Classification of Breast Cancer Cytological Images using Transfer Learning and Deep Convolutional Neural Networks

Mohammad Amin Shamshiri

A thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Masters (Computer Science)

Concordia University

Montréal, Québec, Canada

July 2022

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Mohammad Amin Shamshiri

Entitled: Classification of Breast Cancer Cytological Images using Transfer Learning and Deep Convolutional Neural Networks

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Charalambos Poullis — Chair

Dr. Sudhir Mudur — Examiner

Dr. Charalambos Poullis — Examiner

Dr. Adam Krzyzak — Thesis Supervisor(s)

_____ Thesis Supervisor(s)

Approved by _____
Dr. Lata Narayanan — Chair of Department or Graduate Program Director

Dr. Mourad Debbabi — Dean of Faculty

Abstract

Classification of Breast Cancer Cytological Images using Transfer Learning and Deep Convolutional Neural Networks

Mohammad Amin Shamshiri

Microscopic analysis of breast cancer images is the primary task in diagnosing cancer malignancy, which requires high expertise and precision. Recent attempts to automate this highly subjective task have employed deep learning models whose success has depended on large volumes of data while acquiring annotated data in biomedical domains is time-consuming and may not always be feasible. A typical strategy to address this is to apply transfer learning using pre-trained models on a large natural image database (e.g., ImageNet) instead of training a model from scratch. This approach, however, has not been effective in several previous studies due to fundamental differences in patterns, size, and data features between natural and medical images. In this study, we propose and compare several transfer learning approaches that, in the pre-training phase, use both unrelated natural images and related histopathological images to our target data (i.e., cytological images) in order to classify breast cancer cytological biopsy specimens. To our best knowledge, this is the first reported effort to employ a histopathology data source in transfer learning to classify cytological images of breast cancer. Despite intrinsic differences between histopathological and cytological images, we demonstrate that the features learned by the deep networks during the pre-training are compatible with those obtained throughout fine-tuning with the target data set. To thoroughly investigate this assertion, we explore three different strategies for training as well as two different approaches for fine-tuning

deep learning models. The proposed method is compared with five state-of-the-art studies previously conducted on the same data set of cytological biopsy images, and we demonstrate that the proposed approach significantly outperforms all of them in terms of classification accuracy. Specifically, the proposed method boasts of improved classification accuracy by 6% to 17% compared to the state-of-the-art studies which were based on traditional machine learning techniques, and also enhanced accuracy by roughly 7% compared to those who utilized deep learning methods, eventually achieving 94.55% test set accuracy and 98.73% validation accuracy. Experimental results show that our approach, despite using a very small number of training images, has achieved performance comparable to that of experienced pathologists and has the potential to be applied in clinical settings.

Acknowledgments

Before all else, in memory of my dear father who never saw the end of this adventure, I would like to dedicate my dissertation to him. The journey of pursuing my master's degree began with his encouragement, and before he left this world, he always kept me hopeful for the future with his endless support.

My sincere thanks goes to my family members, especially my mother. Despite the great distance, I always felt their presence by my side, and this success would never have been possible without their support.

I would also like to thank my supervisor, Dr. Adam Krzyzak, for his technical advice and cooperation during my research at Concordia University.

Finally, I would like to thank all the individuals who directly or indirectly assisted me in this research.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Description
AUC	Area Under the ROC Curve
BreakHis	Breast Cancer Histopathological Database
CAD	Computer-Aided Diagnosis
CCD	Charged-Coupled Device
CNN	Convolutional Neural Network
CT	Computed Tomography
CV	Computer Vision
DL	Deep Learning
DM	Digital Mammography
FC	Fully Connected
FNB	Fine Needle Biopsy
H&E	Hematoxylin & Eosin
HAR	Human Activity Recognition
ICIAR	International Conference on Image Analysis and Recognition
IDC	Invasive Ductal Carcinomas
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
ISBI	International Biomedical Imaging Symposium
LSTM	Long Short Term Memory
ML	Machine Learning
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
PCA	Principal Component Analysis
ReLU	Rectified Linear Activation Unit
RGB	Red Green Blue
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SOA	State-of-the-Art
SVM	Support Vector Machines
TIFF	Tagged Image File Format
TL	Transfer Learning
US	Ultrasound
WHO	World Health Organisation

Chapter 1

Introduction

In this chapter, we first briefly state the subject under study and explain the motivation behind this thesis (Section 1.1). Next, the main objectives pursued in this research are summarized in Section 1.2. The contributors to this study are then listed in Section 1.3. Lastly, we present the organized structure of this thesis (Section 1.4), and the published articles are mentioned in Section 1.5.

1.1 Motivation

Breast cancer is the most common invasive cancer among women and has long been recognized as a major public health burden worldwide (Sung *et al.* (2021) [79]). Curing cancer is strongly dependent on the early detection of the disease and the implementation of appropriate treatment. One reliable way to effectively diagnose breast cancer is to analyze cell samples acquired by the Fine-needle Biopsy (FNB) (Mitra *et al.* (2016) [52]). FNB involves obtaining material directly from a tissue suspected of having cancer. The collected biopsy specimens are examined under a microscope by a pathologist to determine the prevalence of cancer cells. Nevertheless, even for qualified and experienced pathologists, the task is tedious, time-consuming, and susceptible to error. Performing this task automatically can potentially improve

efficiency and reliability, allowing it to be accomplished on a large scale.

Advances in medical imaging techniques, on the one hand, and the development of computer-aided diagnosis systems in recent years, on the other, have enabled the automatic analysis of medical images, helping physicians to expedite performing the task. Despite some successes, it has not been feasible to develop a universal system that could be used in routine diagnostics. Due to the complex nature of the problem (image segmentation, feature extraction, classification, etc), the approaches are mainly based on Machine Learning (ML) and Deep Learning (DL) techniques. Especially, DL has proved to be effective in the segmentation and classification of histopathological and cytological images [14, 15, 35, 86].

Given the capability of deep Convolutional Neural Networks (CNNs) to learn generalizable descriptors directly from images, many previous studies have employed them to perform image analysis tasks. Despite the CNNs' high potential for analyzing pattern recognition problems, their application in biomedicine does not always yield state-of-the-art (SOA) results. The success of CNNs in achieving satisfactory results depends on the availability of large volumes of data, while acquiring annotated data in the biomedical domain is time-consuming and may not always be possible. Furthermore, training deep CNNs from scratch is computationally expensive and requires extensive memory resources, and more importantly, it is often associated with over-fitting problems. An approved idea to address this is to apply Transfer Learning (TL) using pre-trained models on a large natural image database (e.g., ImageNet [21]) instead of training a model from scratch. This approach, however, has not been effective in several previous studies [4, 7, 60] due to primary divergences in pattern, size and data feature between natural and medical images. All of these issues motivated us to propose a new TL approach that, by addressing the challenges outlined above, could achieve SOA accuracy in classifying cytological images of breast cancer.

The scope of this dissertation concerns a fully automated breast cancer classification using the TL approach. The work includes issues related to all stages of data acquisition, data preprocessing, nuclei segmentation, and binary image classification. This research was carried out on real-case cytological images of breast cancer prepared by the specialists of the Regional Hospital in Zielona Góra, Poland.

1.2 Main Goals

The main goal of this work is to develop an efficient medical decision framework based on DL models for classifying cytological images of breast cancer. Our focus is on designing a new TL approach that can address the weaknesses of the methods presented in the literature and achieve satisfactory performance in the classification task. The performance of the proposed framework is tested on real-life medical data accumulated from a number of anonymous patients. The dissertation thesis can be formulated as follows:

Automatic binary classification of breast cancer images with high accuracy (more than 98%) is possible even with the availability of a small amount of annotated data using the TL technique. The main idea is to use an auxiliary data set compatible with cytological images, instead of using natural images (e.g. ImageNet), which makes the features learned by the model useful for deciding on cancer malignancy.

The task at hand is challenging, seeing that several sub-tasks must be performed, each with its own complexities:

- Accumulating medical images taken from biopsy specimens and constructing a database of cytological images of breast cancer.
- Performing all required preprocessing operations on the collected images.

- Performing the segmentation task and preparing corresponding binary maps in order to isolate meaningful regions of the image (cell nuclei) from other background patterns that are not worth analyzing.

- Experimentally evaluating different thresholding techniques as part of the segmentation task in order to select the most efficient one. If existing solutions do not yield satisfactory results, a new approach needs to be developed.

- Designing a pipeline for binary classification of breast cancer cytological images, which includes the stages of data acquisition, image patching, patch selection, network pre-training, and image classification.

- Operating six SOA deep CNNs for the classification task and comparing their performance to find the most efficient network in terms of accuracy.

- Exploring three different scenarios for training, as well as two different approaches for fine-tuning deep CNNs.

To demonstrate the superiority of the proposed framework over those presented in previous studies, the experimental results obtained from this research are finally compared with five SOA studies previously conducted on the same data set. The next section outlines the main contributions of this thesis.

1.3 Novel Contributions

The main novel contributions of this thesis can be summarized as follows:

(1) Elimination of the need to have a large number of annotated cytological images for the binary classification;

(2) Significant reduction of the model training time and ability of the model to achieve high accuracy early in the training process;

(3) Examining nine different thresholding techniques to perform the segmentation task and determining the most efficient algorithm by comparing the results of their application on sample images.

(4) Employing six SOA deep CNNs and determining the most efficient network for the binary classification of breast cancer cytological images.

(5) Exploring five different scenarios for training, as well as fine-tuning networks and determining the most effective approach by analytical comparison of the results obtained.

(6) Implementation of several pre-trained deep CNNs on histopathological datasets and making their weights publicly available for use by researchers in this field to initialize the network;

(7) Effective learning by any DL model with a very limited number of labeled images;

1.4 Thesis Structure

This dissertation consists of five chapters which are organized as follows:

- Chapter 1; briefly explained the subject under study and stated the motivation for conducting this research. The main objectives and novel contributions of this research were summarized in the continuation of this chapter. Meanwhile, the last part of this chapter was dedicated to mentioning the published articles.
- Chapter 2; provides the necessary background on the subject of diagnosing breast cancer using computer-aided systems, and then reviews recent related works in this area to outline the contributions each has made. The application of transfer learning to solve medical image classification problems, specifically breast cancer images, is discussed later in this chapter, and related studies in this area are reviewed.
- Chapter 3; is dedicated to a comprehensive explanation of the proposed method. In this chapter, we describe all the stages of the designed pipeline separately.
- Chapter 4; presents the experimental results of all the strategies applied in this research and then discusses the results analytically.
- Chapter 5; outlines conclusions of the work along with future research directions.
- Appendix A; in addition to all the chapters mentioned, this thesis also includes an appendix, in which the confusion matrices for different scenarios are presented.

1.5 Publications

- *__M. A. Shamshiri__, A. Krzyżak, M. Kowal and J. Korbicz, Compatible-domain Transfer Learning for Breast Cancer Classification with Limited Annotated Data, Medical Image Analysis, 2022.

  *This article has been submitted to the Medical Image Analysis Journal, currently under review.

- M. Lazo-Cortes, J. Martnez-Trinidad, J. Carrasco-Ochoa, V. Valev, **M. A. Shamshiri** and A. Krzyżak, Taking advantage of typical testor algorithms for computing nonreducible descriptors, 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM'2022), Feb. 3-5, 2022. Published 15.11.2021 (online).

Chapter 2

Background and Literature Review

2.1 Introduction

This chapter consists of two main sections. In the first part (Section 2.2), we briefly introduce breast cancer and further refer to the global statistics of this disease among women. The screening and diagnosis methods of breast cancer are also explained in the very last part of this section. In the second part (Section 2.3), we review the latest techniques presented on the subject of Computer-aided Diagnosis (CAD) systems for breast cancer. The application of TL for medical image analysis is also explained in the last part of this section.

2.2 Breast Cancer

2.2.1 Basic Concepts

Cancer is caused by abnormal changes or mutations in the genes responsible for cell growth [8]. These genes are located in the nucleus of cells and are constantly active. Cells also have a specific lifespan in which they multiply and divide to replace old cells with new ones. The unrestrained cell growth in the body prevents the proper

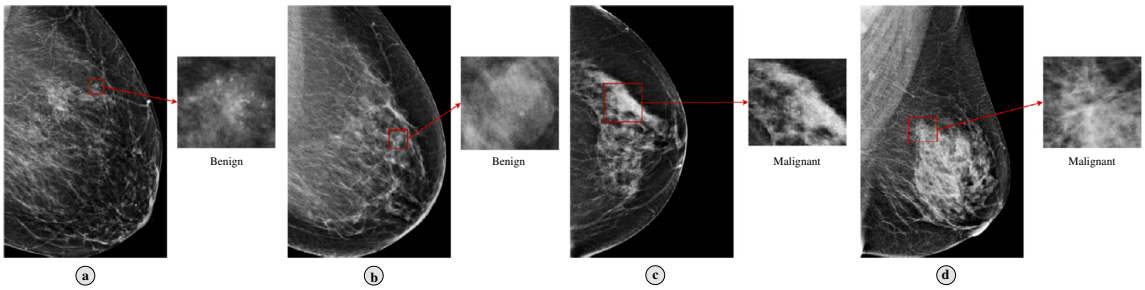


Figure 1: Four examples of biopsied breasts, the first two specimens (a and b) being identified as benign masses and the last two samples (c and d) being diagnosed as malignant masses (the image is taken from [73]).

replacement of dead cells with new ones, and as a result, the process of cell growth in the body gets out of control and leads to further problems. This may occur in one cell or a small group of cells in the body that eventually form a mass of tissue, so-called a tumor.

Although all tumors are caused by uncontrolled cell growth in the body, not all of them are necessarily cancerous. A tumor can be benign or malignant. A malignant tumor (cancerous tumor) spreads through the lymphatic system to various organs in the body and extracts nutrients from the body's tissues. While a benign tumor (a non-cancerous tumor) does not invade neighboring tissue and is therefore controllable (Mahmood *et al.* (2020) [49]). Figure 1 shows benign and malignant masses on breast mammography images.

Breast cancer is a type of cancer that includes a group of diseases in which cells in the breast tissue are affected and divided uncontrollably, eventually leading to a mass or tumor. There are four main subtypes of breast cancer that have different characteristics and therefore require different treatments (Yersal *et al.* (2014) [93]). The female's breast structure is mainly composed of 3 components, namely lobules, ducts, and fatty tissue. Lobules are milk-producing glands. The milk is carried to the nipple through small tubes called ducts. All these components (lobules and ducts) are

covered with fat that gives the breast shape and size. This structure is illustrated in Figure 2. Research has shown that most breast cancers start in the lobules, and have the potential to spread to other parts of the body, most commonly the liver, lungs, bones, and/or brain (Shumway *et al.* (2020) [75]). This highlights the importance of early detection of breast cancer, as the effectiveness of treatment largely depends on the timely detection of cancer.

2.2.2 Epidemiology

According to the World Health Organization (WHO) [79], in 2020, cancer was the second most common cause of death worldwide after cardiovascular disease, accounting for every sixth death across the world. In terms of new cases, breast cancer was the most common type of disease worldwide, with 2.26 million cases. Among women, breast cancer is the most frequently diagnosed cancer (Ji, *et al.* (2020) [37]),

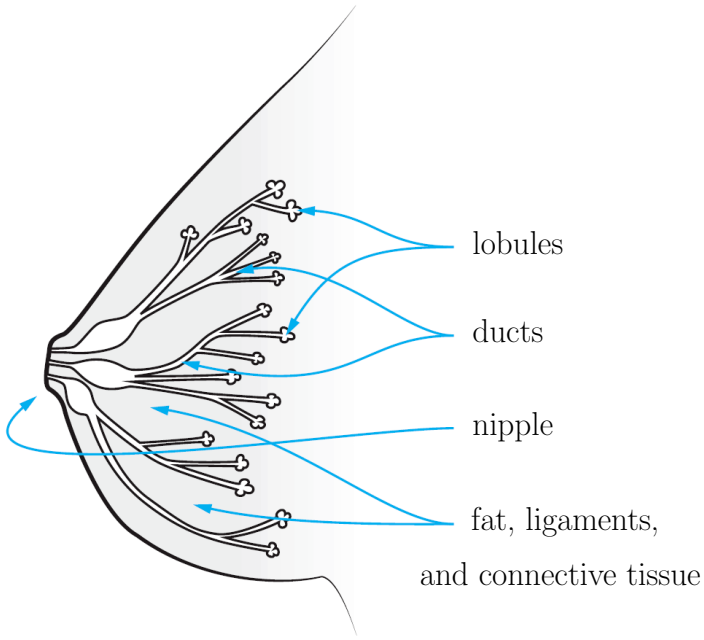


Figure 2: Female breast structure consists of lobules, ducts and adipose tissue.

accounting for 1 in 4 cancer cases and, according to Xu, *et al.* [90], ranked first in 155 countries (out of 185 countries) in 2018. Mortality profile among women shows that breast cancer is the fifth leading cause of cancer death in women and is recognized as the most important cause of cancer death in 103 countries in 2018.

