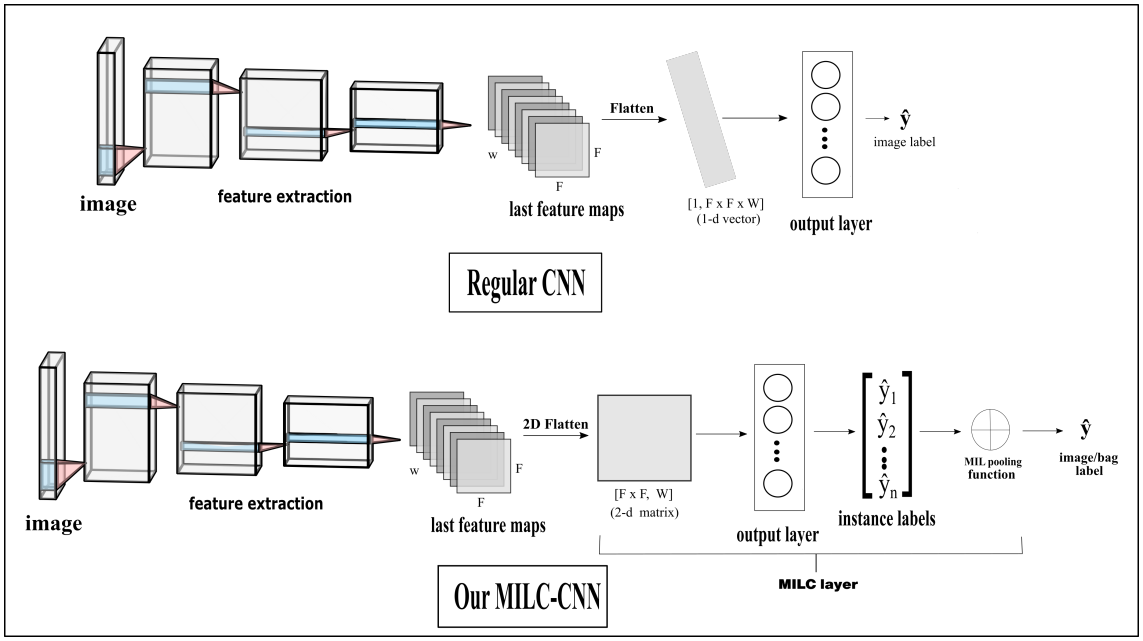

Figure 4.6: The regular CNN and the proposed MILC-CNN

Based on the ResNet50, which is a sort of residual neural network with 50 layers [38], we developed our CNN architecture. Recognized for their shortcuts or skip connections, which allow users to bypass some layers in order to avoid vanishing gradients and degraded accuracy by reusing activations from previous layers in the next layers. ResNet architectures are extremely powerful for classifying complex images such as images with multi-objects. ResNet's design also won the ImageNet competition, where the authors won the first prize in the 2015 ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which was sponsored by Google. For initial convolution and max-pooling, the ResNet50 employs 7x7 and 3x3 kernel sizes, respectively. Following that, the network contains 16 residual blocks, which are built up of (3+4+6+3) blocks with a variety of filter sizes, which are distributed across the network. The convolutional layers in each block are three 1x1, 3x3, and 1x1 convolutions. A flattening layer is also included, which flattens the feature maps into a 2D vector to send them to the activation function (Softmax) to generate instances labels and finally to the MIL pooling function to aggregate instances scores and generate the bag label. The proposed MILC-ResNet50 architecture and methodology is

illustrated in figure 4.7 and in a pseudo algorithm 1.

Weights initialization;
for *j < number of total iterations* **do**
  for *i < number of total training images* **do**
    Feature Extraction;
    2D Flatten;
    Instances' labels generation using Softmax;
    Bag's label generation using MIL pooling function;
    Backpropagation and weights update;
  end
end

Algorithm 1: MILC-ResNet50 pseudo algorithm

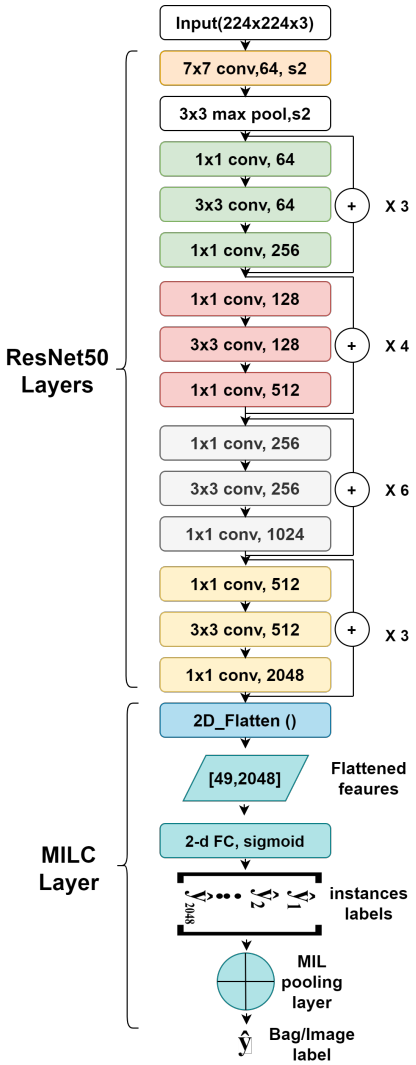

Figure 4.7: Our proposed architecture overview

Our MILC-ResNet50 will generate last feature maps of size 7x7x2048 that will be

flattened to a matrix of size 49x2048, and each features of the 2048 ones - which here represent the instance x_j ($j \in [1, 2048]$) of our model- will be classified using a binary function $p(t_i = 1|x_j)$, where t_i represent the class label, that can be positive or negative ($t_i \in \{0, 1\}$). Then, the instances predictions p_{ij} are combined using a MIL aggregation function $g(.)$, the generalized mean, to get the probability of the final image label (bag label) $p(t_i = 1|x_j, ..., x_{2048})$ (See figure 4.8. So, we can resume these formulations by the following equations:

$$\hat{y} = g(Z) = g(\hat{y}_1, ..., \hat{y}_{2048}) \tag{4.1}$$

Where f represents the Softmax function, X represents the bag, W the output layer weights, Z the predicted labels of the instances, g the MIL pooling function, and $\hat{y}$ represents the predicted bag label.

Our model made use of the Generalized Mean function (see function 3.29)) as the primary MIL pooling function. The Generalized Mean function (GM) is an extension of the mean pooling function, to which a constant 'r' was introduced. The GM function is distinguished by the fact that it allows all examples to contribute equally to the prediction of the bag label using the instances' scoring values. In addition, we compared the results using the various MIL pooling functions (max, ISR, LSE, and Noisy-And), which are explained in detail in section 3.3.

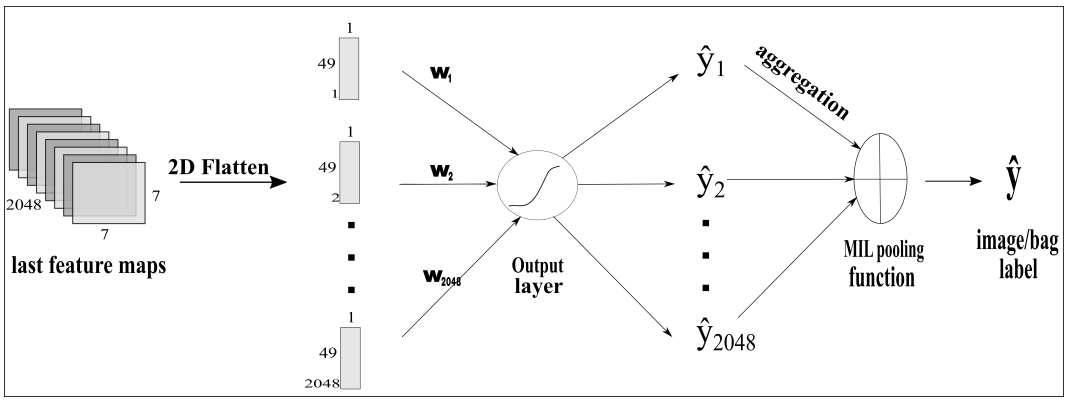

Figure 4.8: Our proposed MILC layer

3.2 The loss function

The traditional CNN loss function evaluates the predicted weights and parameters over each training example like the following equation[33]:

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} f_{\text{loss}}\left(\mathbf{y_i}, F\left(\mathbf{W}, \mathbf{x_i}\right)\right) \tag{4.2}$$

L is the total cost, N is the total number of training examples, $y_i = \{0, 1\}$ is the label of the input (ground truth), W represents the trainable parameters, x_i is the input matrix, F is the activation function (which is the Softmax in our case), $F(\mathbf{W}, x_i)$ is the predicted label of the input.

In our work, we used the loss function as presented in the following equations:

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} f_{\text{loss}}(\mathbf{y_i}, P) \tag{4.3}$$

Where f_{loss} represent the binary cross-entropy loss function that calculated as follows:

$$\begin{aligned} f_{\text{loss}} &= -y_i * \log(P_i) - (1 - y_i) * \log(1 - P_i) \\ &= \begin{cases} -\log(1 - P_i), & \text{if } y_i = 0 \\ -\log(P_i), & \text{if } y_i = 1 \end{cases} \end{aligned} \tag{4.4}$$

And P_i is the predicted label of the bag (image) i, and it's calculated as follows:

$$\begin{aligned} P_i = G(F(\mathbf{W}, x_i)) &= G(\{p_{k=1}, p_{k=2}, ...p_{k=2048}\}) \\ &= \left(\frac{1}{2048} \sum_{k=1}^{2048} p_k^r \right)^{\frac{1}{r}} \end{aligned} \tag{4.5}$$

Where G is the MIL pooling function (generalized mean), and $p_{k=1}, p_{k=2}, ...p_{k=2048}$ are the scores of the instances (predicted labels of the instances) that are calculated using the equation (4.6), and r is a constant of the generalized mean.

$$\begin{aligned} p\left(y = j \mid \theta^i\right) &= \frac{e^{\theta^{(i)}}}{\sum_{j=0}^{k} e^{\theta_k^{(i)}}} \\ where\, \theta^i &= W^T X^i \end{aligned} \tag{4.6}$$

W are the learnable parameters of the output layer, X^i are the different instances (the flattened features), $j \in \{0 or 1\}$, and $k = 2$.

3.3 Dataset description

In this research, we employed the BreakHis [74] dataset, which has emerged as one of the community's gold standard datasets in the last several years. Breast cancer histological images were gathered from 82 patients using an Olympus BX-50 system microscope and a digital color camera to create BreakHis, which is a collection of histopathological

images of breast cancer. The slides were acquired after staining with hematoxylin and eosin (H&E) at four distinct magnifications: 40X, 100X, 200X, and 400X, which corresponded to objective lenses of 4X, 10X, 20X, and 40X, respectively. The BreakHis is divided into two categories: benign and malignant, with eight distinct types of benign and malignant tumors (Malignant: lobular carcinoma, ductal carcinoma, mucinous carcinoma, and papillary carcinoma. Benign: phyllodes tumors, fibroadenoma, tubular adenoma, and adenosis.)[74]. All the images are in the RGB color format with a resolution of 700x460 pixels [74]. In addition, statistics on the dataset are presented in the table 4.1.

Class	Sub-Class	Magnification Factor				Total
		X40	X100	X200	X400	
Benign	Adenosis	114	113	111	106	444
	Fibroadenoma	253	260	264	237	1014
	Tubular Adenoma	109	121	108	115	453
	Phyllodes Tumors	149	150	140	130	569
Malignat	Ductal Carcinoma	864	903	896	788	3451
	Lobular Carcinoma	156	170	163	137	626
	Mucinous Carcinoma	205	222	196	169	792
	Papillary Carcinoma	145	142	135	138	560
Total		1995	2081	2013	1820	**7909**

Table 4.1: BreakHis dataset statistics

3.4 Hardware and software

Python 2.7 is used to implement all of the trained models in this study, which is done on a high-performance computing (HPC) system with an Intel(R) Xeon(R) E5-2660 v3 processor and 64 GB RAM. The tests and training are carried out with the help of Keras[1], Sci-kit Learn[2], and Tensorflow [3]. The datasets are randomly separated into two sections, with 30% of the datasets being used for the test set and 70% of the datasets being used for the training set, with no overlap.

3.5 Data augmentation and transfer learning

One of the hardest and most difficult points of training deep learning models is the lack of ground truth data, particularly medical images, due to the time and effort required to label these data, as well as the patient's privacy, which many hospitals reject patient data

[1]https://github.com/fchollet/keras
[2]https://scikit-learn.org/
[3]https://github.com/tensorflow/tensorflow

requests. To solve this problem, we used online data augmentation rather than offline data augmentation to apply data augmentation to all training set images during the training process, using horizontal and vertical flip, zoom, brightness, and shear transformations. To avoid the overfitting problem, all images in the training and test sets are rescaled to the [0,1] range rather than the [0,255] range. We also used transfer learning to initialize the weights using the ImageNet [20] weights, which is a well-known dataset with over 1.2 million images and 1000 different categories. Without freezing any layers, the pretrained weights are applied to all CNN layers (except the classification layer).

3.6 Data Balancing

As shown in table 4.1, the BreakHis dataset has imbalanced classes, where classes do not have the same number of images or even an approximate number of images. As a result of the disproportionate ratio, directly training deep learning models on this dataset will result in overfitting. To solve this issue, we used the class weighting technique, in which each class is assigned a weight in the loss function based on their class. We calculated the exact weight for each class using the Scikit-Learn Python library for this task.

The class weights are calculated using the follow equation:

$$w_j = \frac{N}{N_c * N_{sc}} \tag{4.7}$$

Where N is the total number of samples, N_c is the number of classes, and N_{sc} is the number of samples per class.

4 Results and discussions

We will present and discuss the results of our various breast cancer classification models in this section. Using our MIL-CNN, we developed a model for classifying images as benign or malignant regardless of their magnification factor. In addition, we used different MIL polling functions to train models and compare the results. To compare the two strategies, we trained models for each magnification factor (X40, X100, X200, and X400). We also compare our results to the state-of-the-art results in this section.

4.1 Evaluation metrics

To evaluate these models we used different metrics, the accuracy, precision, recall, F1-score, and the patient recognition rate.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.8}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.9}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.10}$$

$$F1score = 2 * \frac{precision * recall}{precision + recall} \tag{4.11}$$

Where TP(true positive) represents the situation in which the model properly predicts that a sample belongs to the positive class, when the model properly predicts a sample as belonging to the negative class, this is referred to as TN (true negative). A false positive (FP) is a situation in which the model mistakenly forecasts a sample as belonging to the positive class. When the model mistakenly forecasts a sample as belonging to the negative class, this is referred to as FN (false negative).

The patient recognition rate is calculated as follows:

$$PatientRecognitionRate = \frac{\sum_i^p PatientScore_i}{Totalnumberofpatients} \tag{4.12}$$

where the patient score is defined as:

$$PatientScore_i = \frac{N_r}{N} \tag{4.13}$$

Where N is the number of images of a patient, and N_r is the number of correctly classified images.

4.2 Combined magnification factors models

In order to address the issue of varied magnification factors in histopathological pictures, we developed a model that was trained on merged images from all of the magnification factors in the BreakHis dataset, which has four distinct magnification factors in total (X40, X100, X200, and X400). Regardless of the magnification factor, this model will be able to predict whether the images are cancerous or nonmalignant. To train the model, we employed a variety of MIL pooling functions (max, generalized mean, ISR, LSE, and NoisyAnd function) and compared them to determine the one that was the greatest match

for this purpose. It was decided to initialize the generalized mean (GM) and log-sum-exp (LSE) with (r=5), which is the best value recommended by the researchers who developed these functions [60], and to initialize the NoisyAnd with a=10, which is the best value recommended by the researchers who developed these functions [47].

The results of our MIL-CNN architecture are presented in table 4.2 with the different MIL pooling functions and based on the combined images, where we can clearly see that the generalized mean outperformed the other functions in all metrics. And if we sort the results by accuracy, we'll find that the GM MIL pooling function comes first, followed by NoisyAnd, LSE, Max, and finally, the ISR function. The results are nearly identical, with the exception of the decimal part, where there is a significant difference. According to the results, our models are successful in classifying breast cancer histopathological images regardless of their magnification factor (X40, X100, X200, or X400).