Over the past 30 years, even with the rapid development of prevention and treatment methods, the incidence of breast cancer has steadily increased globally (Ruan, *et al.* (2021) [66]). Despite the increase in incidence, breast cancer survival rates have improved over the past three decades thanks to effective diagnostic and therapeutic methods (Tabár, *et al.* (2019) [83]). However, there are still areas in the world where the mortality rate due to breast cancer is increasing. This reflects the fact that breast cancer mortality rates vary around the world for a variety of reasons, the analysis of which is beyond the scope of this study.

2.2.3 Screening and Detection

Breast cancer is usually diagnosed either before the onset of symptoms, during screening, or after a lump is detected in a patient's breast (Shumway, *et al.* (2020) [75]), which is the most common symptom of this cancer. Since breast cancer is usually asymptomatic in the early stages (Siegel, *et al.* (2021) [76]), early detection of breast cancer with mammography plays an important role in reducing the risk of mortality. Delays in the diagnosis and treatment of breast cancer, on the other hand, necessitate the use of more invasive treatment procedures and potentially increase the mortality risk.

One of the most reliable and accurate methods of evaluating breast masses is known as the triple test, which is based on three medical examinations including physical examination (palpation), mammography, and fine-needle biopsy. Fine-needle biopsy without aspiration, which is one of the most popular modalities for diagnosing breast cancers [91], consists in obtaining material directly from the breast tumor.

This method offers fast results without significant discomfort and scarring, and also allows treatment options to be clarified for the physician before any surgery [8]. The collected material is finally examined under a microscope by a cytologist to determine the prevalence of cancer cells, which is clearly a task that requires extensive knowledge and experience. This is exactly where a CAD system, along with the medical imaging techniques, can help cytologists as an assistant to accurately diagnose cancer cells. The advantage of such a system is that it allows the analysis of biopsy images on a large scale automatically, leaving only the uncertain cases that require further examination to the cytologists. In this regard, we intend to employ the TL technique to design a new DL-based framework for classifying breast cancer biopsy specimens into benign and malignant categories, which can perform this task as accurately as possible and achieve SOA results.

2.3 Related Literature

This section is divided into two parts. In the first part, the motivation of designing a computer-aided system for diagnosing breast cancers is summarized, and then we review recent related works in this area to reveal the contributions each has made. The second part explains the idea of using the TL technique in medical imaging, and then it deals with recent related works in the literature.

2.3.1 Computer-Aided Diagnosis Systems

Computer-aided diagnosis systems are widely used in biomedical engineering and have been developed to assist physicians in the early detection of breast cancer. As the population ages, epidemics increase, and medical personnel decline, the development of computer-based systems that support and expedite the diagnostic process has become essential. Thanks to advances in electronics, computer science, and

physics, various medical imaging technologies such as digital mammography (DM), ultrasound (US), and magnetic resonance imaging (MIR) have been introduced into modern health care systems. As a result, it is now possible to visualize pathogenic changes with great precision. Paradoxically, high-resolution imaging poses new problems due to the fact that physicians must carefully analyze large volumes of raw data. To support medical staff in performing this time-consuming task, it is essential to develop effective image processing methods to extract knowledge that helps diagnose diseases. Here, we discuss the latest related research and proposed techniques in the development of such systems, specifically designed to diagnose breast cancer.

To definitely confirm cancer and determine its type, cell material must be collected by biopsy and analyzed by pathologists. To facilitate the work of pathologists, many different computer methods are developed to improve and automate histopathological and cytological diagnostics. For example, Roy, *et al.* (2019) [65] studied two patch-based classification strategies for automatic classification of histopathological breast cancer images using 5 custom CNN models. In the first strategy, in order to classify the image without error, all image patches had to be classified correctly. In the second strategy, a label of the image was determined by a majority voting among patches from that image. Proposed techniques were tested on ICIAR 2018 breast cancer histopathology image data set. For the first strategy, accuracy was 77.4% for 4-class and 84.7% for 2-class classification problems. The second strategy based on majority voting gave an accuracy of 90% for 4-class and 92.5% for 2-class classification. Xu, *et al.* (2020) [89] developed a method of deep selective attention to select valuable regions from histopathological images to facilitate the task of classifying these images. The proposed model was composed of an LSTM-based decision network (DeNet) and a soft attention (SaNet) classification network based on residual units. The decision network decided where to crop and whether the cropped patch was salient, then it fed the classification network, which in turn provided feedback to the decision network to

adapt its selection policy. The approach's effectiveness was evaluated on BreakHis image collection, where it achieved approximately 98% classification accuracy while only taking 50% of the training time of the hard-attention approach. Miselis, *et al.* (2019) [51] explored the potential of SOA deep neural networks architectures to classify breast cancer based on cytological images of FNBs. Five different CNN architectures have been evaluated: AlexNet, GoogleNet, SqueezeNet, DenseNet, and Inception-V3 in terms of the binary classifications of breast cancer. They found Inception-V3 the best model reaching 91.86% accuracy and 0.97 value for area under the ROC curve (AUC). Most of the presented approaches classify whole images or cropped patches, but Kowal, *et al.* (2021) [45] suggested an alternative approach based on the single-cell nuclei classification. Breast cancer cytological images were first segmented using the U-Net neural network and marker-controlled watershed transform. Then, individual cell nuclei were classified as benign or malignant based on the set of handcrafted features. For 2-class classification, SVM classifier reached 88.2% accuracy.

Pramanik, *et al.* (2015) [56] proposed an automatic CAD approach to recognize early breast cancer by analyzing breast thermograms. The proposed system was based on an artificial neural network and consisted of three stages: image segmentation, feature extraction, and classification. Using the initial feature point image of each breast thermogram along with principal component analysis (PCA), they were able to extract the most important features and minimize computations and information redundancy. A feed-forward multilayer perceptron (MLP) consisting of four layers was finally used for the classification task. They eventually achieved 90.48% accuracy, which was comparable to existing methods. Shi, *et al.* (2022) [74] proposed an efficient multi-task network for breast cancer diagnosis in US images designed for use on mobile devices. The great advantage of the proposed method is its lightness, which requires only 20 Megabytes (MB) to deploy, while being able to quickly detect breast cancer. They incorporated an auxiliary task (segmentation), into their

primary task (tumor classification), which allows the network to focus only on the regions where the tumor is present, and to extract and learn the descriptive features. The classification task was performed on a data set of US images (consisting of four different public data sets), and they attempted to minimize system error by proposing a new numerically weighted cross-entropy loss function. Comparison of the results obtained from the proposed method with a single-task classification (Single-CLF) and segmentation model (Single-SGM) built by MobileNet-V1 showed that the proposed method significantly increased the classification accuracy, although no improvement was made in the segmentation task.

Unlike most studies of breast cancer classification in US images that focus on the binary classification of benign versus malignant lesions, Behboodi, *et al.* (2021) [13] came up with the idea of increasing the number of classes to achieve more satisfactory performance. In this regard, they proposed a DL-based approach to address a multi-class classification problem in US images having limited annotated data. In addition to considering three subtypes of breast cancer named fibroadenoma, cyst, and invasive ductal carcinomas (IDC) in their classification problem, they explored the idea of considering image backgrounds as a fourth class and showed that this improved the accuracy of the model. To evaluate the effect of adding more classes on their network performance, they compared the results of the 4-class classification problem with the binary (IDC vs. other classes) classification problem. Their results ultimately showed that adding more classes improved the AUC score for the networks employed (i.e. ResNet-34 and MobileNet-V2) by the factor of 31% and 9%, respectively. Qiu, *et al.* (2017) [59] applied end-to-end DL approach to classify breast masses without their segmentation and feature extraction. The effectiveness of the constructed classifier was assessed for a set of 560 images (280 benign and 280 malignant) using 4-fold cross-validation. Experimental results revealed that the proposed CAD yields an overall AUC value of 0.79. Obtained results demonstrate that end-to-end DL models

can avoid the potential errors or biases introduced in conventional CAD by lesion segmentation and suboptimal feature extraction. However, authors warn that the performance of developed DL-based CAD schemes may not be statistically higher than the conventional schemes and further studies are necessary. Nascimento, *et al.* (2016) [54] extracted and selected morphological features from 100 US images and applied SVM and shallow neural network for binary classification of breast cancer. The best results obtained for accuracy and AUC were 96.98% and 0.98, respectively, both with neural networks using the whole set of features. The best AUC obtained in this study is higher than the value of 0.942 received by Flores, *et al.* (2015) [25] and the value of 0.565 obtained by Alvarenga, *et al.* (2007) [3]. Rasti, *et al.* (2017) [61] extracted DCE-MRI (Dynamic Contrast-enhanced Magnetic Resonance Imaging) features from segmented ROIs and classified breast cancer using a mixture of CNN networks. ME-CNN (Mixture Ensemble Convolutional Neural Network) comprises multiple CNN experts and one CNN gating member which combines the experts' responses. To test the effectiveness of the proposed CAD, authors have used 112 DCE-MRI breast examinations from high- or intermediate-risk patients. The best classification accuracy of 96.39% was obtained by the model composed of three CNN experts and one convolutional gating network. Compared with existing classifiers and ensemble methods, a proposed mixture of neural networks achieved competitive classification performance while preserving a compact structure and fast execution time.

2.3.2 Transfer Learning for Medical Imaging

Transfer learning is a widely used technique by which previously learned knowledge can be intelligently applied to solve new problems more accurately. The recent studies on TL in medical imaging have mainly adopted two different approaches to take advantage of this technique. We can therefore divide the recent research in the

literature into two groups and review each separately.

The first group includes research that has employed a pre-trained CNN as a feature extractor. This essentially means that an input image is given to a pre-trained CNN and then the output of the CNN, which is a feature vector, is extracted from a specific layer of the network. The feature extraction is typically done from the intermediate layers of the CNN, as it has been verified that the outputs of these layers are highly discriminative features that can be used to achieve great performance in visual classification tasks (Razavian, *et al.* (2014) [62]). Then, in the next step, the extracted feature vector is given to a new classifier (e.g. SVM) and the final prediction about the image is revealed. On this subject, Ursuleanu, *et al.* (2021) [85] adopted a similar approach and utilized the high capability of DL for pathology detection in chest radiograph data. For this purpose, after feeding the input image to a CNN that had previously trained on the ImageNet data set, they extracted several descriptors. Then, a linear kernel SVM was used for the final classification phase. They ultimately showed that the best performance can be achieved by using the features obtained from the deep network along with some auxiliary features. A very similar technique was employed by Ginneken, *et al.* (2015) [26], who extracted 4096 features from the penultimate layer of a pre-trained (ImageNet) CNN, and then used an SVM estimator with linear kernel for the nodule classification in computed tomography (CT) scans. They showed that CNN features have great potential for use in diagnostic tasks in volumetric medical data. Nevertheless, they eventually concluded that systems built on CNN features alone performed worse than advanced optimized CAD systems and that integrating CNN-based features with the handcrafted features could improve the performance. As an alternative to handcrafted features to address the classification of mass lesions in mammography images, Arevalo, *et al.* (2016) [9] designed a framework in which they used a deep CNN to generate high-level features of the input image. They finally used an SVM classifier for the classification stage and

showed that the proposed approach outperforms the SOA methods by 6% in terms of AUC.

The second category involves studies that are not directly dependent on the features extracted from the previously trained models, but in which, in addition to replacing the final layer with a new classifier, some of the previous layers are also selectively retrained. In general, the role of the initial layers in deep neural networks is to capture generic features, while the succeeding layers focus more on the specific task at hand. Thus, higher-order feature representations in the base model are typically fine-tuned to be more relevant for the specific task. Fine-tuning allows the model to apply past knowledge to the new problem and learn some task-related things by updating the network weights. As an example of research conducted on this basis, we can refer to Chen, *et al.* (2015) [16] who proposed a TL strategy for localizing fetal abdominal standard planes in US images. To transfer knowledge, they leveraged a CNN trained on a large database of natural images, so they fine-tuned five layers of the base CNN, as well as trained three fully connected (FC) layers appended to the end of the network. Compared with previous works based on low-level features, their approach was able to represent the complicated appearance of the input data and achieved a better classification performance. Alzubaidi, *et al.* (2021) [6] proposed a novel TL approach, utilizing a large number of unlabeled images to provide a pre-trained model compatible with the features of medical images. The idea was to use unlabeled medical images to pre-train and adapt the model to the characteristics of medical data in the first step, and then to train and fine-tune the model on the limited data of the target set. They then demonstrated the effectiveness of their method in classifying histopathological images of breast and skin cancers into benign and malignant categories. This approach significantly improved the performance of both their classification scenarios compared to when the model was trained from scratch. One of the most obvious advantages of their method is the elimination of the need to

annotate large volumes of medical data, which avoids a time-consuming and costly process. It has also been noted that their method is not limited to skin and breast cancer scenarios and can be effective in the same-domain tasks dealing with images with similar texture and pattern (e.g., histopathological images of colon and bone cancer). Despite the aforementioned advantages of their method, there is still a need to acquire large volumes of unlabeled images for the pre-training phase of the model. Considering that obtaining high-quality medical images is not an effortless task and in some cases may not even be feasible, proposing a method that does not depend on a large number of unlabeled images can certainly be a major contribution to this area. Chen, *et al.* (2019) [17] dealt with 3-dimensional medical data and designed a novel DL framework to address the task of classifying pulmonary nodules as well as segmenting CT scans of the liver. They combined multiple data sets of several medical challenges with diverse modalities and carried out effective TL by leveraging different CNNs trained on the acquired data. The results of their experiments showed that the proposed method accelerated the training convergence speed 10 times compared with training from scratch and also improved the accuracy of the model, ranging from 3% to 20%. Samala, *et al.* (2017) [68] applied multi-task TL to classification of mammograms by CNNs. The idea of the proposed approach was to transfer the knowledge learned from non-medical images to medical diagnostic tasks through supervised training, and to increase the generalization capabilities by simultaneously learning auxiliary tasks. The study shows that multi-task TL outperforms single-task TL in terms of AUC. The results demonstrated that learning efficiency and prediction accuracy could be significantly improved thanks to enriching the training set with heterogeneous mammograms acquired from different imaging technologies. Shahidi, *et al.* (2020) [72] verified the effectiveness of several DL models in the task of classifying histopathological samples. The study identified the most accurate models in terms of the binary, four, and eight classifications of breast cancer for publicly available

collections of histopathological images. Fine-tuned SENet-54 model pre-trained on ImageNet has shown the highest accuracy (99.87%) for binary classification of images from the BreakHis database. Several other SOA models have achieved accuracies that are only by a small fraction lower to the best result. The highest accuracy (97.5%) for the four-class classification problem based on the BACH image collection was achieved by as many as 5 models: SENet-154, ResNet-V2, ResNeXt-101, NASNet-A-Large, and DPN-131. Examination for eight-class classification using BreakHis database showed that the best, in this case, is InceptionResNet-V2 model, which gained accuracy 97.25%. As another example of a study conducted on the BreakHis database, Qi, *et al.* (2019) [58] proposed a deep active learning framework for the classification of breast cancer histopathological images. This approach aims to maximize the learning accuracy using a limited number of annotated images. The DL model (AlexNet pre-trained on ImageNet) was iteratively updated using the most valuable images. Thanks to this approach, the annotation costs were reduced to 66.67%, with almost unchanged classification accuracy at 89-91% for images and 91-93% for patients.