	Precision	Recall	F1 score	Support	Accuracy	PRR
Max	98,59%	98,96%	98,77%	2373	98,31%	98,37%
$GM_{r=5}$	**99,38%**	**99,14%**	**99,26%**	2373	**98,88%**	**99,00%**
ISR	98,94%	97,79%	98,36%	2373	97,83%	97,58%
$LSE_{r=5}$	99,37%	98,59%	98,98%	2373	98,53%	98,46%
$NoisyAnd_{a=10}$	99,25%	98,77%	99,02%	2373	98,63%	98,45%

Table 4.2: Models' results trained on various MIL pooling functions for the combined magnification factors. PRR: Patient Recognition Rate

To make sure our findings were correct, we compared our new MILC-ResNet50 (with a generalized mean function) to ResNet50. Regular ResNet can be used with either the flatten [4] or the global average pooling layer [5]. We used the same hyper-parameters for all three models, which were also trained on combined images with different magnification factors. For training the models, we used a batch size of 16 and a learning rate of 0.0001 for the Adam optimizer. We also used transfer learning, real-time data augmentation, and class weighting. As the table4.3 shows, our MILC layer did better than both other layers in every way. This means that our layer is more efficient and can generalize better. These abilities of the MILC layer come from the separation of the feature maps. Each feature map is directly and separately classified using the activation function. This helped the model predict the class of the image by combining the different labels of the feature maps (instances) that have different kernels. This means that each instance label is predicted using a different shape and edges of the convolution (see Figure: 4.8).

[4] A layer that reshapes the spatial dimensions of the input into a 1-D vector

[5] A layer that reduces the dimensionality of the features maps to 1 dimension

	Precision	Recall	F1 score	Support	Accuracy	PRR
ResNet50 with MILC layer	**99,38%**	**99,14%**	**99,26%**	2373	**98,88%**	**99,00%**
ResNet50 with Flatten layer	94,23%	91,31%	93,77%	2373	91,49%	90,25%
ResNet50 with Global Average Pooling	95,61%	95,03%	95,32%	2373	93,59%	92,02%

Table 4.3: MILC, Flatten, and Global Average Pooling results compared to each other. PRR: Patient Recognition Rate

4.3 Model for each magnification factor

In addition, we trained a model for each magnification factor in order to compare our results to those of other researchers. Table 4.4 shows the results of each magnification factor model. As can be seen, the results of magnification factor X100 produced the best precision, recall, F1 score, and accuracy metrics. This advantage is due to two main factors. The first is the size of the training data, with the X100 model trained on 1455 images compared to 1396, 1409, and 1273 images for X40, X200, and X400, respectively. The second reason is that as we delve deeper into histopathological images, more complex images emerge, making X100 images less complex than X200 and X400 images.

	Precision	**Recall**	**F1 score**	**Support**	**Accuracy**	**PRR**
X40	99,27%	99,76%	99,51%	599	99,50%	**99,71%**
X100	**99,77%**	**99,77%**	**99,77%**	**626**	**99,60%**	99,28%
X200	98,56%	98,80%	98,68%	604	98,68%	97,78%
X400	99,18%	98,65%	98,92%	547	98,72%	98,49%

Table 4.4: Models' results trained on each magnification factor, utilizing generalized mean MIL pooling functions. PRR: Patient Recognition Rate

We can see the comparison between our results and the results of other works that used the same dataset and multiple instances learning approaches in table 4.5. In terms of accuracy and all magnification factors, our results outperformed both Das et al[19] and Sudharshan et al[77], with 99,50 %, 99,60 %, 98,68 %, and 98,72 % for X40, X100, X200, and X400, respectively. For the majority of the magnification factors, the other works had accuracies under 90%.

Work	**Metrics**	**Results**			
		X40	X100	X200	X400
Das et al [19]	Accuracy	89,52%	89,06%	88,84%	87,67%
Sudharshan et al [77]	Accuracy	92,1%	89,1%	87,2%	82,7%
Our Proposed Method	Accuracy	**99,50%**	**99,60%**	**98,68%**	**98,72%**

Table 4.5: Comparative evaluation of the suggested techniques versus existing methods that employed the same dataset and MIL strategy.

We compared our results to the recent results of researches that used the same dataset BreakHis in table 4.6. Our results, as shown in the table, outperformed all of the previous studies in terms of accuracy and patient recognition rate even works based on deep learning [73, 30, 9, 8] in comparison to our work based on multiple instance learning. This means that our proposed architecture is well suited to classifying microscopic and histopathological images without it being an object-centered image. The multiple instances approach was added to the ResNet50 architectures, and each feature (instance) of the last layer of the CNN participated directly in the prediction of the input image class (bag class). Instead of mapping all of the features to the dense layer or using the global pooling function directly, the success key of this method is passing each feature map to the activation function, which will give us different predicted classes for the different features (max-pooling or average pooling).

We also believe that the multiple instances learning improved the results because it follows James Surowiecki's [78] philosophy of crowds wisdom, which states that the aggregation of information in groups is often better than any single member of the group could have made. As a result, having different decisions from different instances of the bag will improve the results more than mapping feature maps directly to the output layer, which will give us directly one final decision.

Work	Metrics	Results			
		X40	X100	X200	X400
Spanhol et al [74]	PRR	83,8%	82,1%	84,2%	82,0%
	Accuracy	Not evaluated	Not evaluated	Not evaluated	85,62%
Spanhol et al [73]	PRR	84,0%	83,9%	86,3%	82,1%
	Accuracy	84,6%	84,8%	84,2%	81,6%
Gandomkar et al [30]	PRR	**98,77%**			
	Accuracy	98,6%	97,9%	98,3%	97,6%
Bayramoglu et al [9]	PRR	79,40%	78,69%	83,72%	80,83%
	Accuracy	Not evaluated	Not evaluated	Not evaluated	Not evaluated
Bardou et al [8]	Accuracy	98,33%	97,12%	97,85%	96,15%
Our Proposed Method	PRR	**99,71%**	**99,28%**	97,78%	98,49%
	Accuracy	**99,50%**	**99,60%**	**98,68%**	**98,72%**

Table 4.6: Comparative evaluation of the suggested techniques versus existing methods that employed the same dataset

5 Conclusion

We proposed a new strategy for classifying breast cancer histopathological images based on Multiple Instances Learning (MIL) and convolutional neural networks in this chapter. After the last feature extraction layer, we used the MIL approach to divide the feature

maps into different instances, which we then mapped to the activation function to get the labels of the instances. The final bag label is created using a MIL pooling function. The following are some of the benefits of our strategy: first, the MIL strategy improved precision by dividing extracted features into different instances to capture deeply the variable and complex micro-objects of the images. Our MILC-ResNet50, on the other hand, is a very powerful architecture for complex histopathological images that avoids the vanishing gradients problem, resulting in higher accuracy. Also, we trained models for each magnification factor of the histopathological images (X40, X100, X200, X400) and a model for combined magnification factors with different MIL pooling functions to demonstrate the efficiency of the generalized mean function in predicting the bag labels. The important part is that our MILC-ResNet50 achieved 99,50%, 99,60%, 98,68%, and 98,72% as accuracy for X40, X100, X200, X400 respectively, which is the best state-of-the-art results in the malignancy breast cancer classification. All the methodologies and the results of this chapter have been published in an international journal of Imaging Systems, and Technology [3].

In the next chapter, we described the problem of grading breast cancer using histopathological images, which is the second step after classifying the malignancy of the tumors.

Chapter 5

Breast cancer grading

1 Introduction

Mammography aids pathologists in locating the tumor, and a biopsy procedure is required to confirm the tumor's malignancy. Doctors must examine the histological images generated by this procedure to determine whether the tumor is malignant or benign. The pathologists define three different grades to allow understanding of how breast cancer cells differ from normal breast cells and how quickly they grow [11]. The cancer cells in grade 1 resemble normal cells in size and uniformity, and they grow slowly. Grade 2 cells are more significant, have more varied shapes, and expand at a faster rate than normal cells. The cells in grade 3 are very large and spread quickly, according to [29].

In this chapter, we illustrate our work in which we used deep learning for grade classification, which has shown good performance in feature extraction and classification compared to traditional methods in recent years (support vector machine, K-nearest neighbor..., etc.). Though since deep learning requires a large amount of data, we focused on the issue of dataset scarcity, which is a common constraint that leads to less generalizability and accuracy in medical image classification models. So, using two different datasets, we propose a new strategy based on multi convolutional neural networks (CNN) to classify breast cancer grades. Our contribution is the addition of a new class called grade zero, which contains benign histological images from the well-known BreakHis [75] dataset combined with the invasive ductal carcinoma (IDC) [22] dataset, which contains the grades classes (grade 1, grade 2, grade 3) of our whole dataset. The ResNet50 model was used because it is the best fit for histological images with large and complex textures. This model uses a skip connection to perform feature extraction, which helps to avoid the vanishing gradients problem [38], which leads to classification overfitting. We also compared the results of ResNet50 architectures with the MobileNet, which has a small number of trainable parameters.

This chapter is organized as follows, related work are discussed in section 2. The methods are more detailed in section 3. Section 4 elaborates on the experiments and results in detail; also, the discussion and comparisons with the recently published methods are described. Finally, section 5 concludes this chapter.

2 Related work

Despite the absence of labeled data, researchers have been working on breast cancer grade categorization for some time. For example, in Petushi et al. [58], they developed a system for identifying and detecting histological characteristics based on 24 H&E stained slide images to assist pathologists in determining the grade of breast cancer based on the

Nottingham score. Segmentation techniques, features extraction methods, and other techniques are used to detect the different cell types, nucleus types, and other cell structures. According to the authors, the density of cell nuclei number and the density of tubular sections determined by image processing are the two characteristics that distinguish between tumor grades.

Doyle et al. [24] employed texture-related features such as grey level features, Haralick features, and Gabor filter features in their work [24]. On the basis of 48 breast biopsy tissue images, they used architectural image features including the Voronoi diagram, Delaunay triangulation, minimum spanning tree, and nuclear features to classify the malignancy and grade (low/high grade) of breast cancer tumors, with the best accuracy on BC grade classification being 93.3%. Dalle et al. [18] used different resolutions of histopathological images (low and high resolution) to assess the grade of breast cancer using segmentation techniques for neoplasm localization and detection, tubular formation detection, and mitotic cell detection and scoring, all based on the Nottingham score, in order to determine the grade of breast cancer. They generated 2000 image frames from a patient's sample, which they then analyzed.

A breast cancer grades dataset consisting of 300 histopathology images, as well as two classification models for malignancy and grade identification, was proposed by Dimitropoulos et al. [22] in 2017. The image representation on the Grassmann manifold is represented by a Vector of Locally Aggregated Descriptors (VLAD) vector. Various patching procedures and image sizes were used to categorize and organize the images. The patch size 8x8 combined with an overlapping method is the most effective strategy, as seen by its 95.8 % accuracy for grade classification. On the basis of 859 histological images, Couture et al. [14] developed a model for breast cancer grades classification. Their methodology divides tumors into just two categories: low-intermediate tumor grade and high tumor grade, which are the most common in breast cancer grading.

Pan and colleagues [57] employed the Xception architecture (See figure 5.1), one of the convolutional neural networks fed by both the Breakhis[74] and the breast cancer grades data sets, to propose a fine-grained multi-tasks model for breast cancer malignancy and grade classification. The BC grade classifier was 94.54% accurate based on 300 histology images. For small and medium-sized biomedical datasets, the Maguolo et al.[52] trained an ensemble of CNNs models using a range of activation functions which was suggested to increase the performance. The researchers analyzed a number of biological datasets, including the breast cancer grades dataset[22], where their algorithms obtained a 95.33% accuracy rate in predicting breast cancer grades.

The above methods used conventional methods, which are weak in features representations as done in [24, 18, 22, 58]. For this, deep learning algorithms have been promised

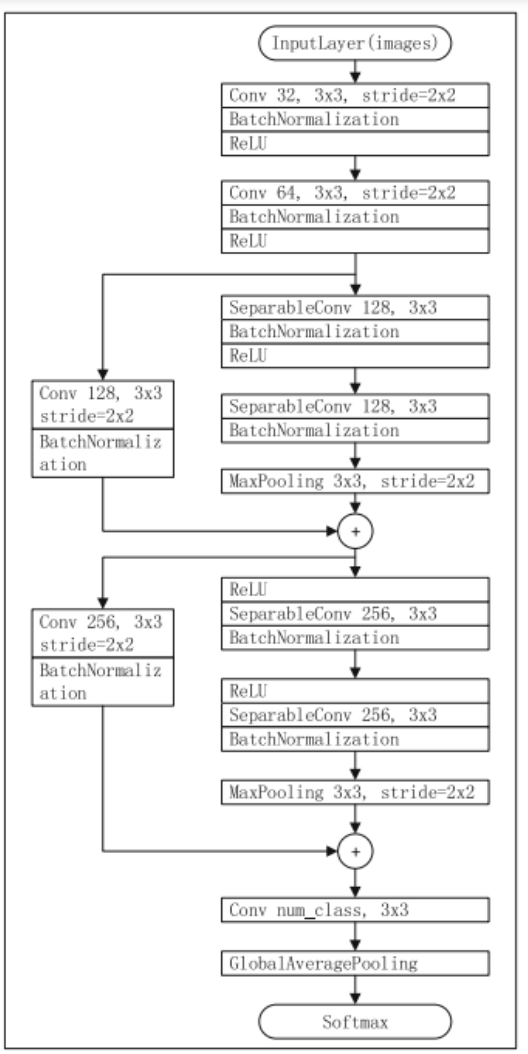

Figure 5.1: The Xception based network architecture in [57]

in computer vision images by using different powerful architectures for complex objects [68, 71, 9, 10]. Thus, the multi-tasks model in [57] achieved the best results compared to the single models due to the feeding of two different datasets to the multi-tasks model[57].

3 The proposed model for breast cancer grading

In this part, we discuss a novel breast cancer grade categorization system that has been developed. To train two different convolutional neural networks, we present a technique that relies on combining data from two different datasets: the grades dataset and the BreakHis[74] dataset. To that end, we created a new class named "Grade 0" in our dataset, which reflects the images that were found to be benign in the BreakHis dataset. All of the

images utilized in this study are from four different classes: adenosis, fibroadenoma, phyllodes tumor, and tubular adenoma. And all of the images used in this study have the same magnification factor as the grades dataset, which is a factor of X400. After we produced a dataset that allowed the CNN models (ResNet50 and MobileNet) to classify the tumor malignancy and grade rather than using two different models that predicts the malignancy of the tumor first and then the grades. In order to do this, we created a class for benign cases because the three prior classes (grades 1, 2, and 3) represent the malignant class (see Figure 5.2).

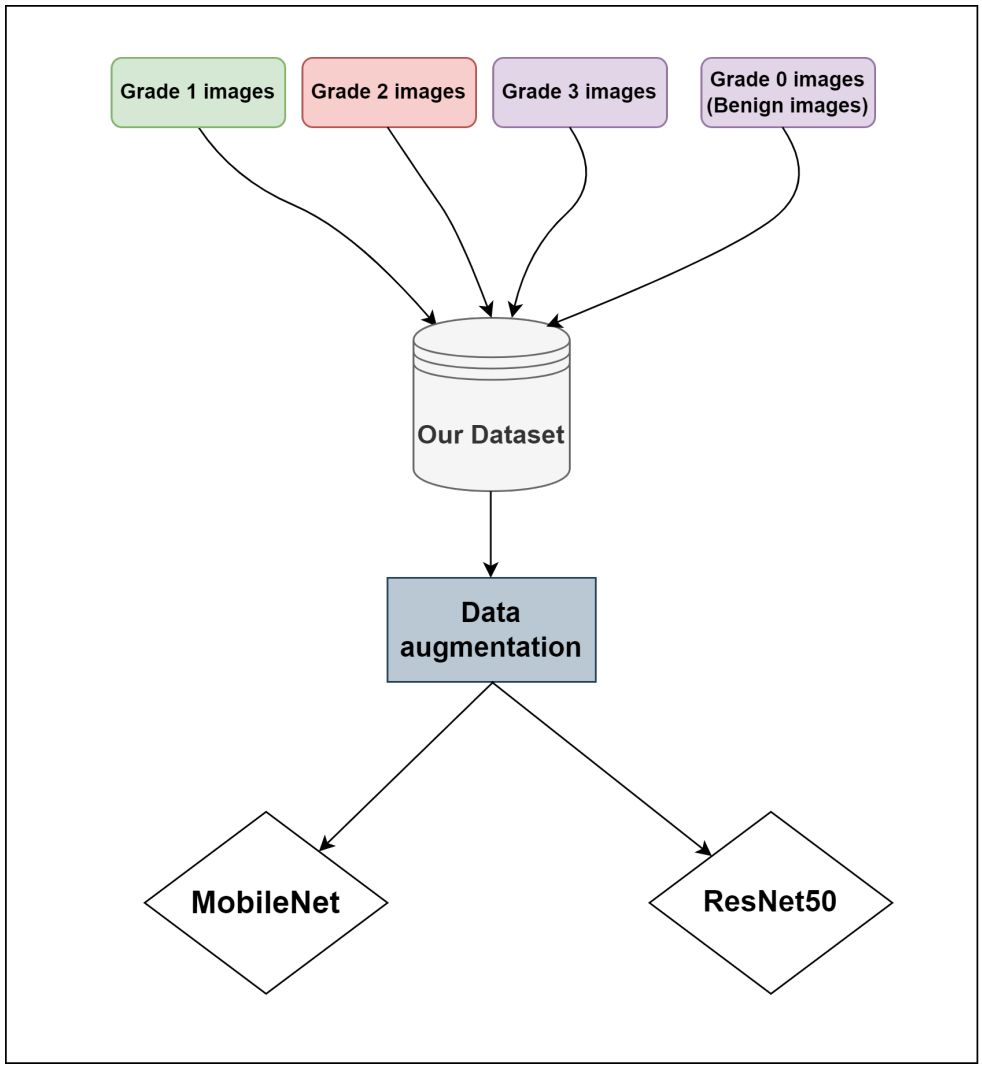

Figure 5.2: Our proposed grading strategy

3.1 ResNet50 Architecture

ResNet, or Residual Neural Networks, is a neural network architecture that performs residual training rather than learning some features. This means that shortcut connections are

used to subtract the learned feature from a layer's input. So, H(x) denotes the underlying mapping function that will be fit by a few stacked layers, with x denoting the first of these layers' inputs. Instead of expecting the normally stacked layers to approximate H(x), these layers explicitly approximate F(x):= H(x) - x. So, F(x)+x replaces the original function. These skip connections (shortcuts) are used to jump over some layers and avoid the problem of vanishing gradients, as well as reusing activation from previous layers and combining low-level and high-level features.