2.4 Summary

The first part of this chapter (Section 2.2) provided basic knowledge about the biology and epidemiology of breast cancer. As discussed, breast cancer has long been a serious health problem in modern civilizations that has affected the entire world. Then, the significance of timely detection of the disease was emphasized, and FNB was introduced as one of the most common diagnosis methods of breast cancer.

In the second part of this chapter (Section 2.3), it has been shown that, thanks to the use of ML/DL techniques, the task of detecting and classifying breast cancer images can be done automatically, although in some cases the accuracy of the systems

is still not satisfactory so that they can be operated in clinical settings. This highlights the need to propose more efficient techniques that can be based on TL concepts. In the last part of this section, recent studies on the application of TL for medical image analysis were discussed. Research on this topic in the literature has been divided into two groups, including those that utilize CNNs as feature extractors, and the second group, which focuses on fine-tuning the networks. All of the reviewed research confirms that the extraction of high-level features from CNNs can help improve the performance of diagnostic systems in a variety of tasks, although as we have seen, the use of these features alone for achieving high performance in some scenarios may not be sufficient.

Chapter 3

Methodology

3.1 Introduction

This chapter comprehensively explains the methodology used in this thesis. In order to classify breast cancer cytological images into benign and malignant categories, several steps need to be performed. After presenting an overview of the proposed method (Section 3.2), we will explain each step separately. Briefly, the first step is to acquire medical images from the biopsy specimens described in Section 3.3. In the second step, in order to isolate important areas of the image from other background textures, the segmentation task is performed (Section 3.4). Due to the fact that training deep CNNs with large-sized images is computationally expensive, the images are partitioned into smaller patches according to the specified sizes (Section 3.5). However, not all patches generated are worth analyzing, and undesired patches must be filtered through a mechanism (Section 3.6). At this point, aiming to train deep CNNs and accomplish the classification task, a set of desired patches is formed, which is split into sub-sets of training, validation, and test (Section 3.7). Finally, after initializing the networks with the updated weights obtained as a result of pre-training on the BreakHis data set, the task of classifying the desired patches into benign and

malignant categories is performed by deep CNNs. The architecture of each of the deep CNNs employed in this study is described separately in Section 3.8.

3.2 Overview

We propose a new TL approach to classify breast cancer cytological images into two categories: benign and malignant. Taking into account the ineffectiveness of employing natural images in TL to solve biomedical-domain problems, we propose the idea of compatible-domain TL. This means that instead of using natural images (i.e. ImageNet) that are not essentially compatible with medical data, the pre-training phase of the model is performed employing histopathological images. We then fine-tune pre-trained models on the target data set containing limited cytological images. Figure 3 illustrates a schematic of the typical approach to using the TL technique, along with the approach presented in this thesis.

Despite the fact that histopathological and cytological images are inherently inconsistent in some respects, they both belong to the same image modality (i.e., microscopic images), and we demonstrate in this study that following the proposed method makes the features learned by deep CNNs during the pre-training procedure relevant to what is obtained from the target images. Given that many collections of histopathological images have become publicly available in recent years, acquiring this type of medical data for pre-training models is not problematic. Hence, this approach helps to address the sparsity of training data as we no longer need to allocate a significant portion of the target data set (i.e. cytological images) to train the model. We will explore three distinct scenarios for training deep CNNs and will apply two different approaches to fine-tuning the models, and finally compare and analyze the results of these five different approaches to find the best and most efficient strategy. The six networks we used in this study include 1) DenseNet-169, 2) InceptionResNet101-V2,

3) Inception-V3, 4) ResNet-101, 5) VGG-16, and 6) VGG-19. These are well-known deep CNNs that have already proven their capability to solve classification tasks in ImageNet Challenge.

To demonstrate the effectiveness of the proposed approach, we propose and compare various scenarios for pre-training and fine-tuning the models, which are listed below:

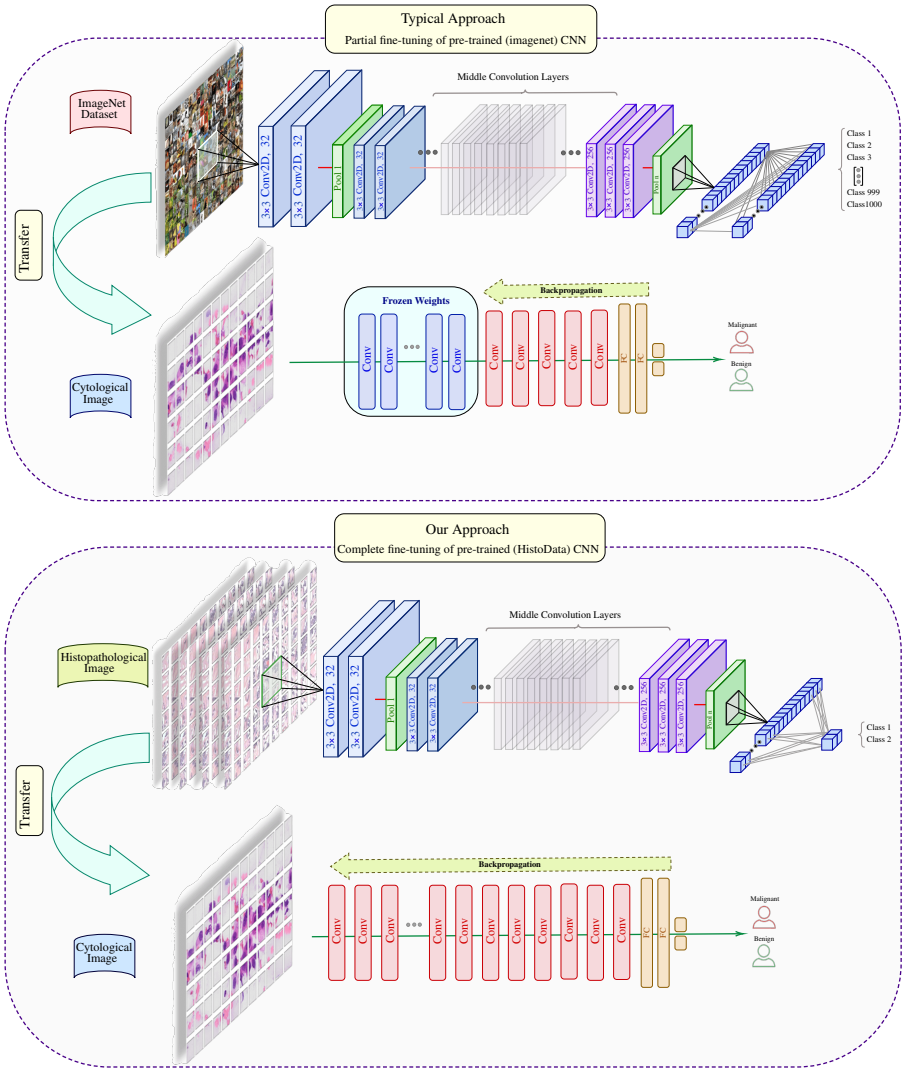


Figure 3: Representation of a typical strategy alongside our proposed approach (CDTL) to apply TL technique in the classification of breast cancer cytological images. In the typical approach (upper one), a set of natural images is used as a data source for the pre-training phase, while in the proposed method (lower one), histopathological images are employed as an auxiliary data source for this purpose. The network fine-tuning is also performed differently in these two approaches.

(1) Pre-training six SOA deep CNNs using breast cancer histopathological database, then fine-tuning the networks with cytological images of the target data set.

(2) Pre-training six SOA deep CNNs using ImageNet data set, then fine-tuning the networks with cytological images of the target data set.

(3) Training six SOA deep CNNs from scratch with cytological images of the target data set.

We also perform *complete* and *partial* fine-tuning approaches over the networks in each of scenarios 1 and 2, the former meaning the weights of all layers are updated, and the latter meaning that the backpropagation operation is performed only over the last few fully connected layers of the network.

To be more precise, as shown in Figure 4, we perform the pre-training phase of the models using two separate data sets, namely BreakHis and ImageNet. It should be noted that since various DL frameworks have made multiple pre-trained networks on the ImageNet data set publicly available, we utilize those networks, and obviously, we do not run the model training process on the ImageNet data set again. Then, the model weight initialization is accomplished in three different ways (modes) and the model is fine-tuned on the cytological images of the target data set. In the first mode, the network is initialized using the updated weights as a result of model pre-training on BreakHis data set. In this mode, the initialized weights of the network are frozen at first, and the fine-tuning operation is performed only on the output FC layers attached to the end of the network. Then, the last few layers of the network are unlocked so that they can be updated during the backpropagation operation. Unlocking the network layers continues incrementally to witness the effect of updating more network weights on its performance. The second mode is almost the same as the first one, except that this time the network is initialized using the updated weights as a result of model pre-training on the ImageNet data set. The

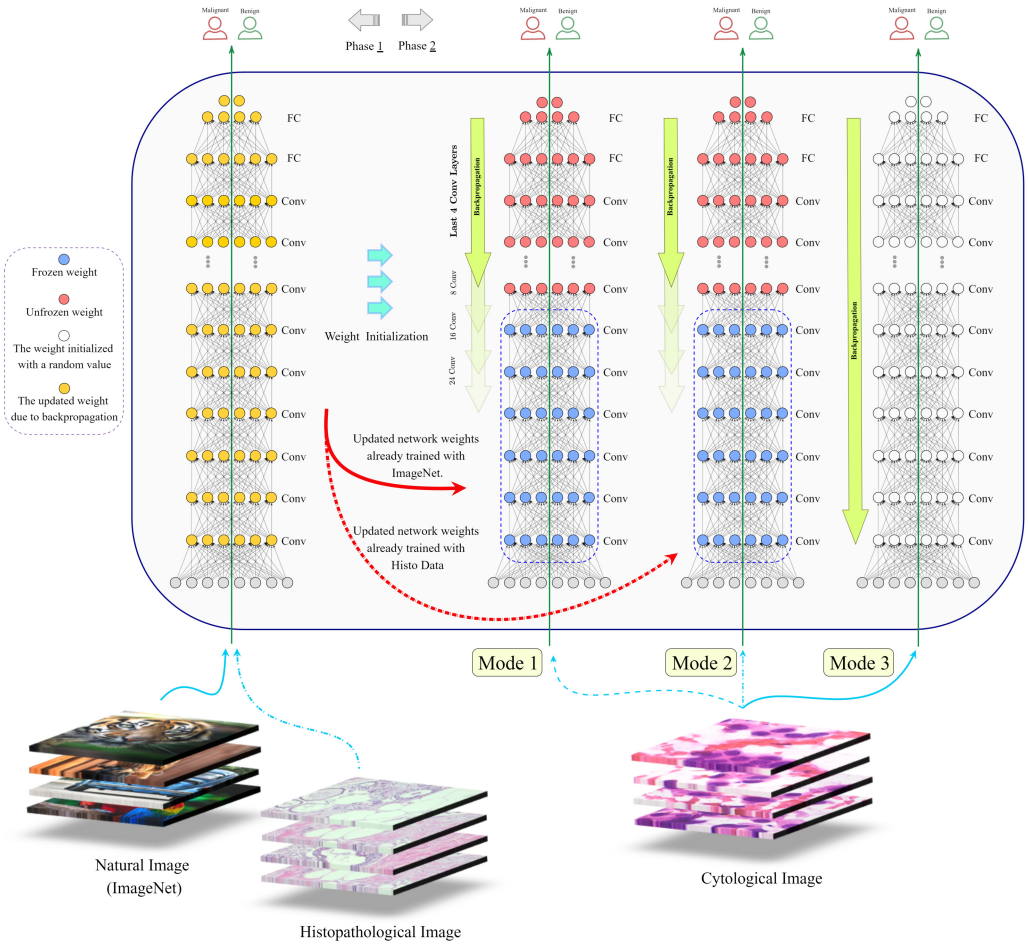


Figure 4: Approaches used for initialization and fine-tuning of the network. The network initialization is done in 3 different modes: 1: using weights updated as a result of pre-training on histopathological images (BreakHis), 2: using weights updated as a result of pre-training on a natural image data set (ImageNet), 3: using random values.

fine-tuning operation is followed with the same approach adopted in the previous mode. Finally, the third mode involves initializing the deep CNN with random values and training from scratch on the cytological images of the target data. Investigating this mode helps to observe the effect of using the TL technique employed in the first two scenarios.

Addressing the problem of classifying virtual cytology slides into benign and malignant categories is done through the pipeline we designed, which consists of eight separate modules as depicted in Figure 5. Briefly, the first step is to perform image

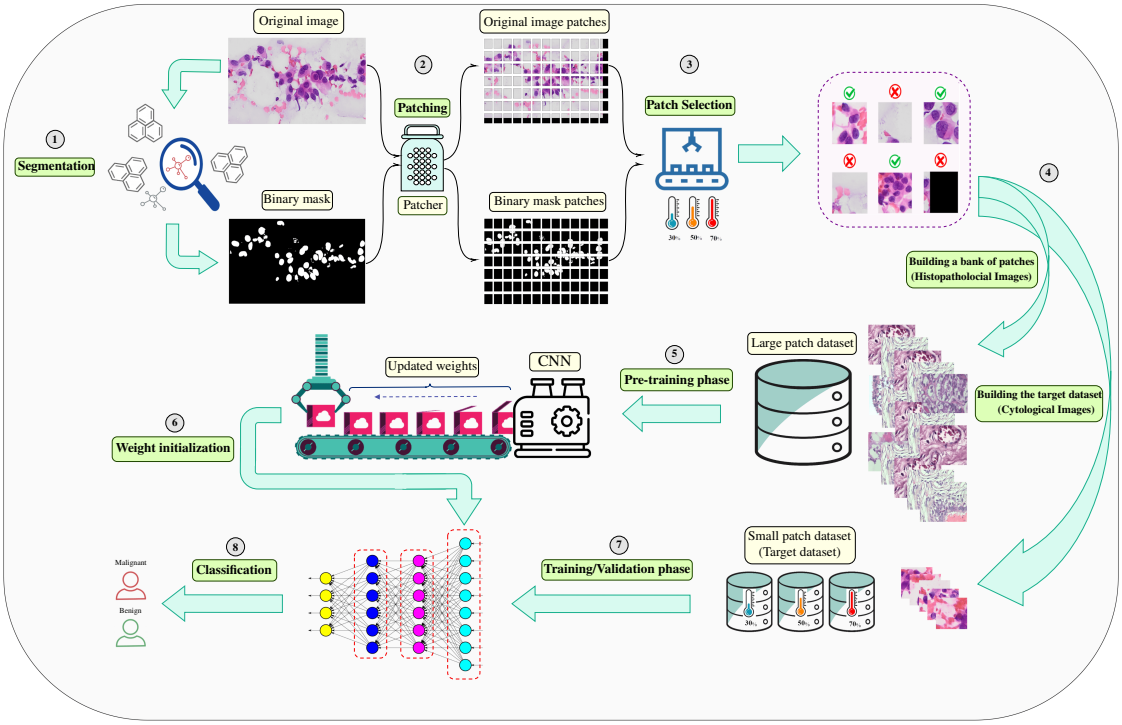


Figure 5: The method pipeline consisting of eight separate stages. 1) Segmentation, 2) Image patching, 3) Patch selection, 4) Building banks of patches, 5) Pre-training phase, 6) Wight initialization, 7) Training/validation phase, and 8) Classification.

segmentation to determine the exact location of the cancer cell nucleus in the target images. Then, in the next step, the images are partitioned into smaller image patches so that they can be used as input data to train the model. Seeing that not all patches constructed contain useful features for diagnosing the type of cancer, in the third step, the desired patches are selected through a mechanism based on the extent to which they contain determinative pixels. The appropriate patches obtained from the previous step are then used to form the model training data set. Next, the pre-training phase of the model is performed using histopathological image patches, as a result of which the model weights are updated. Finally, a new deep CNN is initialized using the updated weights, and the binary classification task is performed on cytological image patches. The proposed method algorithm is described in Algorithm 1. In the continuation of this section, we will explain each of these steps in detail separately.