ResNet50, as shown in figure 5.3, is a residual neural network with 50 layers. This architecture uses 7x7 and 3x3 kernel sizes for initial convolution and max-pooling, respectively. Following that, the network has 16 residual blocks made up of (3+4+6+3) blocks with various filter sizes. There are three convolutional layers in each block (1x1, 3x3, and 1x1 convolutions). Finally, there is a global average pooling layer, which is followed by a fully-connected layer with N neurons, where N is the number of classes [38].

A fully connected layer of four neurons (N=4) is added to our ResNet50 model in our four grades strategy. Then, over 100 epochs, we trained our model with 16 batch size and input RGB images resized to 224X224 using Adam optimizer and cross-entropy loss.

3.2 MobileNet architecture

Because there are fewer parameters and multiplications/ additions, the MobileNet described in table 5.1 is a smaller model that is useful for mobiles and embedded systems. These characteristics are the result of using depthwise convolution, which uses a single filter for each input channel rather than combining them to create new features. The new features are then generated using filters and a 1 x 1 convolution (pointwise). Depthwise separable convolution [40] is a combination of depthwise convolution and pointwise convolution.

To make the comparison between the architectures possible and fair, we trained our MobileNet model with the same hyper-parameters used in ResNet50 (Adam optimizer, cross-entropy loss, 16 batch size, 100 epochs, and 224x224 RGB input image).

3.3 Datasets description

In this section, we present the datasets used in our implementation, where we used two different datasets, the BreakHis and the breast cancer grades dataset.

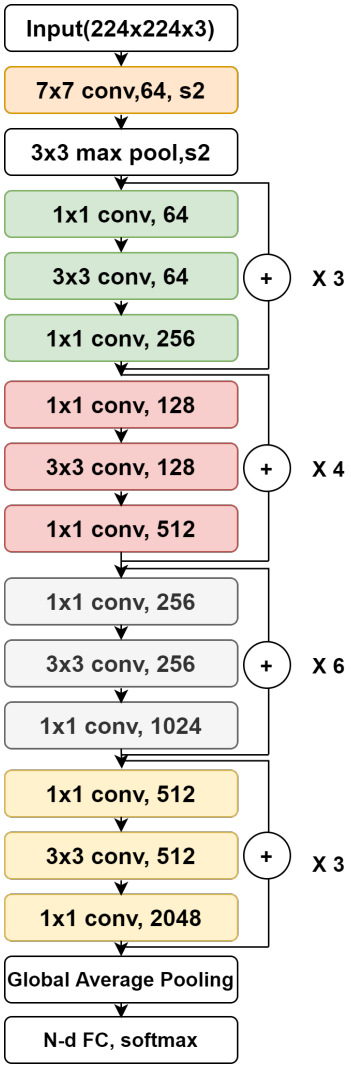

Figure 5.3: The ResNet50 architecture

3.3.1 BreakHis dataset description

BreakHis is a collection of 82 patients' histopathological images from breast cancer. After staining with hematoxylin and eosin (H&E), images were captured using a digital color camera and an Olympus BX-50 system microscope with a relay lens at a magnification of 3.3x. The data was collected with four magnification factors: 40X, 100X, 200X, and 400X, which correspond to objective lenses 4X, 10X, 20X, and 40X, respectively. It also includes eight different types of benign and malignant tumors (Benign: adenosis, fibroadenoma, tubular adenoma, and phyllodes tumors. Malignant: ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma) [75]. And each image has a resolution of 700x460 pixels. The statistics for the entire dataset can be found in table 5.2.

Type / Stride	Filter Shape	Input Size
Conv / s2	3 x 3 x 3 x 32	224 x 224 x 3
Conv dw / s1	3 x 3 x 32 dw	112 x 112 x 32
Conv / s1	1 x 1 x 32 x 64	112 x 112 x 32
Conv dw / s2	3 x 3 x 64 dw	112 x 112 x 64
Conv dw / s1	1 x 1 x 64 x 128	56 x 56 x 64
Conv / s1	3 x 3 x 128 dw	56 x 56 x 128
Conv dw / s2	1 x 1 x 128 x 128	56 x 56 x 128
Conv dw / s2	3 x 3 x 128 dw	56 x 56 x 128
Conv / s1	1 x 1 x 128 x 256	28 x 28 x 128
Conv dw / s1	3 x 3 x 256 dw	28 x 28 x 256
Conv / s1	1 x 1 x 256 x 256	28 x 28 x 256
Conv dw / s2	3 x 3 x 256 dw	28 x 28 x 256
Conv / s1	1 x 1 x 256 x 512	14 x 14 x 256
5x Conv dw / s1	3 x 3 x 512 dw	14 x 14 x 512
Conv / s1	1 x 1 x 512 x 512	14 x 14 x 512
Conv dw / s2	3 x 3 x 512 dw	14 x 14 x 512
Conv / s1	1 x 1 x 512 x 1024	7 x 7 x 512
Conv dw / s2	3 x 3 x 1024 dw	7 x 7 x 1024
Conv / s1	1 x 1 x 1024 x 1024	7 x 7 x 1024
Avg Pool / s1	Pool 7 x 7	7 x 7 x 1024
FC / s1	1024 x 1000	1 x 1 x 1024
Softmax / s1	Classifier	1 x 1 x N

Table 5.1: The fully MobileNet architecture [40]

3.3.2 Breast cancer grades dataset description

This dataset contains histopathological images collected and labelled by pathologists at the General Hospital of Thessaloniki in Greece [22], which were captured using a microscope with an x40 magnification objective lens attached to a Nikon digital camera following the hematoxylin and eosin (H&E) staining phase. The dataset contains 300 annotated images corresponding to 21 different patients in three different grades (grade 1: 107, grade 2: 102, and grade 3: 91 images).

3.4 Data augmentation and transfer learning

One of the most difficult aspects of training deep learning models is the lack of large training data. Furthermore, professional pathologists must label these medical images, which is difficult and time-consuming. To overcome these challenges, we used online data augmentation instead of traditional offline data augmentation to save physical storage space and speed up the training process. During the training of the model, online data

Class	Sub-Class	Magnification Factor				Total
		X40	X100	X200	X400	
Benign	Adenosis	114	113	111	106	444
	Fibroadenoma	253	260	264	237	1014
	Tubular Adenoma	109	121	108	115	453
	Phyllodes Tumors	149	150	140	130	569
Malignat	Ductal Carcinoma	864	903	896	788	3451
	Lobular Carcinoma	156	170	163	137	626
	Mucinous Carcinoma	205	222	196	169	792
	Papillary Carcinoma	145	142	135	138	560
Total		1995	2081	2013	1820	**7909**

Table 5.2: BreakHis dataset statistics

augmentation is used to apply augmentation techniques to a batch of images directly.

We used different data augmentation techniques, which are:

1. The horizontal and vertical flip: is a technique for creating new images by mirroring image pixels across the horizontal and vertical axes.

2. The zoom, where the image is zoomed in by replicating the pixels.

3. Changing the brightness of the image by making it darker or lighter compared to the original image.

4. Shear transformation, which is a type of image rotation where images are rotated using a rotation angle.

In addition, instead of using random initialization, we used transfer learning to train the models, with parameters initialized using ImageNet [20] weights (which include more than 1.2 million images in 1000 different categories). We used the pretrained ImageNet models on the entire CNN architectures in this work, with no layers frozen.

3.5 Data Balancing

As seen in table 5.2, the datasets are unbalanced, with classes not having the same or an approximate number of images, which causes neural networks to be misled and the model to bias towards particular classes. This issue was solved using the class weighting approach, which entails giving a weight to each class on the loss function by multiplying each example's loss by a specified factor that is dependent on the class in which they are found to be. To determine the precise weight of each class, we utilized the Scikit-Learn Python package, which contains integrated functions for calculating the weights.

Each class weight is calculated as follow:

$$w_j = \frac{N}{N_c * N_{sc}} \tag{5.1}$$

Where N is the total number of samples, N_c is the number of classes, and N_{sc} is the number of samples per class.

The regular Softmax loss (Categorical Cross Entropy) is calculated as follows:

$$CE = -log\left(\frac{e^{s_p}}{\sum_j^C e^{s_j}}\right) \tag{5.2}$$

Where C are the classes, S_p is the positive output score, and S_j are the other classes output scores.

On the other hand, the weighted Softmax loss is calculated as follows:

$$WCE = -w_j * log\left(\frac{e^{s_p}}{\sum_j^C e^{s_j}}\right) \tag{5.3}$$

And w_j are the classes weights.

4 Results and discussion

Python 2.7 is used to implement all of the trained models in this study, which is done on a high-performance computing (HPC) system with an Intel(R) Xeon(R) E5-2660 v3 processor and 64 GB RAM. The tests and training are carried out with the help of Keras[1], Sci-kit Learn[2], and Tensorflow [3]. The datasets are randomly separated into two sections, with 30% of the datasets being used for the test set and 70% of the datasets being used for the training set, with no overlap.

[1]https://github.com/fchollet/keras
[2]https://scikit-learn.org/
[3]https://github.com/tensorflow/tensorflow

4.1 Evaluation metrics

Evaluation is performed using the accuracy, precision, recall, and F1-score metrics, which are calculated as follow.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.5}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.6}$$

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \tag{5.7}$$

Where TP(true positive) represents the situation in which the model properly predicts that a sample belongs to the positive class, when the model properly predicts a sample as belonging to the negative class, this is referred to as TN (true negative). A false positive (FP) is a situation in which the model mistakenly forecasts a sample as belonging to the positive class. When the model mistakenly forecasts a sample as belonging to the negative class, this is referred to as FN (false negative).

4.2 The proposed model for grading breast cancer

In this section, we present the results of our proposed four-grade strategy, as well as the three-grade strategy results based on a grades dataset containing only 300 histological images divided into three classes (grades 1, 2, and 3) as defined by the Nottingham et al.[29] scoring system, which is currently the most widely used. The results of the three grades strategy are presented in table 5.3 grades results, where accuracy, precision, recall, F1 score, and support (number of images used for testing) are shown for each CNN architecture (ResNet50, MobileNet).

		Accuracy	Precision	Recall	F1 score	Support
ResNet50	**Model**	**92,39%**	**92,46%**	**92,39%**	**92,29%**	92
	Grade 1	84,85%	93,33%	84,85%	88,89%	33
	Grade 2	100%	91,18%	100%	95,38%	31
	Grade 3	92,86%	92,86%	92,86%	92,86%	28
MobileNet	**Model**	**93,48%**	**93,76%**	**93,48%**	**93,51%**	92
	Grade 1	87,88%	93,55%	87,88%	90,62%	33
	Grade 2	96,77%	100%	96,77%	98,36%	31
	Grade 3	96,42%	87,10%	96,43%	91,53%	28

Table 5.3: Results of CNN models on the 3 grades dataset

The table 5.4 presents the results of the four grades strategy using the same CNN architectures with the same metrics.

		Accuracy	Precision	Recall	F1 score	Support
ResNet50	**Model**	**97,03%**	**97,41%**	**97,03%**	**97,05%**	269
	Grade 0	100%	100%	100%	100%	177
	Grade 1	96,97%	82,05%	96,97%	88,89%	33
	Grade 2	93,55%	96,67%	93,55%	95,08%	31
	Grade 3	82,14%	100%	82,14%	90,20%	28
MobileNet	**Model**	**94,42%**	**94,99%**	**94,42%**	**94,49%**	269
	Grade 0	100%	100%	100%	100%	177
	Grade 1	90,91%	73,17%	90,91%	81,08%	33
	Grade 2	77,42%	88,89%	77,42%	82,76%	31
	Grade 3	82,14%	95,83%	82,14%	88,46%	28

Table 5.4: Results of CNN models on the 4 grades dataset

To begin, table 5.3 shows that MobileNet outperformed ResNet50 in the three-grade classification, with MobileNet receiving 93,48 % and ResNet50 receiving 92,39 % due to MobileNet's effectiveness on limited datasets. The grade 2, on the other hand, had the best accuracy across all CNN architectures, with a score of 100 % in ResNet50 and 96.77 % in MobileNet. Because it has more data than grade 3, and its images are more predictable than grade 1, which are difficult to distinguish, whereas grade 2 took the lead in the results.

Due to the size of the dataset used in this classification, ResNet50 received 97,03% accuracy and outperformed the other CNN architectures in the four-grade classification, as shown in Table 5.4. Furthermore, the grade 0 received 100% accuracy in both the ResNet50 and the MobileNet due to the large number of images used in this class, making our models powerful in the classification of benign breast cancer tumors.

The MobileNet required less training time than the ResNet50, which required 40.15 minutes and 114.57 minutes on the grade dataset and combined datasets, respectively, whereas the ResNet50 required 92.98 minutes and 192.4 minutes on the grade dataset and combined datasets, respectively, for 100 epochs. When it comes to inference time, which is the time it takes to load the model and test it on a single image, the MobileNet performed significantly better than the ResNet50. On the grade dataset and the combined datasets, the MobileNet performed 9.87 seconds and 10.66 seconds, respectively, while the ResNet50 performed 24.97 seconds and 25.89 seconds, respectively. As a result of its light-weighting and the limited number of trainable parameters, the MobileNet beat the ResNet50 both in terms of training time and inference time. Furthermore, the models trained on mixed datasets took longer to train than the models trained on a single data set.

The results of our 4 grades strategy model with and without Data Augmentation and Transfer Learning (DA&TL) are presented in Table 5.5, where the model that was trained with data augmentation and transfer learning outperformed the model that was trained without these two techniques, as shown in the results of table 5.5. Because of the efficacy of transfer learning, these differences in outcomes may be explained by the fact that it

is more effective in initializing the weights of neural networks using ImageNet weights rather than the random initialization method. Additionally, by training the model on a greater variety of data, the data augmentation improves the accuracy of the model due to the generation and adding more training data.

Method	Grade 0	Grade 1	Grade 2	Grade 3	All
ResNet50 (with DA&TL)	100%	96,97%	93,55%	82,14%	**97,03%**
MobileNet (with DA&TL)	100%	90,91%	77,42%	82,14%	**94,42%**
ResNet50 (without DA&TL)	97.77%	57.58%	80.65%	64.29%	87.36%
MobileNet (without DA&TL)	92.09%	39.39%	93.55%	67.86%	83.27%

Table 5.5: Comparison between our 4 grades strategy model with and without Data Augmentation and Transfer Learning (DA&TL) using accuracy metric

In table 5.6 and figure 5.4 we compared our results with the recent works that used the three grades dataset[22]. Our ResNet50 model trained on three grades outperformed the model of [57] in grade 2 classification, in which our model got 100% as accuracy in this class. Also, our ResNet50 model trained on four grades outperformed the model of [57] in grade 1 classification, where it got 96,97% as accuracy, which is very promising results since grade 1 is hardly differentiated. In term of accuracy, our ResNet50 trained on the four grade got the best accuracy 97,03%, better than [57, 22, 52] which got 94.54%, 95.8%, and 95.33% respectively.