Notation	Description
$\mathcal{T}$	Threshold algorithm.
$\mathcal{S}$	Image Segmenter: Produces a binary map of an RGB image given a threshold algorithm.
$\mathcal{P}$	Image Patcher: Partitions a large image into several square patches with regard to the given size.
$\mathcal{L}$	Patch Selector: Selects the desired patches w.r.t the given cellular coverage.
$\mathcal{D}_I^a$	An auxiliary data set (i.e. ImageNet), used to pre-train models.
$\mathcal{D}_H^a$	An auxiliary data set (i.e. BreakHis), used to pre-train models.
$\mathcal{D}^t$	The target data set containing cytological images of breast cancer.
$\mathcal{W}_H$	The weights of the pre-trained models updated on ImageNet data set.
$\mathcal{W}_I$	The weights of the pre-trained models updated on BreakHis data set.
$\mathcal{W}_C$	The weights for networks that are fine-tuned to the target data set.
$\mathcal{W}_R$	Random weights.

Table 1: Notations used in describing the proposed TL Framework.

Algorithm 1: Proposed TL Framework

Initialize: PatchBanks $\leftarrow$ [*EmptyList*]
Initialize: FilteredPatchBanks $\leftarrow$ [*EmptyList*]
for $image \in \mathcal{D}^t$ **do**
  for $size \in \{64_{px}, 128_{px}, 256_{px}\}$ **do**
    PatchBanks.append($\mathcal{P}(image, size)$);
    for $\mathcal{PB} \in PatchBanks$ **do**
      for $\mathcal{C} \in \{30_{\%}, 50_{\%}, 70_{\%}\}$ **do**
        FilteredPatchBanks.append($\mathcal{L}(\mathcal{PB}, \mathcal{C})$);

PHASE 1: Training deep CNNs with auxiliary data sets, $\mathcal{D}_i^a$, where $i \in$ [ImageNet, BreakHis] to get the updated weights.
for *number of iterations* **do**
  for *number of mini-batches* **do**
    Minimize $Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log \widehat{y} + (1 - y_i) \cdot \log (1 - \widehat{y})$;
    Update parameters (weights) of the network by gradient descending;

PHASE 2: Fine-tuning the networks with cytological images of the target data set, $\mathcal{D}^t$.
$\mathcal{M} = \{1$: 'Histopathological', 2: 'ImageNet', 3: 'random'$\}$;
for $mode \in \mathcal{M}.key()$ **do**
  if *mode == 1* **then**
    Initialize: $\mathcal{W}_C \leftarrow \mathcal{W}_H$;
  else if *mode == 2* **then**
    Initialize: $\mathcal{W}_C \leftarrow \mathcal{W}_I$;
  else
    Initialize: $\mathcal{W}_C \leftarrow \mathcal{W}_R$;
for *number of iterations* **do**
  for *number of mini-batches* **do**
    Minimize $Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log \widehat{y} + (1 - y_i) \cdot \log (1 - \widehat{y})$;
    Update parameters (weights) of the network by gradient descending;

3.3 Medical Image Acquisition

The first step in constructing the data set required to train the models is to acquire medical images from extracted cytological specimens. For this purpose, a specialized camera is placed on top of a light microscope. Then, to observe the structure of cancer cells by a specialist, the smears of cytological material obtained by FNB are fixed in spray fixative. The time between preparing the smears and their preservation in fixative never exceeds three seconds. Considering that most cells are colorless and transparent, the samples are stained in such a way that makes the cell tissue, specifically the cell nuclei, more visible. In this regard, the cytological material were stained with one of the most common staining agents, called Hematoxylin-Eosin (H&E [41]). Hematoxylin is a dye called hematein that is used to stain acidic (or basophilic) structures a purplish-blue. Eosin is an acidic dye that stains basic (or acidophilic) structures such as the cytoplasm red or pink. This technique can either be performed in a non-specific way, i.e., most cells are stained in almost the same way, or specific, meaning that specific chemical groups or molecules of cells are selectively stained. At this stage, the stained cytological material is scanned and digitized into virtual slides using the Olympus VS120 Virtual Microscopic System [12].

The system consists of a 2/3" charge-coupled device (CCD) camera and 40× objective, for a total of 0.172 µm/pixel resolution. A virtual slide is a massive digital image with an average size of 200000 × 100000 pixels. Capturing slides is done by scanning multiple times with the focus plane located at different positions along the Z-axis. These slides are then compiled, while only retaining the regions in each frame that are in sharp focus. Since not all parts of a slide necessarily contain useful medical information for analysis, a cytologist manually selected 11 distinct regions of interest (ROI) which were converted to 8 bit/channel RGB TIFF files of size 1583 × 828

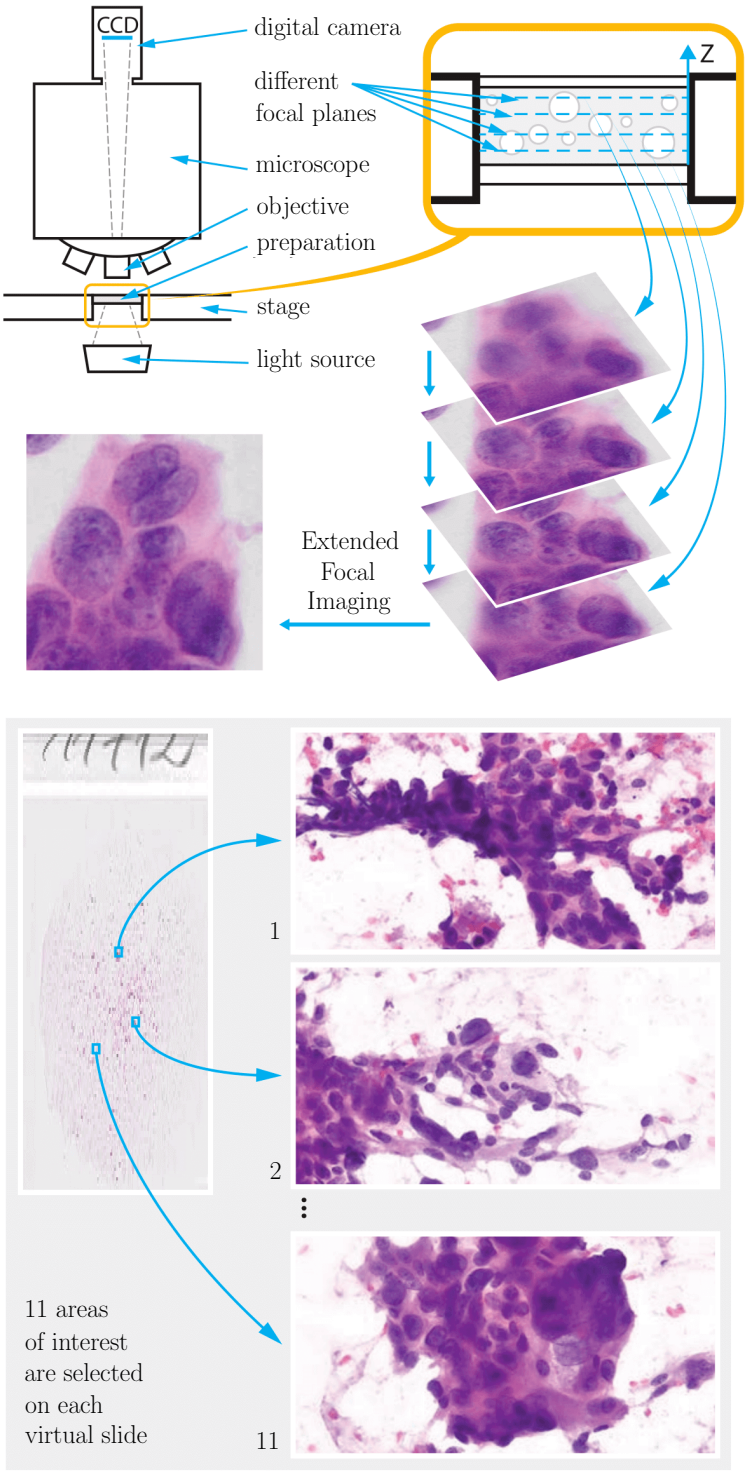


Figure 6: The process of scanning biopsy specimens using EFI, and selecting ROIs from the virtual cytological slides (the image is taken from [24]).

pixels.

The main criteria for selecting regions by the specialist was to ensure:

- The ROIs taken from malignant specimens contain only malignant cell nuclei
- The number of cell nuclei is sufficient (at least 10)
- Only select areas where the cell nuclei are clearly visible

It has to be acknowledged that all the materials tested in this study have been clinically examined by cytologists to diagnose cancer, and therefore there is no doubt about the type of cancer in the samples. Figure 6 summarizes the acquisition process using the virtual microscopy system.

3.4 Segmentation

The segmentation task is of great importance in the process of analyzing medical images. The purpose of this task is to isolate areas of the image that contain useful diagnostic information from other parts that are not worth analyzing. Given the fact that cell nuclei mostly contain useful diagnostic information, their analysis is an essential step in recognizing cancer malignancy. Malignant nuclei differ significantly from benign nuclei in various aspects such as shape, size, and textural pattern (Bankes, *et al.* (2007) [11]). Therefore, at this stage, our main goal is to isolate the cell nuclei from other objects on the image (e.g., red blood cells) so that the analysis can be performed only on important parts of the image. This task, however, is not straightforward as it is to be done automatically without the help of a human expert, and there are challenges in between. Generally, the cell clusters seen in cytological images overlap and their colors are mixed together so that they are not easy to distinguish from each other. This leaves no clear boundary between the cell nuclei, making the segmentation task more difficult. Since all subsequent analyzes will be based on

the results obtained from the segmentation stage, performing this step accurately will be effective in achieving high performance in the final classification. So far, various approaches and methods have been proposed for image segmentation [50, 96], each of which has advantages that are appropriate to apply in specific scenarios.

Segmentation, however, is not performed in a single phase and requires steps to pre-process the images. These steps include the following:

Color channels separation

As previously described, during the process of staining cytological images, cell nuclei react mainly with hematoxylin (blue), while the cytoplasm and red blood cells react mainly with eosin (red). Consequently, separating the colors of the RGB image helps us to easily set the cell nuclei apart from other image components and perform further analyzes solely on the cell nuclei. To extract the stain concentrations at each pixel, we use the matrix provided by the color deconvolution plugin for FIJI software [70], which is tuned to absorb a specific set of stains. The color deconvolution operation involves the decomposition of the input image into its constituent channels, each representing the concentration of each stain used (Haub, *et al.* (2007) [30]). As a result of applying the deconvolution matrix, three distinct images are obtained, including 1) deposition areas stained with hematoxylin, 2) deposition areas stained with eosin, and 3) residual areas. Figure 7 depicts a sample of the resulting images after the deconvolution operation.

Image normalization

There are a number of preprocessing techniques for normalizing medical images before any image analysis. A common approach, known as standardization, is done by

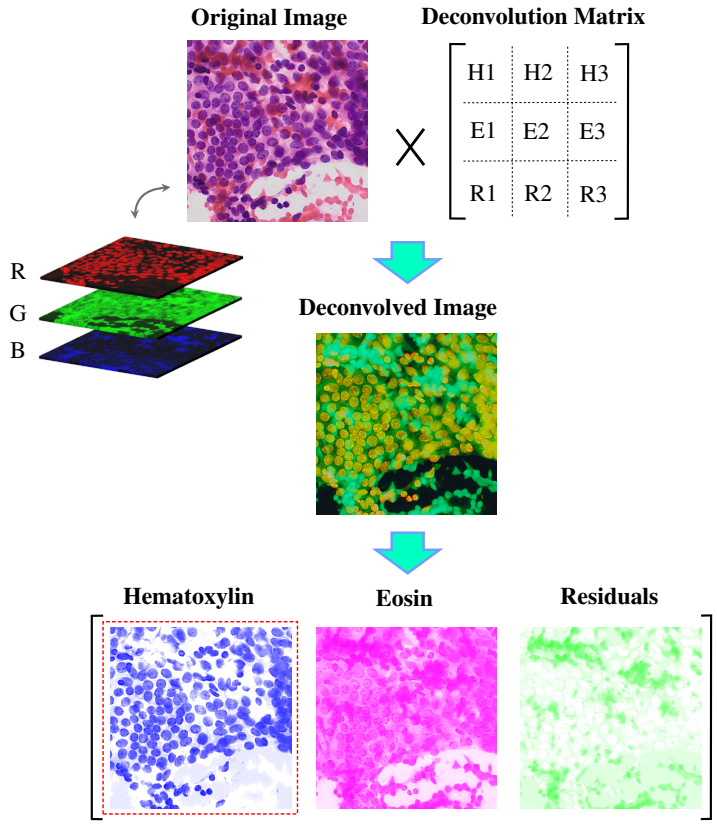


Figure 7: Representation of images obtained after performing H&E color deconvolution operation. The deconvolution matrix is first applied to the original image (RGB color channel). The color channels of the consequent image are then separated, resulting in Hematoxylin, Eosin, and Residuals images.

subtracting the mean value of the image intensity from each pixel value and then dividing the result by the standard deviation of the image intensity. The main purpose of standardization is to adjust the values measured at different scales to a common scale, since non-standard coefficients are not directly comparable. Out of the three images obtained as a result of the color separation step, we consider only the image stained with hematoxylin in which the cell nuclei are more prominent, so normalization is performed only on the bluish map.

3.4.1 Semantic Segmentation

After completing all the necessary preprocessing operations on the images, in this stage, in order to completely isolate the cell nuclei from other components of the image, the semantic segmentation task is performed. Semantic segmentation of an image is essentially a sort of classification of each pixel into a predefined category. Here we have considered 3 different categories, each pixel of the image belongs to one of them: a) the cell nuclei interior, b) the cell nuclei edge, or c) the background. The complex structure and heterogeneity of cell nuclei, as well as the overlap of cell nuclei with other image textures, make the process of labeling image pixels quite difficult and time-consuming. Doing this process based only on domain knowledge seems practically impossible, and using some form of ML techniques can help solve the problem. However, applying these techniques is not without its challenges and usually requires a large amount of data. This requires us to prepare a large number of manually segmented images, which is a very time-consuming task and requires a great deal of human effort. Nevertheless, it has already been shown by Ronneberger, *et al.* (2015) [64] that neural networks can be trained using relatively small data sets, yet without over-fitting problems. This was made possible by the use of a special neural network architecture called the U-Net network, along with the use of image augmentation techniques.