Method	Grade 0	Grade 1	Grade 2	Grade 3	All
ResNet50 (3 grades strategy)	-	84,85%	100%	92,86%	92,39%
MobileNet (3 grades strategy)	-	87,88%	96,77%	96,42%	93,48%
ResNet50 (4 grades strategy)	100%	**96,97%**	93,55%	82,14%	97,03%
MobileNet (4 grades strategy)	100%	90,91%	77,42%	82,14%	94,42%
Multi-task Xception [57]	-	89.39%	98.41%	96.30%	94.54%
Dimitropoulos et al. 8x8, overlapping [22]	-	-	-	-	95.8%
Maguolo et al.[52]	-	-	-	-	95.33%

Table 5.6: Comparison of the proposed methods against state of the art approaches based on the invasive breast carcinoma grades dataset

5 Conclusion

In this chapter, we present a set of experiments using different convolutional neural network architectures to classify breast cancer grades using data from two different datasets (ResNet50, MobileNet). In the state-of-the-art, we achieved the best results. Furthermore, experimental results show that the more histological images added to neural nets, the more accurate they are. Our models that were trained on the combined datasets outperformed models that were trained on a single limited dataset of breast cancer grades. All

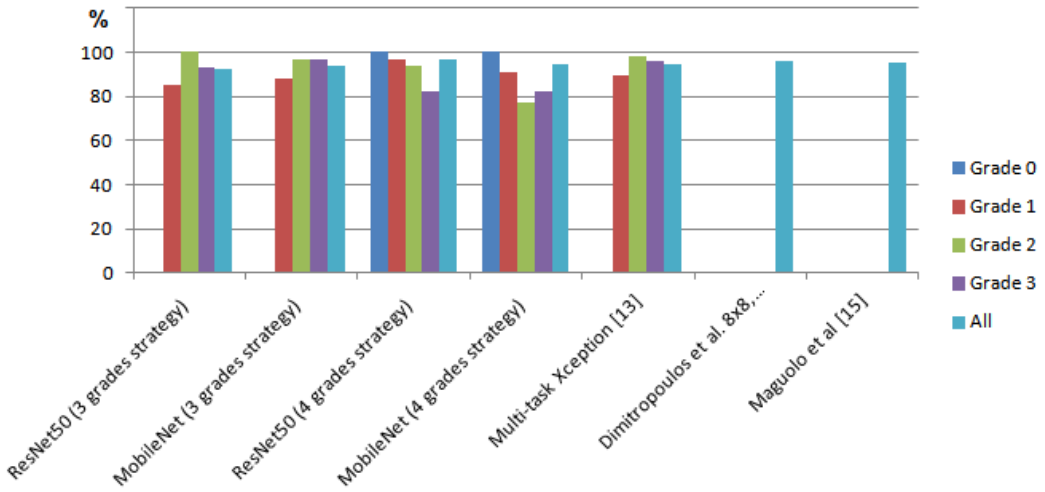

Figure 5.4: A comparison of the suggested approaches with already available techniques

the methodologies and the results of this chapter have been published at an international conference on Image Processing Theory, Tools and Applications (IPTA) [2].

This research will serve as the foundation for our future work, which will include enriching the dataset by labelling more malignant examples at various magnifications with the help of experienced breast cancer pathologists.

Chapter 6

Conclusion

1 Summary of contributions

In this thesis, we studied the challenge of building computer-aided diagnostic systems for classifying breast cancer histopathological images. So, we worked on two different models, the first one is for classifying the histopathological images into benign or malignant, and the second is about classifying the grade of breast cancer, which could be a grade 1, grade 2, or grade 3; according to the most used scoring system of Nottingham[29]. The proposed methods are based on deep learning architectures and, more precisely, on convolutional neural networks and multiple instance learning.

Building a Computer-Aided System (CAD) for breast cancer classifications using histological images is challenging due to the high variability and complex textures of these images, which vary depending on the tumor's growth and the types of cells from which the biopsy is taken. Furthermore, depending on the microscope zoom level from which they are taken, histological images have varying magnifications, making computer classification difficult.

In chapter 2, We provided a comprehensive overview of breast cancer, the anatomy of the female breast, and how breast cancer develops. We also showed the different types of breast cancer (ductal and lobular carcinoma, invasive and in situ, triple-negative breast cancer, and inflammatory breast cancer) as well as the various grades of breast cancer (grade 1, grade 2, and grade 3). Furthermore, we presented the breast cancer screening imaging used by clinicians in the diagnosis of breast cancer in this section, which included mammograms, ultrasound, magnetic resonance imaging, and histopathology images.

Chapter 3 is a brief introduction to deep learning, focusing on convolutional neural networks and multiple instance learning. So, in this chapter, we discussed the various blocks that make up convolutional neural networks, as well as the convolution operation, which is CNN's primary function. This section also covers non-linear activation functions, pooling, and fully-connected layers, as well as the two main CNN algorithms, the forward and backward algorithms. We presented different approaches to the multiple instance learning methods (instance-based approach, embedding-based approach, and bag-based approach) with different multiple instance learning pooling functions (max, mean, noisy-or, generalized mean, integrated segmentation and recognition, log-sum-exp, and the noisy-and).

In chapter 4, We discussed how we combined the multiple instance learning method and convolutional neural networks to create a new strategy for classifying breast cancer histopathological images. First, we demonstrated various works in the field that used both conventional and deep learning methods in this chapter. Second, we showed the loss function we used in our model, as well as the dataset and the techniques (data augmentation, transfer learning, and data balancing) we used to improve the model's robustness. Our

MILC-ResNet50 achieved 99,50%, 99,60%, 98,68%, and 98,72% as accuracy for X40, X100, X200, X400 respectively, which is the best state-of-the-art results in the malignancy breast cancer classification.

In chapter 5, we demonstrate our second contribution to breast cancer classification using histopathological images. The chapter introduces the various related works and methods used to classify cancer grades, then demonstrates the methodology and dataset used, as well as various techniques for improving the results (data augmentation, transfer learning, data balancing). Our ResNet50 model achieved 97,03% as accuracy and outperformed the results of the state-of-the-art.

2 Perspectives

In this thesis, we addressed several issues: breast cancer classification and grading, the lack of data, data balancing, and how to surpass the problem of complex histopathological images using multiple instance learning for binary classification. We intend in the future to work on the multi-classification of breast tumor subtypes which is a challenging task. Also, we'd like to create our own pooling function to use in our MIL layer, which this function must be more generalizable and work for the multi-classification tasks.

On the other hand, we will also work on the breast cancer grading by using the segmentation techniques to help the pathologists in counting and detecting the different cellular objects using the different and latest techniques of deep learning.

3 List of publications

Here is a list of my publications that have been mentioned in this thesis to express my research contributions:

Journal Articles

- **ABDELLI, Adel,** SAOULI, Rachida, DJEMAL, Khalifa, et al. Multiple instance learning for classifying histopathological images of the breast cancer using residual neural network. International Journal of Imaging Systems and Technology, 2022.

Conference Articles

- **ABDELLI, Adel,** SAOULI, Rachida, DJEMAL, Khalifa, et al. Combined datasets for breast cancer grading based on multi-cnn architectures. In : 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE, 2020. p. 1-7.

Bibliography

[1] Ahmed M Abdel-Zaher and Ayman M Eldeib. Breast cancer classification using deep belief networks. *Expert Systems with Applications*, 46:139–144, 2016.

[2] Adel Abdelli, Rachida Saouli, Khalifa Djemal, and Imane Youkana. Combined datasets for breast cancer grading based on multi-cnn architectures. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–7. IEEE, 2020.

[3] Adel Abdelli, Rachida Saouli, Khalifa Djemal, and Imane Youkana. Multiple instance learning for classifying histopathological images of the breast cancer using residual neural network. *International Journal of Imaging Systems and Technology*, 2022.

[4] Key statistics for breast cancer in men. *American Cancer Society*, 2021.

[5] Andreea Anghel, Milos Stanisavljevic, Sonali Andani, Nikolaos Papandreou, Jan Hendrick Rüschoff, Peter Wild, Maria Gabrani, and Haralampos Pozidis. A high-performance system for robust stain normalization of whole-slide images in histopathology. *Frontiers in medicine*, 6:193, 2019.

[6] Teresa Araújo, Guilherme Aresta, Eduardo Castro, José Rouco, Paulo Aguiar, Catarina Eloy, António Polónia, and Aurélio Campilho. Classification of breast cancer histology images using convolutional neural networks. *PloS one*, 12(6), 2017.

[7] Boris Babenko. Multiple instance learning: algorithms and applications. *View Article PubMed/NCBI Google Scholar*, pages 1–19, 2008.

[8] Dalal Bardou, Kun Zhang, and Sayed Mohammad Ahmad. Classification of breast cancer based on histology images using convolutional neural networks. *IEEE Access*, 6:24680–24693, 2018.

[9] Neslihan Bayramoglu, Juho Kannala, and Janne Heikkilä. Deep learning for magnification independent breast cancer histopathology image classification. In *2016 23rd International conference on pattern recognition (ICPR)*, pages 2440–2445. IEEE, 2016.

[10] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

[11] HJG Bloom and WW Richardson. Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years. *British journal of cancer*, 11(3):359, 1957.

[12] William Bloom and Alexander Maximow. *A textbook of histology*. WB Saunders, 1952.

[13] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[14] Heather D Couture, Lindsay A Williams, Joseph Geradts, Sarah J Nyante, Ebonee N Butler, JS Marron, Charles M Perou, Melissa A Troester, and Marc Niethammer. Image analysis with deep learning to predict breast cancer grade, er status, histologic subtype, and intrinsic subtype. *NPJ breast cancer*, 4(1):30, 2018.