U-Net is a convolutional network architecture developed by Ronneberger, *et al.* specifically for the segmentation of biomedical images. The architecture won the cell tracking challenge at the International Biomedical Imaging Symposium (ISBI) in 2015 in a variety of categories, and has so far outperformed many of the previous methods proposed for the segmentation task. These facts convinced us to use it as our baseline to perform the segmentation task in this research. The architecture consists of two main paths. The first path, known as the encoder or downsampling path, comprises several convolution and max-pooling layers designed to capture the context of the

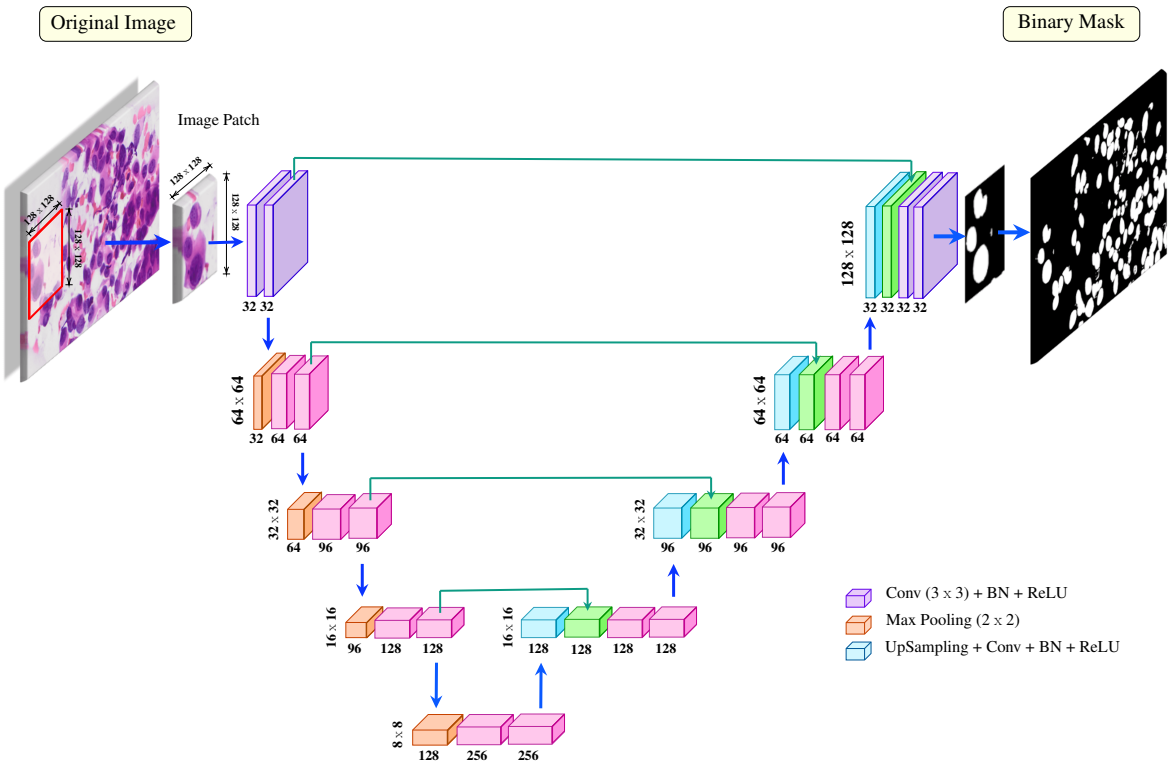


Figure 8: Scheme of the U-Net neural network.

image. The second path, known as the decoder or upsampling path, is designed to enable precise localization using transposed convolutions.

The U-Net neural network architecture is shown in Figure 8. As illustrated in the figure, the spatial size of the feature map is reduced by a factor of 0.5 at each downsampling step, and the number of feature channels is doubled. The convolutional blocks operated in the downsampling path consist of a convolution layer (kernel size 3×3) with rectified linear unit (ReLU) activations followed by batch normalization and dropout layers. Conversely, in the upsampling path, the spatial size of the feature map is doubled and the number of feature channels is halved in each step. In the last step of the upsampling path, the convolution layer (kernel size 3×3) with ReLU activation followed by a softmax activation function is responsible for generating the network output in the form of 3 feature maps. The spatial size of these maps is the same as the input image, and for each pixel a class probability distribution is defined

as shown in Figure 9.

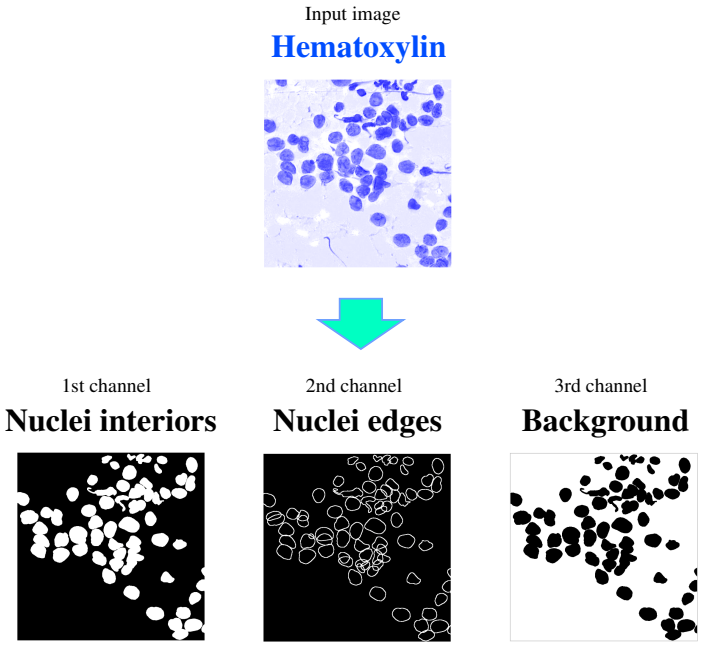


Figure 9: The resulting images of semantic segmentation using the U-Net network. The image representing hematoxylin concentration is normalized and is given as input to the network. The output of the network is the probability distribution of the classes (i.e., nuclei interiors, nuclei edges, and background).

To perform the segmentation task with U-Net, the network needs to be trained using a number of manually segmented cytological images. Due to the fact that the time costs associated with manually segmenting the entire set of images (550 images) were extremely high, we decided to train the U-Net with fewer images. For this purpose, we constructed a smaller set of images (set B) by selecting two ROIs for each patient, resulting in 50 ROIs for the benign category and 50 ROIs for malignant cases (a total of 100 ROIs for 50 patients). The data set was then divided into two parts: training and validation. The training set (B1) consisted of 50 manually segmented ROIs from 12 benign patients as well as 13 malignant patients (two ROIs per patient). The validation set (B2) also included 50 manually segmented ROIs,

randomly selected from 13 benign patients and 12 malignant patients. It should be noted that due to the limited number of manually segmented images, we have not formed any test set, as this reduces the size of the training and validation sets. The performance of the U-Net network was evaluated on the remaining ROIs of set A, none of which included in set B. Consequently, the quality of the segmentation can only be assessed visually because the images in our test set (set A) have not been manually segmented. The quantitative information of the assembled sets is shown schematically in Figure 10.

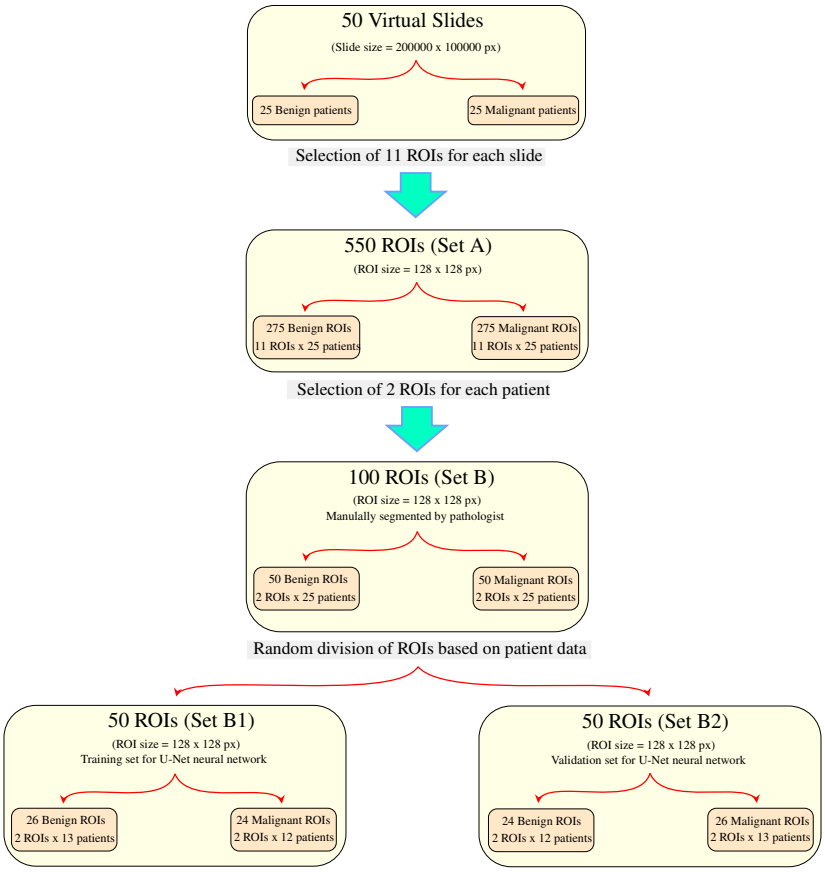


Figure 10: Schematic of the data distribution for training the U-Net network.

Despite the high capability of the U-Net network in conducting semantic segmentation, performing this task using U-Net causes us to encounter challenges when dealing with the subsequent task (i.e., classification). As mentioned earlier, using the

U-Net network depends on having a large amount of manually segmented images to be trained on. This prompted us to devote a sizable portion of the data set (i.e., 100 images out of 550 images) to network training, while not being able to use these images in the classification stage. Although there is technically no barrier to using the images employed for the segmentation task in the classification stage, this is not theoretically admissible. Using the same images to train the U-Net network, as well as to train SOA deep CNNs, the former to accomplish the segmentation task, and the latter to perform the classification task, makes the system biased against our data set. For the sake of this issue, we must discard those 100 images employed in semantic segmentation and use the rest of the images to train deep CNNs for binary classification. Obviously, as a result of this action, the number of images available to train deep CNNs becomes more limited than before, increasing the risk of overfitting as well as making it more difficult to achieve high classification accuracy. In addition, as stated earlier, quantitative evaluation of U-Net performance is not attainable since the test set images have not been manually segmented, thus the network performance can only be assessed visually. Because of the issues outlined above, and to have a more reliable system, we finally decided to operate an alternative solution, i.e. intensity thresholding, to perform the segmentation task instead of using the U-Net network. A detailed description of this approach is provided in the next section.

3.4.2 Intensity Thresholding

Image thresholding is a simple form of image segmentation in which each pixel value is replaced by 0 (black) if the intensity is less than a constant value, or otherwise by 255 (white). The resulting image will be a binary map in which the cell nuclei (which are white) are completely isolated from the other components of the image (which are black). In this study, we employ nine different thresholding algorithms presented in the Scikit-learn library, namely MINIMUM (Prewitt *et al.* (1966)

[57]), LOCAL (Chow *et al.* (1972) [18]), TRIANGLE (Zack *et al.* (1977) [95]), ISODATA (Ridler *et al.* (1978) [63]), OTSU (Otsu (1979) [55]), LI (Li *et al.* (1993) [48]), MEAN (Glasbey (1993) [27]), YEN (Yen *et al.* (1995) [92]) and SAUVOLA (Sauvola *et al.* (2000) [69]). It is worth noting that the thresholding algorithms provided in the Scikit-learn library are not limited to these nine techniques, and a number of other algorithms are also available to make use of. The main purpose of considering several thresholding algorithms in this study was to be able to select the most efficient one by comparing the results, thus making the designed framework more reliable. Despite the simplicity of the logic used by these algorithms, the effective application of these techniques has been confirmed in many recent studies to perform the segmentation task [20, 23, 36, 42].

To evaluate the performance of the above algorithms, we applied them to six original sample images and then compared the obtained binary masks to select the most efficient thresholding algorithm. Figure 13 shows the binary maps obtained by applying different thresholding algorithms to the RGB images. In order to quantitatively evaluate the performance of the applied algorithms and select the most efficient one for the segmentation task, we use Intersection-Over-Union (IoU), also known as the Jaccard Index, which is one of the most commonly used metrics in semantic segmentation. The Jaccard Index between two sets is the size of the intersection of the sets divided by the size of their union, as defined in the following formula:

$$J\left(X,Y\right)=\frac{\mid X\cap Y\mid}{\mid X\cup Y\mid}, \tag{1}$$

where X is the set of binary map pixels manually segmented by the specialist (the ground truth), and Y is the set of binary map pixels obtained as a result of applying the threshold algorithm (see Figure 11). If the two sets are disjoint, meaning they have

no common members, the Jacquard Index is 0, and when they are exactly identical, it is equal to 1.

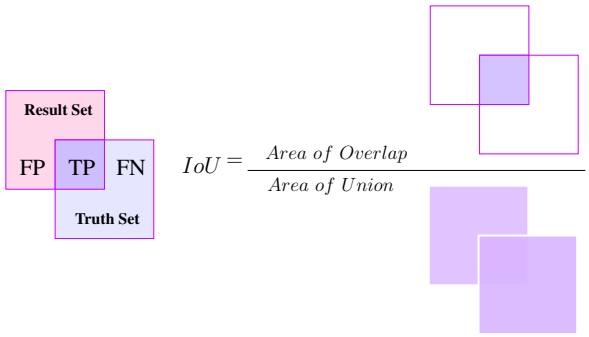
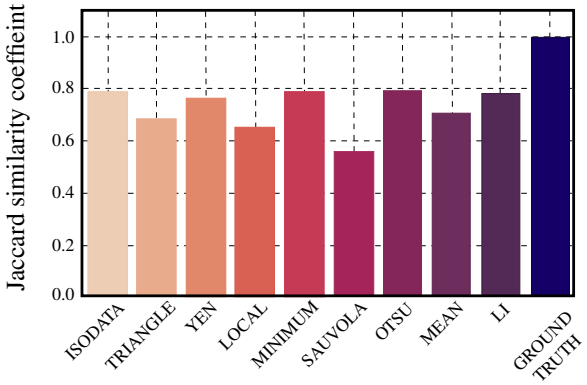


Figure 11: Intersection over Union.

After calculating the Jacquard indices according to the above formula for all images, they were averaged, and the results are shown as a bar graph in Figure 12. Based on the results listed in the table, the MINIMUM algorithm performed better and more accurately than the others, so we used this algorithm in the segmentation task. All operations performed to generate a binary map of an original image using the intensity thresholding approach are shown as a pipeline in Figure 14. As shown in the figure, the pipeline consists of four steps: deconvolution operation, color channel separation, image normalization, and binary mask generation.



No.	Algorithm	Jaccard Similarity Index			
		Micro	Macro	Weighted	Binary
1	**MINIMUM**	**0.7934**	**0.7516**	0.7994	**0.6735**
2	LI	0.7834	0.7482	0.7941	0.6728
3	MEAN	0.7899	0.7499	0.8014	0.6654
4	ISODATA	0.7894	0.7464	**0.8026**	0.6521
5	OTSU	0.7926	0.7448	0.8017	0.6493
6	TRIANGLE	0.6856	0.6538	0.6964	0.5934
7	YEN	0.7645	0.7045	0.7683	0.5843
8	LOCAL	0.6515	0.6193	0.6836	0.5161
9	SAUVOLA	0.5645	0.5506	0.5896	0.4855
	Ground Truth	1.00	1.00	1.00	1.00

Figure 12: Comparison of Jaccard index (Micro). The results of applying 9 thresholding algorithms to sample images were averaged and are displayed as a bar graph.

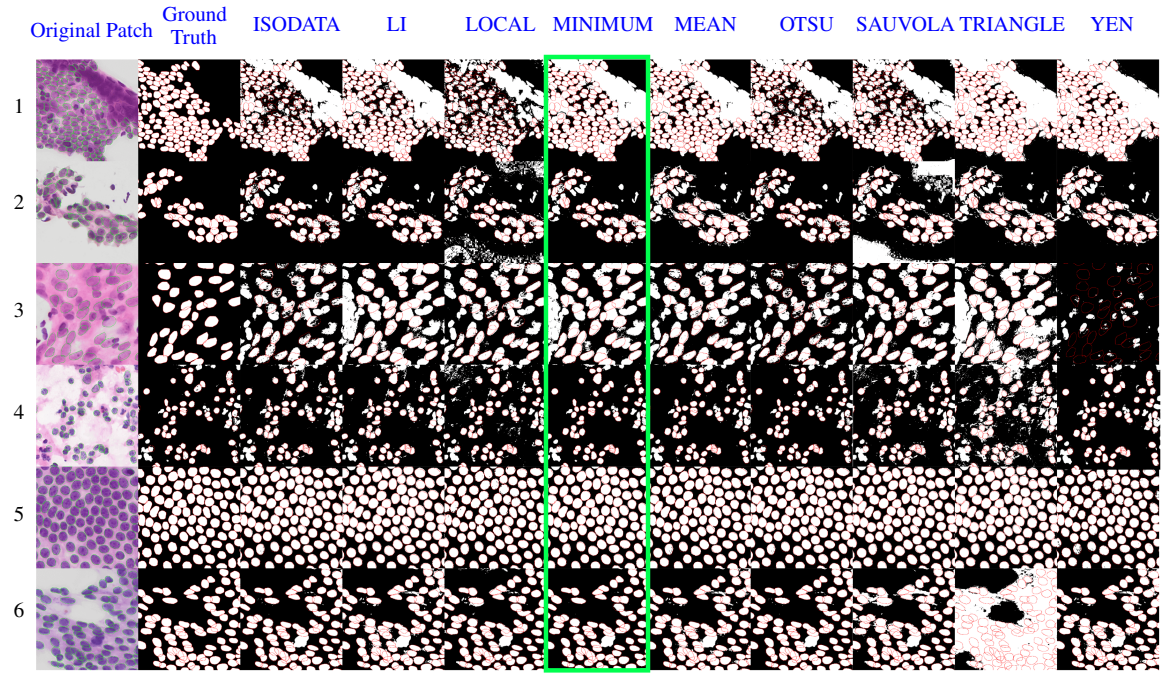


Figure 13: Thresholding techniques' comparison in segmentation of cytological images. Each column corresponds to a thresholding technique, and each row represents a sample of the original cytological image. The red contours are the boundaries of the ground truth that are manually drawn by a human expert. Of all the techniques, the MINIMUM algorithm has the best performance.

With the help of the obtained binary map, we identify the important parts of the original (RGB) image, and the image analysis is performed only on those areas. These RGB images, however, can not currently be fed as input data to the neural networks due to their large size and need to be divided into smaller pieces, which we will discuss in detail in the next section.

3.5 Image Patching

The analysis of cytological images to identify tissue characteristics is a tedious task due to their large size. These images obviously cannot be fed directly to the neural

network because processing such large images requires extensive memory resources and high-capacity processing units, and will also be very time-consuming. On the other hand, resizing and shrinking them can lead to the loss of useful information for diagnosis, so it does not seem to be a good solution to address this problem. A commonly used approach to deal with this issue is to segment large slides into smaller pieces called patches, so that the analysis can be performed separately on each small image patch.

The image patching task can be performed in such a way that the patches have overlap, or it can be done without overlapping by considering the step size equal to the width of each patch when segmenting the image. Partitioning the image into overlapping patches allows for more patches to be produced of each image, although many will eventually have the same texture. Non-overlapping patches, however, each represent a fragment of the large original image and will have a unique texture. In this study, in order to provide the appropriate data required for the pre-training phase of the model, we segmented the histopathological images of the BreakHis into overlapping patches so that we could produce a large number of images. Besides, for the training and evaluation phases, the cytological images of the target data set were segmented into non-overlapping square patches. Since image size may affect the performance of the system, we decided to use different sizes for patching to observe and evaluate this issue. Contrary to recent research (Miselis, *et al.* (2019) [51]) that has segmented images into patches with dimensions of 256 $\times$ 256 pixels, we have

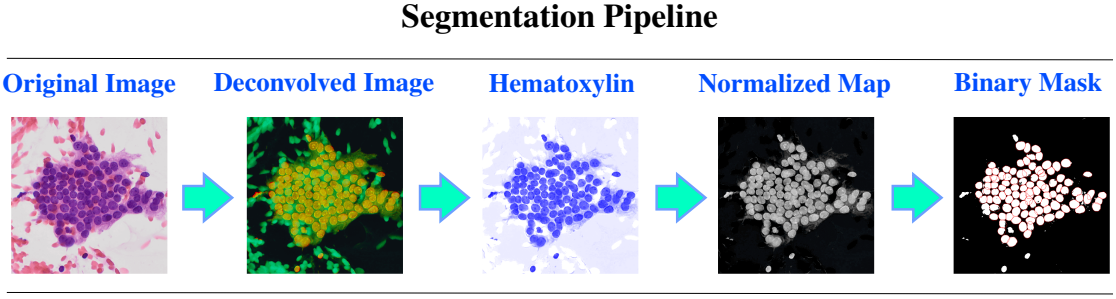


Figure 14: Segmentation pipeline using the intensity thresholding approach.

considered three different sizes of 64, 128, and 256 pixels for patching, and then we formed 3 separate bags of patches according to these sizes. This means that we performed the cancer malignancy diagnosis process 3 times, each time using one of these patch bags, and finally determined the optimal size for the patch by comparing the results. In order to be able to partition an image into several non-overlapping square patches with fixed dimensions (e.g. 256×256 px), it is obvious that the width and height of the image must be a multiple of that constant value, otherwise, the border patches that are close to the edge of the image will be smaller than the others. Failure to fulfill this condition will result in the system inputs, which are patches here, not being the same size, creating another challenge as we later plan to feed the data to the network. To address this issue, and to have patches of the desired width and height, one way is to crop part of the image and diminish its size, which results in the loss of some pixels, and we may end up losing useful information. Alternatively, we can enlarge the image to a fixed size, which allows us to partition the image into equal-sized patches without any problems. There are generally two ways to enlarge a small image to a fixed size: zero padding and image scaling using interpolation. While each of these methods has its pros and cons, we decided to use zero padding to enlarge images. The advantage of zero padding over scaling is that there is a possibility of image pattern deformation in scaling, whereas there is no such risk in the zero-padding approach. However, we must note that as a result of zero padding, an area is attached to the image that contains entirely black pixels, so we will end up with a small number of black patches. The presence of these black patches in the data set not only does not help in diagnosing cancer malignancy, but they will mislead the model in this regard, as these patches do not belong to any type of cancer malignancy. To tackle this issue, we filter out unwanted patches through a mechanism, and the model training process is done only with patches containing useful information. The next section is where we explain in detail how to filter and

select the appropriate patches.

3.6 Patch Selection

The breast cancer cytological images examined in this study are mainly composed of regions that can be grouped into three general categories. The first group belongs to the areas where the components of the cell nuclei are located, which are the most important part of the image in the diagnosis of cancer malignancy. The second category is the areas where red blood cells and other background tissues are found. And the third group consists of regions that are made up entirely of black pixels, which have been added to the image as a result of enlarging the original image. Therefore, after partitioning the image into small-sized patches done in the previous step, we end up with the patches, only a few of which possess useful information, and the rest of which contain pixels whose analysis does not help our ultimate goal of image classification. Figure 15 illustrates an example of patches belonging to each of these 3 regions in cytological images. This prompted us to select the desired patches instead of using all the available patches, and to train the models only with those that contained useful diagnostic information. Doing so obviously requires us to set a criterion for eliminating unwanted patches. Considering that the determining factor in choosing a patch is the presence of sufficient cellular material in that patch, we calculated the percentage of nuclei pixels for each patch, and then by setting a threshold, we selected only those that contained more than the specified percentage of cellular material. Unlike previous research [51] that trained an SVM classifier to identify suitable patches, we calculated the ratio of nuclei pixels to total patch pixels using the corresponding binary map for each patch according to the following equation:

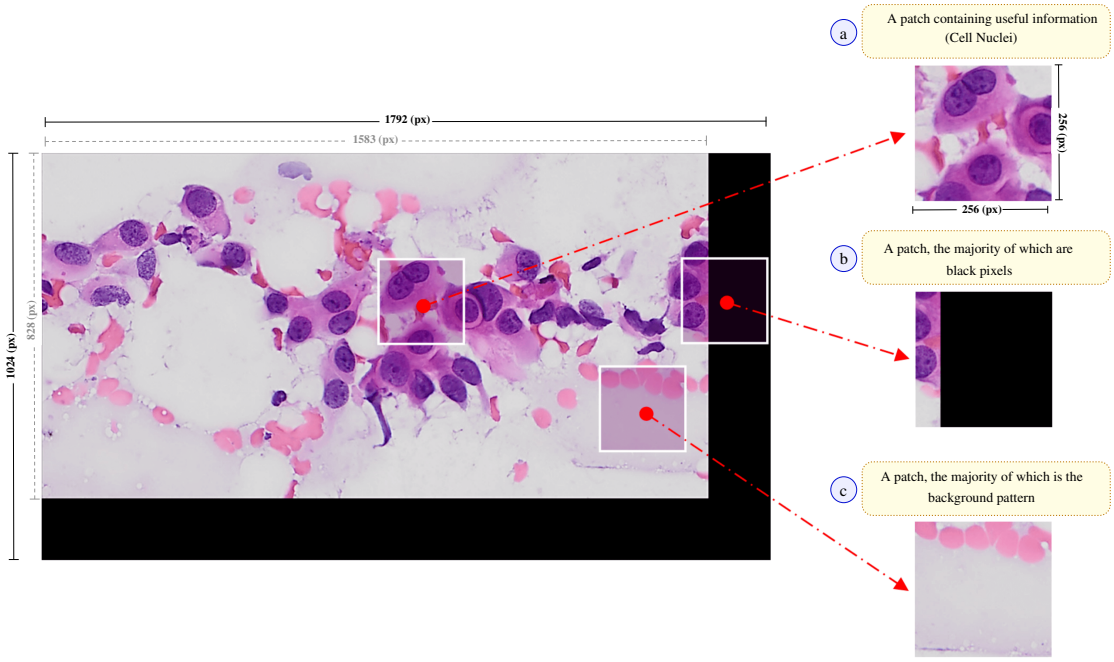


Figure 15: Patches belonging to different regions of the cytological image. a) Indicates a patch that contains a number of cell nuclei (needs to be analyzed). b) Displays the patch, which is mostly made up of black pixels. c) Represents the patch, which is mainly composed of background textures.

$$P = \frac{n_{nucleus}}{n_{all}} \times 100, \tag{2}$$

where $n_{nucleus}$ is the number of pixels for which the corresponding binary map value is 1 (the white pixels), and n_{all} is the total number of pixels that make up the patch.

The purpose of employing this method was to minimize system error as much as possible, since the use of the SVM classifier in previous research was associated with an error of about 8% in detecting cellular material for each patch. We then continued the process of selecting the desired patches by setting different thresholds so that only patches with cellular coverage equal to or above the specified thresholds are selected. Accordingly, we considered three different thresholds of 30%, 50%, and 70% and formed separate sets of patches with a percentage of cellular coverage in

accordance with the defined limits. A few examples of patches with different cellular coverages are shown in Figure 16. Although the higher the threshold, the larger cellular coverage the candidate patches will have, this will eventually result in fewer patches being selected, and obviously, a smaller data set will be formed. Thus, there seems to be a clear trade-off between the number of obtained patches and their cellular coverage. So, the reason for considering different thresholds is to find a set of patches that maximize system performance. Observing the fact that cell nuclei in benign and malignant cancer specimens are not uniformly distributed across the slides, the number of patches obtained after thresholding will not be the same in benign and malignant cases. Basically, the pattern of benign cytology specimens is structured in such a way that the cell nuclei are spread out in large numbers across the image, while the cell nucleus clusters in malignant specimens are dense and usually found in a small region of the image. This has led to the number of patches containing the cell nuclei belonging to these two sets are not exactly equal, and there is a need for further efforts to balance the data set.

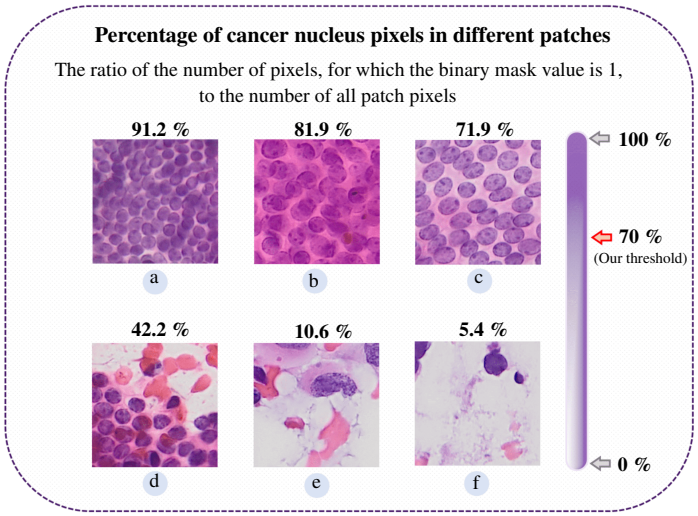


Figure 16: Six examples of patches with different cellular coverage. The top row patches (a, b, and c) have a percentage of cancerous pixels that exceed the threshold value. While the cellular coverage percentages of the bottom row patches (d, e, and f) are below the threshold value.

The next section describes the data set we formed from the selected patches, which are eventually fed to the models.

3.7 Building Dataset

In order to take advantage of the TL technique to address image analysis tasks, two separate phases must be performed to train the model, each of which requires a distinct data set. The first phase, known as the pre-training phase, is usually performed on a large data set and initializes the model to a point in the parameter space, which in a way makes the model optimization process more efficient (Erhan, *et al.* (2010) [22]). Then, in the next phase, the pre-trained model is fine-tuned on the target data set, which can potentially make significant progress by incrementally adapting the network features to the new data. Using a pre-trained network will generally work best when both tasks or data sets have something in common. Accordingly, in this research, we proposed the idea of using histopathological images for the pre-training phase of deep CNNs, and then used cytological biopsy data as the target data set for the binary classification task. Despite the gap between histopathological and cytological data in some respects, we claim that this approach can be much more effective in improving image classification accuracy than when employing natural image data for the pre-training phase of the model. In the rest of this section, each data set is described in detail separately.

3.7.1 Target Dataset

The target data set investigated in this research is digital cytology images of breast cancer, which is an archival collection of samples taken from patients at the Regional Hospital in Zielona Gora, Poland. As a result of the data acquisition process, a collection of 550 ROIs related to 50 patients was formed, in which 275 ROIs belonged

to benign patients and 275 ROIs to malignant cases.

3.7.2 Pre-train Dataset

Unlike many medical image analysis studies that use large-scale natural annotated images for the model pre-training phase, in this study we intend to use breast cancer histopathological images as a data source for this purpose. In addition to this approach, we also use CNNs already trained on the ImageNet data set to compare the two approaches to ultimately reveal the impact of using compatible domain data on system performance. Basically, the data set used to pre-train the model can be selected from related domains or from unrelated fields. Research has shown that the greater the dissimilarity and gap between the two data sets, the less effective the pre-training [5, 88]. Although there are clear differences between histopathological and cytological images, since the nature and structure of these two types of images are almost identical and both come from the similar domains, it is expected that the features extracted from them will be compatible with each other. In general, the analysis of cytological images is easier compared to histopathological specimens due to their special characteristics. For example, cell clusters are clearly separated in cytological images, and more complex structures such as glands are rarely seen in these types of images. Histopathological images, on the other hand, exhibit a more comprehensive view of the disease and its effect on tissues, as the process of their preparation is such that the structure of the underlying tissue is preserved. The samples of these types of images can be found in Figure 17. Thanks to the publicly available histopathological images, there is no problem in obtaining the images needed in our pre-training phase. In this research, we use one of the most popular collections of histopathological images of breast cancer, BreakHis [78], to pre-train the models.

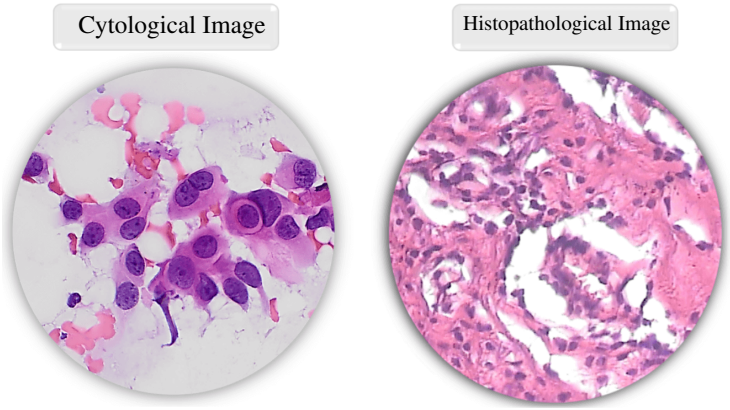


Figure 17: Sample of cytological and histopathological images. While cell nuclei in histopathological images usually do not have clear boundaries, cell clusters are clearly separated in cytological images.