[15] Angel Cruz-Roa, Ajay Basavanhally, Fabio González, Hannah Gilmore, Michael Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, and Anant Madabhushi. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *Medical Imaging 2014: Digital Pathology*, volume 9041, page 904103. International Society for Optics and Photonics, 2014.

[16] Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, and Fabio Augusto González Osorio. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection.

In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 403–410. Springer, 2013.

[17] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[18] Jean-Romain Dalle, Wee Kheng Leow, Daniel Racoceanu, Adina Eunice Tutac, and Thomas C Putti. Automatic breast cancer grading of histopathological images. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3052–3055. IEEE, 2008.

[19] Kausik Das, Sailesh Conjeti, Abhijit Guha Roy, Jyotirmoy Chatterjee, and Debdoot Sheet. Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 578–581. IEEE, 2018.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[21] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[22] Kosmas Dimitropoulos, Panagiotis Barmpoutis, Christina Zioga, Athanasios Kamas, Kalliopi Patsiaoura, and Nikos Grammalidis. Grading of invasive breast carcinoma through grassmannian vlad encoding. *PloS one*, 12(9), 2017.

[23] Lingraj Dora, Sanjay Agrawal, Rutuparna Panda, and Ajith Abraham. Optimal breast cancer classification using gauss–newton representation based algorithm. *Expert Systems with Applications*, 85:134–145, 2017.

[24] Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John Tomaszewski. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 496–499. IEEE, 2008.

[25] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[26] Christopher W Elston and Ian O Ellis. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, 1991.

[27] Navid Farahani, Anil V Parwani, Liron Pantanowitz, et al. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathol Lab Med Int*, 7(23-33):4321, 2015.

[28] Kunihiko Fukushima and Sei Miyake. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern recognition*, 15(6):455–469, 1982.

[29] Marcus H Galea, Roger W Blamey, Christopher E Elston, and Ian O Ellis. The nottingham prognostic index in primary breast cancer. *Breast cancer research and treatment*, 22(3):207–219, 1992.

[30] Ziba Gandomkar, Patrick C Brennan, and Claudia Mello-Thoms. Mudern: Multi-category classification of breast histopathological image using deep residual networks. *Artificial intelligence in medicine*, 88:14–24, 2018.

[31] Xianjie Gao, Maolin Shi, Xueguan Song, Chao Zhang, and Hongwei Zhang. Recurrent neural networks for real-time prediction of tbm operating parameters. *Automation in Construction*, 98:225–235, 2019.

[32] X Glorot, A Bordes, and Y Bengio. Deep sparse rectifier neural networks in proceedings of the fourteenth international conference on artificial intelligence and statistics (éd. gordon, g., dunson, d. & dudık, m.) 15 (pmlr, 2011), 315-323, 2011.

[33] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[34] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[35] Vibha Gupta and Arnav Bhavsar. Breast cancer histopathological image classification: is magnification important? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 17–24, 2017.

[36] M.N. Gurcan, L.E. Boucheron, A. Can, A. Madabhushi, N.M. Rajpoot, and B. Yener. Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, 2:147–171, 2009.

[37] MN Gurcan, LE Boucheron, A Can, A Madabhushi, NM Rajpoot, and B Yener. Histopathological image analysis: a review. ieee rev biomed eng. 2009; 2: 147–71, 2009.

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[39] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

[40] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[41] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Deep multiple instance learning for digital histopathology. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, pages 521–546. Elsevier, 2020.

[42] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016.

[43] Andrew Karellas and Srinivasan Vedantham. Breast cancer imaging: a perspective for the next decade. *Medical physics*, 35(11):4878–4897, 2008.

[44] James D Keeler, David E Rumelhart, and Wee Kheng Leow. Integrated segmentation and recognition of hand-printed numerals. In *Advances in neural information processing systems*, pages 557–563, 1991.

[45] SanaUllah Khan, Naveed Islam, Zahoor Jan, Ikram Ud Din, and Joel JP C Rodrigues. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognition Letters*, 125:1–6, 2019.

[46] Daniel B Kopans. *Breast imaging*. Lippincott Williams & Wilkins, 2007.

[47] Oren Z Kraus, Jimmy Lei Ba, and Brendan J Frey. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59, 2016.

[48] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[49] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[50] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[51] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, pages 1107–1110. IEEE, 2009.

[52] Gianluca Maguolo, Loris Nanni, and Stefano Ghidoni. Ensemble of convolutional neural networks trained with different activation functions. *arXiv preprint arXiv:1905.02473*, 2019.

[53] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *Advances in neural information processing systems*, pages 570–576, 1998.

[54] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[55] Nandita Nayak, Hang Chang, Alexander Borowsky, Paul Spellman, and Bahram Parvin. Classification of tumor histopathology via sparse feature learning. In *2013 IEEE 10th international symposium on biomedical imaging*, pages 410–413. IEEE, 2013.

[56] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[57] Xipeng Pan, Lingqiao Li, Huihua Yang, Zhenbing Liu, Yubei He, Zhongming Li, Yongxian Fan, Zhiwei Cao, and Longhao Zhang. Multi-task deep learning for fine-grained classification/grading in breast cancer histopathological images. In *International Symposium on Artificial Intelligence and Robotics*, pages 85–95. Springer, 2018.

[58] Sokol Petushi, Fernando U Garcia, Marian M Haber, Constantine Katsinis, and Aydin Tozeren. Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. *BMC medical imaging*, 6(1):14, 2006.

[59] A. Pêgo and P. Aguiar. Bioimaging 2015, 2015. Available from: http://www.bioimaging2015.ineb.up.pt/ dataset.html.

[60] Jan Ramon and Luc De Raedt. Multi instance neural networks. In *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*, pages 53–60, 2000.

[61] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

[62] Stanley L Robbins and Ramzi S Cotran. *Pathologic basis of disease*. Saunders, 1979.

[63] Fredika M Robertson, Melissa Bondy, Wei Yang, Hideko Yamauchi, Shannon Wiggins, Samira Kamrudin, Savitri Krishnamurthy, Huong Le-Petross, Luc Bidaut, Audrey N Player, et al. Inflammatory breast cancer: the disease, the biology, the treatment. *CA: a cancer journal for clinicians*, 60(6):351–375, 2010.

[64] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[65] Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.

[66] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[67] Afiqah Abu Samah, Mohammad Faizal Ahmad Fauzi, and Sarina Mansor. Classification of benign and malignant tumors in histopathology images. In *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 102–106. IEEE, 2017.

[68] Rachida Saouli, Mohamed Akil, Rostom Kachouri, et al. Fully automatic brain tumor segmentation using end-to-end incremental deep neural networks in mri images. *Computer methods and programs in biomedicine*, 166:39–49, 2018.

[69] Jun Shi, Jinjie Wu, Yan Li, Qi Zhang, and Shihui Ying. Histopathological image classification with color pattern random binary hashing-based pcanet and matrix-form classifier. *IEEE journal of biomedical and health informatics*, 21(5):1327–1337, 2016.

[70] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[71] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206, 2016.

[72] American Cancer Society. *Cancer facts & figures 2015*. American Cancer Society, 2015.

[73] Fabio A Spanhol, Luiz S Oliveira, Paulo R Cavalin, Caroline Petitjean, and Laurent Heutte. Deep features for breast cancer histopathological image classification. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1868–1873. IEEE, 2017.

[74] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2015.

[75] Fabio A. Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, July 2016.

[76] W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pages 861–870. International Society for Optics and Photonics, 1993.

[77] PJ Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, and Paul Honeine. Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications*, 117:103–111, 2019.

[78] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[79] P.J. TADROUS. Digital stain separation for histological images. *Journal of Microscopy*, 240(2):164–172, May 2010.

[80] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, 35(8):1962–1971, 2016.

[81] Mitko Veta, Josien PW Pluim, Paul J Van Diest, and Max A Viergever. Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering*, 61(5):1400–1411, 2014.

[82] Tiep Huu Vu, Hojjat Seyed Mousavi, Vishal Monga, Ganesh Rao, and UK Arvind Rao. Histopathological image classification using discriminative feature-oriented dictionary learning. *IEEE transactions on medical imaging*, 35(3):738–751, 2015.

[83] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018.

[84] Robert A Weinberg. How cancer arises. *Scientific American*, 275(3):62–70, 1996.

[85] Yan Xu, Tao Mo, Qiwei Feng, Peilin Zhong, Maode Lai, I Eric, and Chao Chang. Deep learning of feature representation with multiple instance learning for medical image analysis. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1626–1630. IEEE, 2014.

[86] Niyun Zhou, De Cai, Xiao Han, and Jianhua Yao. Enhanced cycle-consistent generative adversarial network for color normalization of h&e stained images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 694–702. Springer, 2019.