BreakHis is a large-scale data set containing 7109 histopathological images of eight breast cancer subtypes acquired from 82 anonymous patients. The data set is divided into two main categories of benign tumors with 2480 samples and malignant cases with 5429 samples, each of which has four different magnification factors: 40X, 100X, 200X and 400X. Each category of benign and malignant breast tumors is classified into four different subgroups (i.e., a total of 8 subgroups) based on cellular aspects identified by pathologists under a microscope. All images have the same dimensions of 700 $\times$ 460 pixels and have three RGB channels. Figure 18 shows an example of images belonging to each subtype in BreakHis. We do not use all the images in

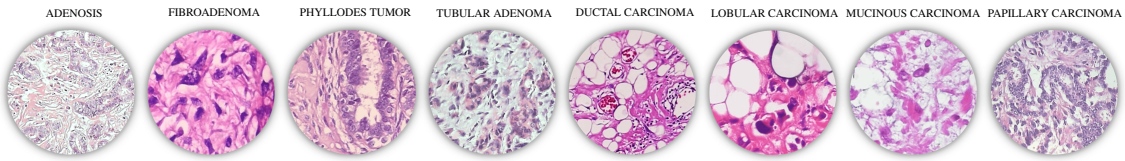


Figure 18: Sample images belong to 8 subgroups in BreakHis. From left to right, the first 4 images belong to the Adenosis (A), Fibroadenoma (F), Phyllodes Tumor (PT), and Tubular Adenoma (TA) subtypes, all of which come from the benign category. And the next 4 images extracted from the malignant category belong to the Ductal Carcinoma (DC), Lobular Carcinoma (LC), Mucinous Carcinoma (MC), and Papillary Carcinoma (PC) subclasses, respectively.

BreakHis for the pre-training phase, but selectively pull out only those that are as structurally similar as possible to cytological images.

Up to this point, all the necessary preprocessing operations have been performed on the images, and they are now ready to be fed to deep CNNs after being partitioned into smaller patches and filtered based on their cellular coverage. In this stage, we intend to build a bank of verified patches to provide the input data of the models. Accordingly, we first create a large bank of patches generated from BreakHis histopathology images to perform the pre-training phase. As previously described, the BreakHis data set includes images of eight different subtypes of breast cancer collected as benign and malignant specimens. Careful observation of the images in the data set revealed that the structure of the images belonging to the subtypes of Fibroadenoma and Lobular Carcinoma is the most similar to the texture of the target cytological images in comparison with the images of other subclasses. The first subtype (Fibroadenoma) is in the benign category, while the second (Lobular Carcinoma) belongs to the malignant group. Therefore, we selected 100 images from each of these two specific subtypes. This actually helps to make the features extracted by the network compatible with those to be obtained from the target data images. After partitioning the images into overlapping patches with dimensions of 64×64, 128×128, and 256×256 we randomly selected 150 patches from each image and ended up with three banks of 30,000 patches. The data set was formed in a balanced way so that there are equal numbers of benign and malignant patches in it (15,000). Since the patches obtained from histopathological images contain an acceptable percentage of cellular material, we can use all the patches generated, and performing the patch selection process is not required in this phase.

To address the problem of classifying cytological images into benign and malignant categories, we now need to form banks of image patches of the target data set. As discussed earlier, the target data set includes cytological images of breast cancer

Patch Size (pixel)	Traning		Validation		Test	
	Benign	Malignant	Benign	Malignant	Benign	Malignant
64×64	53625	53625	17875	17875	17875	17875
128×128	15015	15015	5005	5005	5005	5005
256×256	4620	4620	1540	1540	1540	1540

Table 2: The number of patches in benign and malignant sets after partitioning cytological images into square patches with sizes of 64×64, 128×128, and 256×256 pixels.

taken from 50 patients with an equal number of benign and malignant cases (25). Besides, for each patient, there are 11 images available in the data set. In order to evaluate the performance of the models, we split the images of 25 benign cancer patients into 3 sets of training, validation, and test in such a way that the images of 15 patients were placed in the training set (60%), and validation and test sets each contained those of 5 patients (20%). We then repeated this process for images of 25 malignant patients. It should be noted that we have formed the data set in such a way that all images coming from a particular patient can either belong to the training, validation, or test set, and there is no overlap or commonality between these sets. Then we partitioned the images into non-overlapping patches with dimensions of 64×64 pixels and repeated this process with patches of 128×128 and 256×256 pixels. The details of the number of patches created are given in Table 2.

Cellular Coverage	Patch Size (pixel)	Traning		Validation		Test	
		Benign	Malignant	Benign	Malignant	Benign	Malignant
30 %	64×64	14046	9827	4777	2754	4901	1996
	128×128	3525	2372	1196	657	1206	464
	256×256	888	591	305	153	291	91
50 %	64×64	11294	6272	4024	1766	3812	1199
	128×128	2675	1247	948	358	890	201
	256×256	616	214	205	60	202	22
70 %	64×64	7037	3115	2893	916	2063	576
	128×128	1449	400	646	115	419	55
	256×256	264	28	122	6	82	5

Table 3: The 9 data sets used to train deep CNNs. The square patches with 30%, 50% and 70% cellular coverage with different sizes of 64, 128, 256 pixels have been used to form the data sets.

Now, before feeding these patches to the model, the last step is to select the desired patches according to the specified thresholds (30%, 50%, and 70%), which will result in the formation of 9 separate sets of patches. You can see the details of the assembled sets in Table 3. As can be clearly seen from the table, the number of patches obtained as a result of patch selection is not the same for benign and malignant sets. Furthermore, the number of patches available, especially for the malignant category, has been greatly reduced, which is certainly insufficient to train the model. This prompted us to use the image augmentation technique as a common solution to enlarge patch sets and reduce over-fitting problems. To do so, we applied different geometric transformations to the image patches including horizontal and vertical flipping, horizontal and vertical shifting, and random scaling (zoom)). Also, since microscopic images (e.g. cytological images) are rotationally invariant and can be analyzed from different angles, we also employed random rotation in the augmentation procedure. The image augmentation algorithm is shown in Algorithm 2.

Finally, to have balanced data sets, patch sampling was performed as the last step in preparing the input data. This was done by randomly removing a number of patches from sets that contained more than the required amount of patches. As a result of the augmentation and patch sampling operation, the number of patches in

Algorithm 2: Image Augmentation

Input: Image patches $\{\mathrm{P}\}_{i=1}^{N}$ where N is the number of patches and i^{th} input patch is denoted as P_i
Output: Augmented patches
Initialize: D_j random [0.8:1.2] and D_k random [0:180]
Function `ImgAug()`:
  for $i \leftarrow 1$ *to* N **do**
    $P_i^{Vflip} \leftarrow VFlip(\mathrm{P}_i)$;
    $P_i^{Hflip} \leftarrow HFlip(\mathrm{P}_i)$;
    $P_i^{Vshift} \leftarrow VShift(\mathrm{P}_i)$;
    $P_i^{Hshift} \leftarrow HShift(\mathrm{P}_i)$;
    $P_i^{Zoom} \leftarrow Zoom(\mathrm{P}_i, D_j)$;
    $P_i^{Rotation} \leftarrow Rotation(\mathrm{P}_i, D_k)$;
  return P_i^{Vflip}, P_i^{Hflip}, P_i^{Vshift}, P_i^{Hshift}, $P_i^{Rotation}$, P_i^{Zoom};

the training set eventually became 3000, and the validation and test sets each had 1000 patches, with the number of benign and malignant patches being equal.

3.8 Deep Learning Models

Deep learning techniques have been widely used by researchers as a powerful tool in a wide range of imaging domains - such as classification [2, 10, 84], detection [19, 29, 53], segmentation [33, 87], etc. The extensive applications of DL in various fields have accelerated its development and progress. One of the most popular types of DL models is known as convolutional neural network (LeCun, *et al.* (1995) [47]), which has a different structure compared to a regular neural network. CNNs were primarily developed to process image data, and are perhaps the most flexible type of DL models for image classification problems. Given the capability of deep CNNs to learn generalizable descriptors directly from images, they seem to be the ideal solution to most pattern recognition problems. While CNNs were further developed in the Computer Vision (CV) community, they have rapidly expanded into medical imaging applications and have been introduced as a powerful tool that can assist scientists in this field. The application of CNNs, however, is not limited to image analysis tasks, and they have the potential to achieve SOA accuracy in a variety of challenges such as Natural Language Processing (NLP) (Young, *et al.* (2018) [94]), Human Activity Recognition (HAR) (Jiang, *et al.* (2015) [38]), speech recognition (Abdel-Hamid, *et al.* (2014) [1]), etc, sometimes exceeding human-level performance. The basic CNN architecture is typically composed of three types of layers (or building blocks): convolution, pooling, and FC layers. The idea of the convolution layers is to perform feature extraction. This is basically done by convolving the image with a kernel designed to extract specific features. Pooling layers, on the other hand, form another CNN building block whose task is to reduce the spatial size of feature maps,

thereby reducing the number of learning parameters and the number of computations performed on the network. Lastly, FC layers, which are fully integrated with all activations in the previous layer, map the extracted features into the final output. This layered structure, consisting of several independent components, enables CNNs to extract high-level abstract features as well as learn hierarchical levels of representations from a low-level input vector. Given all the advantages cited for CNNs in image analysis tasks, in this study we examine 6 SOA deep CNNs that have already proven their capability in the ImageNet challenge. With the help of these networks, we solve the problem of classifying cytological images and by analytical comparison of the obtained results, we ultimately introduce the most efficient network. All the deep models used in this study were implemented in Python with Keras and Tensorflow libraries, the architecture of each of which is described in separate sections below.

3.8.1 VGG-16

VGG-16 is a CNN architecture proposed by Simonyan, *et al.* (2014) [77], which was used to win ILSVRC [67] competition in 2014. VGG-16 made the improvement over AlexNet, another famous model submitted to ILSVRC-2012, by replacing the relatively large 11×11 and 5×5 filters with a stack of 3×3 filters with a stride 1. The use of small-size filters had the advantage of low computational complexity by reducing the number of parameters. Simonyan and Zisserman also suggested enhancing the model capacity by deepening the network from 8 layers, previously used in AlexNet, to 16-19 layers, which greatly helped improve model performance. These modifications and presenting a novel deep CNN architecture resulted in a significant increase in the top-5 test accuracy of the model and ranked VGG-16 first in localization and second in classification. 16 in VGG-16 represents the sixteen layers in it, which makes it a relatively large network with about 138 million parameters. Due to the depth and large number of FC nodes, the size of VGG-16 has reached more

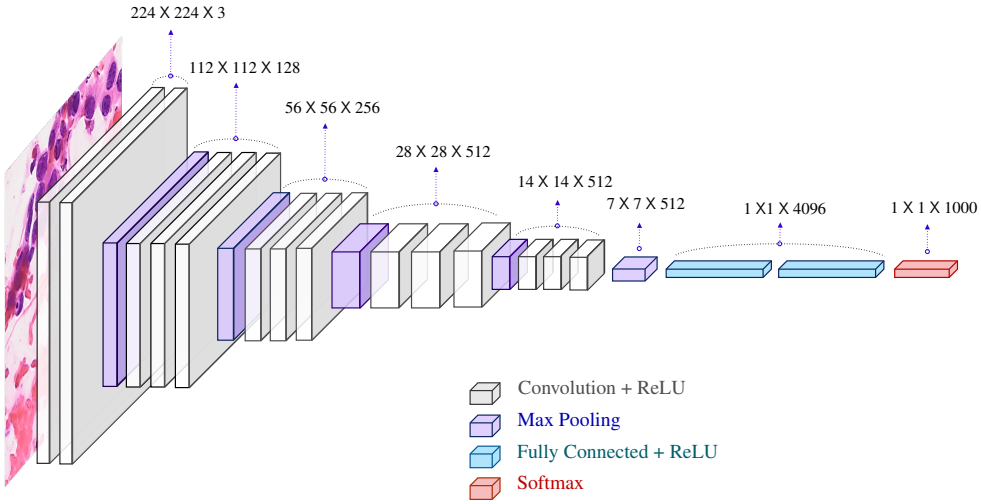


Figure 19: VGG-16 Architecture

than 500 MB, which has made the deployment of the model a somewhat tedious task. Having too many parameters in the model can also increase the risk of overfitting, as well as makes it difficult to pass gradient updates through the entire network. So, in addition to this model, we decided to use smaller and lighter networks in terms of the number of parameters in our experiments so that we can compare the capacity and performance of different models. The architecture of VGG-16 is pretty simple to understand and explain (see Figure 19). Despite the simplicity of the architecture, its performance is such convincing that it is still used as a baseline in various CV tasks [28, 39, 97].

3.8.2 VGG-19

VGG-19, a variant of VGG networks, is a deep CNN architecture, having been trained in the ImageNet challenge (ILSVRC) 1000-class classification task. The architecture of VGG-19 is pretty much similar to that of VGG-16 except that there are three additional convolution layers distributed in the last three blocks of the VGG-19. This brings the total number of network layers to 19, sixteen convolution layers having

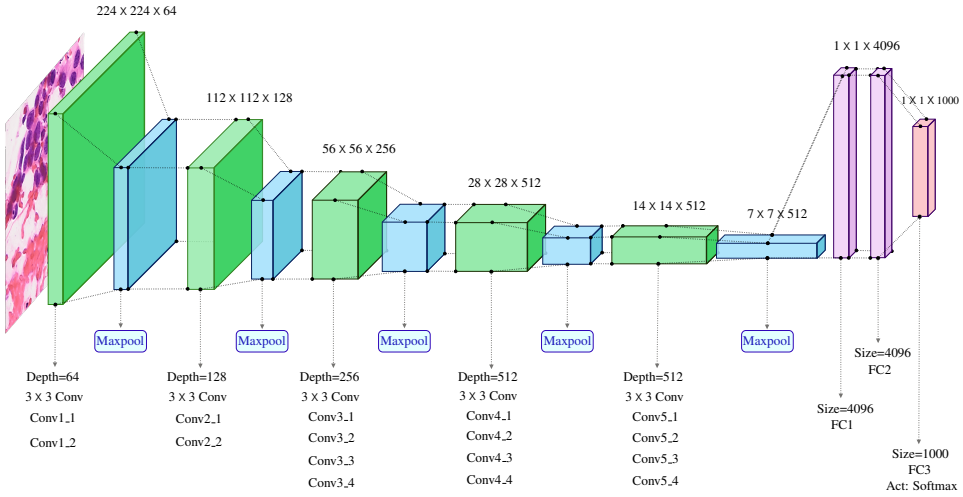


Figure 20: VGG-19 Architecture.

trainable weights followed by three FC layers. The VGG-19 network is used as one of the deep networks for the pre-training stage in this research so that we can investigate the effect of increasing the number of convolution layers on the model performance in the classification task. The VGG-19 architecture is as shown in Figure 20.

3.8.3 ResNet101-V2

The Residual Network (known as ResNet) is a special type of neural network developed by He, *et al.* (2016) [32] that ranked first in the 2015 ILSVRC image recognition and segmentation challenges. The ResNet architecture has certainly been one of the most pioneering works in the CV community in recent years, as it introduces the concept of identity shortcut connections (known as skip connections). The idea of using skip connections in ResNet was to avoid the problem of vanishing gradients in the network and to mitigate the accuracy saturation issues. In general, the deeper a network becomes (containing more layers), the greater its capacity to solve more complex tasks, and usually helps to improve the performance of classification and identification tasks. On the other hand, as we continue to add more layers to the neural network, it becomes very difficult to train, and the accuracy of the model

begins to saturate and then decreases. This is a problem that ResNet has greatly mitigated with residual blocks, allowing gradients to pass through an additional shortcut channel. As illustrated in Figure 21, the residual block has two 3×3 convolution layers, each followed by a batch normalization layer and a ReLU activation function. Besides, there is a direct connection that passes through the layer in between and connects the input x to the addition operator. The ResNet architecture is clearly inspired by the VGG-19, except that skip connections have been added to it.

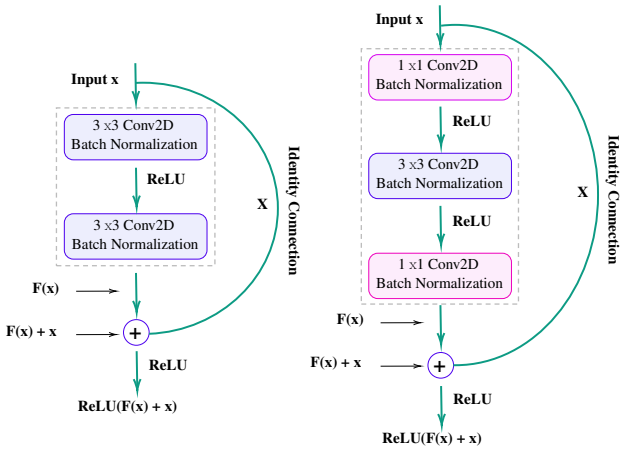


Figure 21: Identity mapping in Residual blocks.

There are a few variants of ResNet architecture, e.g., ResNet-34, ResNet-50, ResNet-101, ResNet-110, ResNet-152, ResNet-164 etc, the difference being essentially the number of layers used in them. The name of the network comes with a number that clearly indicates the number of layers used in that specific architecture. In this research, we use ResNet-101 as another deep model to solve the classification problem and compare its performance with other SOA networks. ResNet-101 is a 101-layer deep network consisting of 44.6 million parameters.

3.8.4 DenseNet-169

Dense convolutional network (DenseNet) developed by Huang, *et al.* (2017) [34] is another type of CNN, similar to ResNet, was proposed to solve the vanishing gradient problem. The DenseNet architecture has a narrow-layered structure that uses cross-layer connectivity, meaning that each layer receives additional inputs from all preceding layers and transmits its own feature maps to all subsequent layers (Khan, *et al.* (2020) [40]). This operation is performed using dense connections between layers through several dense blocks embedded in the network. A dense block itself consists of multiple convolution blocks, each using the same number of output channels. Having such a structure in the network provides direct access of each layer to the gradients through the loss function, and also helps to share the important features learned by each layer, which in turn boost information flow through the whole network. The described mechanism, in addition to strengthening feature propagation, significantly reduces the number of learning parameters across the network, thus making the network training process more efficient. DenseNet has been developed in several versions so far, including DenseNet-121, DeneNet-161, DeneNet-169 and DenseNet-201, of which we employ DenseNet-169 variant in this study. Despite having a depth of 169 layers, DenseNet-169 has relatively few parameters compared to other models and is still able to handle the vanishing gradient problem well.

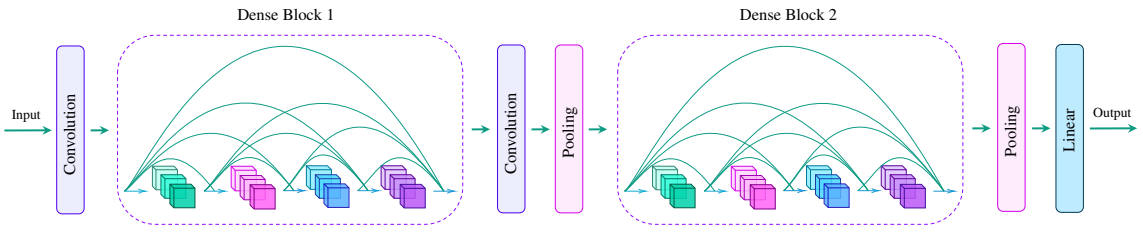


Figure 22: A block diagram of DenseNet-169 consisting of two dense blocks.

A block diagram of DenseNet-169 consisting of two dense blocks is presented in Figure 22.

3.8.5 Inception-V3

Inception networks currently have four versions, namely GoogleNet (also known as Inception-V1) (Szegedy, *et al.* (2015) [81]), Inception-V2, Inception-V3 (Szegedy, *et al.* (2016) [82]), and Inception-V4 (Szegedy, *et al.* (2017) [80]). The GoogleNet, which achieved SOA results for classification and detection in ILSVRC-2014, was the first network to introduce a key innovation called the Inception module. The main objective of the Inception networks was to achieve higher accuracy and at the same time reduce the computational cost, both in terms of the number of parameters and memory resources. In this regard, instead of naively stacking large convolution operations, which are obviously computationally expensive, Szegedy, *et al.* (2017) came up with the novel idea of using Inception modules. The Inception module is a block of parallel convolution layers that encapsulates filters of different sizes (e.g., 1×1, 3×3, 5×5) along with max-pooling layers to perform convolution operations on inputs. This, while helping to capture details at different scales, makes the network generally a bit wider than the previously presented networks. The output feature maps will ultimately be concatenated, and then connected to the next layer Inception modules. In this study, we use Inception-V3, an improved version of the previously introduced Inception networks, whose idea was to reduce the computational cost of the network without affecting generalization. The structure of the Inception module in this version is slightly different, and the large size filters (5×5 and 7×7) have been replaced with small and asymmetric filters (1×7 and 1×5). The illustration of a canonical Inception module is shown in Figure 23. Replacing larger convolutions with smaller ones reduces learning parameters and thus speeds up the network training

process.

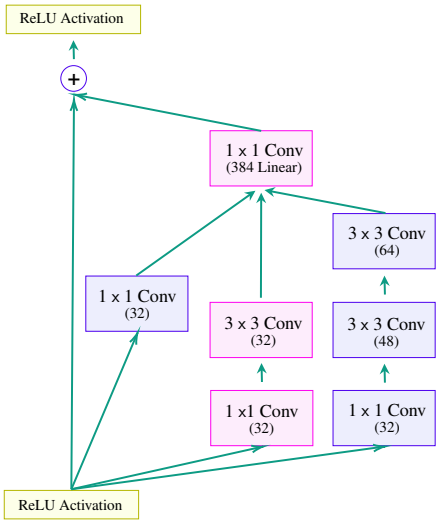


Figure 23: Demonstration of the inception module used in the Inception-V3 network.

3.8.6 InceptionResNet-V2

InceptionResNet is a convolutional neural architecture developed based on a combination of inception structure and residual connection [80]. Inspired by ResNet performance, InceptionResNet has a structure consisting of multiple-sized convolution filters combined with residual connections and employs a hybrid inception module in its architecture. The usage of residual connections not only avoids the degradation problem caused by deep structures, but also reduces the model training time. There are two sub-versions of InceptionResNet, namely V1 and V2. Both sub-versions have the same structure for the residual blocks and the only difference is in the hyper-parameter settings. In this study, we use InceptionResNet-V2, which is significantly deeper than the previous Inception-V3, to classify cytological images. This model also requires roughly twice the memory and computation compared to Inception-V3, and has the potential to achieve higher accuracy in early epochs.

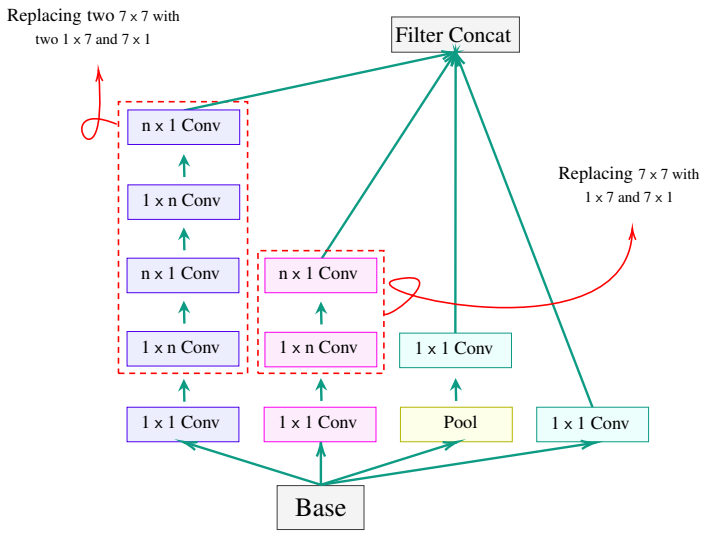


Figure 24: The scheme for the InceptionResNet module.

3.9 Summary

This chapter was dedicated to a comprehensive description of the methodology used in this thesis. We have addressed the problem of classifying breast cancer cytological images by proposing a new TL framework. The proposed TL framework consists of eight consecutive steps: cytological image acquisition, nuclei segmentation, image patching, patch selection, data set formation, network pre-training, weight initialization, and image classification. All these steps were discussed in detail in separate sections. To evaluate the effectiveness of this framework, real medical images captured from cytological biopsy specimens were employed. The process of preparing cytological images using the classical microscope and the virtual microscopy system was described in Section 3.3. Likewise, the task of classifying cytological images into benign and malignant categories was performed using six SOA deep CNNs, including VGG-16 Subsection 3.8.1, VGG-19 Subsection 3.8.2, ResNet101-V2 Subsection 3.8.3, DenseNet-169 Subsection 3.8.4, Inception-V3 Subsection 3.8.5, and InceptionResNet-V2 Subsection 3.8.6. In the next chapter, we present the experimental results of exploring different scenarios for training as well as fine-tuning the networks, and we determine the best strategy by analyzing the results.

Chapter 4

Experimental Results

4.1 Introduction

This chapter presents the experimental results of classifying cytological images of breast cancer and demonstrates the effectiveness of the proposed method in addressing the task. The experimental results are compared in terms of accuracy, as most existing studies conducted on the same data set have used this evaluation criterion, nonetheless, we provide a complete table that, in addition to accuracy, includes information on other criteria. As we explained earlier, in this study we examined 3 different scenarios for training as well as 2 different approaches for fine-tuning deep CNNs, and now we describe the results obtained from each scenario separately.

4.2 Results

4.2.1 Results for the First Scenario

The first scenario is designed to investigate the classification accuracy of deep CNNs when applying TL on a dataset of related medical images (i.e. BreakHis). The CNN architectures we examined are listed in Table 4 (sorted from low to high

based on the number of parameters) whose architectures have already been described in Section 3.8. As it turns out, the selected models include networks that are very different in terms of the number of parameters as well as the number of layers. The deepest networks employed are InceptionResNet-V2 and DenseNet-169, each with 449 and 338 layers, respectively. Despite the large number of layers, DenseNet-169 has the fewest learning parameters among networks, as its convolutions generate fewer output feature maps. Basically, the greater the number of network learning parameters, the greater the model's capacity to learn descriptive features from the input data. However, this certainly increases the risk of overfitting in the model. To evaluate this issue and have an alternative option, we attempted to use models with fewer parameters in our experiments, in addition to deep CNNs with a large number of parameters such as VGG-16 and VGG-19. Selecting DenseNet and Inception-V3 was inspired by the previous work (Miselis, *et al.* (2019) [51]) on the same data set, and the other 4 networks have not yet been used as classifiers on this data set. The pre-training phase of the models in this scenario was performed on a large bank of patches generated from BreakHis histopathology images. The large bank includes 30,000 patches with equal numbers of benign and malignant specimens (15,000), which are the result of partitioning 200 images into overlapping square patches. Since the main objective of this stage is to update the networks' parameters using a compatible data set (i.e. BreakHis), we did not consider any test set and decided to divide the patch bank into two parts: training and validation (with ratios of 90% and 10%). The models' training process was performed with 200 epochs, and we used the Adam (adaptive moment estimation) algorithm to perform the optimization. We also set the learning rate to 10^{-5} after trying a few different values, and considered the batch size of 32. At the end of the pre-training phase, the model weights are stored to be used for initializing the networks in the next phase.

Afterwards, in order to binary classify the images of the target data set and

Model	Parameters	Depth	Size
DenseNet-169	14.3 M	338	57 MB
Inception-V3	23.9 M	189	92 MB
ResNet101-V2	44.7 M	205	171 MB
InceptionResNet-V2	55.9 M	449	215 MB
VGG-16	138.4 M	16	528 MB
VGG-19	143.7 M	19	548 MB

Table 4: The SOA deep CNNs employed in this study. The networks are sorted from low to high based on the number of learning parameters.

estimate the performance of the pre-trained models, the cytological image patch bank was divided into three parts, namely training, including patches of 30 patients (60%), validation, and test sets, each of which accounts for 20% of all patches. For all 5 training strategies, the same splitting pattern was operated to form the data sets. Then the success rate of each model used in each scenario were compared. To observe the effect of patch cellular coverage on the models' performance in the classification task, we formed 3 separate patch banks with cellular coverage of 30%, 50% and 70%. In addition, we considered three different sizes of 64, 128, and 256 pixels to make patches and form data sets. Consequently, the models' training operation was performed using nine different data sets, as listed in Table 5.

The golden standard input size for the employed CNNs when training on the ImageNet data set is 224×224 pixels, although, in this study, we followed a similar

Dataset No.	Cellular Coverage	Patch Size
1	30 %	64×64
2		128×128
3		256×256
4	50 %	64×64
5		128×128
6		256×256
7	70 %	64×64
8		128×128
9		256×256

Table 5: The nine data sets used for training deep CNNs. The square patches with 30%, 50% and 70% cellular coverage with different sizes of 64, 128, 256 pixels have been used to form the data sets.

approach to that used in previous research, using square patches of 64, 128, and 256 pixels as input data. However, there is a restriction to the minimum input size for the networks, e.g., the minimum possible input size for Inception-V3 and InceptionResNet is 75×75 pixels. This made it unattainable for us to train these two networks using 64×64 pixel patches. On the other hand, training aforesaid models using patches of 75×75 pixels made the results incomparable with other networks, so we arranged to train these two CNNs with only 128 and 256 pixel patches.

The CNN experimental settings and parameters used for training are displayed in Table 6. With the exception of the learning rate, default values are selected for network parameters. The learning rate, which is a hyperparameter for controlling the rate at which the model weights are updated, is a factor of great importance in converging towards the minimum loss function. Specifying the learning rate usually requires an experimental process, so we tested 7 different values by selecting 7 numbers with equal logarithmic intervals from the range of $[10^{-7}, 10^{-1}]$, i.e. 10^{-7}, 10^{-6}, ..., 10^{-1}. As a result, the training process of each deep CNN (excluding Inception-V3 and InceptionResNet-V2) was repeated 63 times using 7 different learning rates on the 9 data sets listed in Table 5. Obviously, those networks that could not work with inputs of 64×64 pixels, namely Inception-V3 and InceptionResNet-V2, were run

CNN Training Options	
Parameter	**Value**
Loss function	Binary Cross Entropy
Optimizer	Adam
Learning rate	$[10^{-7}, 10^{-1}]$
$\beta 1$	0.900
$\beta 2$	0.999
Decay	-
Dropout	0.3
Max epochs	100
Mini batch size	32
Execution environment	GPU

Table 6: The experimental settings employed for training deep CNNs.

42 times using only 6 data sets. What is ultimately reported as the result of model classification is related to the best settings operated, which is a learning rate of 10^{-5}, 128×128 pixels patches with 50% cellular coverage. It should also be noted that, in order to make an acceptable comparison, the same hyperparameters were used for training all deep CNNs. The setup was run on two Tesla v100 GPUs in the virya cluster of Concordia University. The pre-training time for each of the deep CNNs was about 2.40 hr, which took a total of 43 hours for the entire pre-training process. Besides, the process of training the networks on the target data set took about 15 minutes for each model. Compared to the previous study [51] conducted on the same dataset whose database contained 51K image patches, in this study the number of patches in the target data set was reduced to 1/17 (3,000).

To evaluate the performance of the models more comprehensively, the classification results are evaluated by some benchmark metrics, including precision (or positive predictive value), recall (or sensitivity), F1-score (the harmonic mean of precision and sensitivity), false negative rate (FNR), and false positive rate (FPR), whose definitions are shown in Equations 3-8 below.

- **Accuracy**: Measures the model ability to identify the whole cases correctly, regardless the cases are being positive or negative, and can be formed as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

- **Precision**: Called positive predictive value (PPV). It shows the rate of true positives among all positive values. It is calculated as

$$Precission = \frac{TP}{TP + FP} \tag{4}$$

- **Recall**: Called the true positive rate (TPR) or sensitivity. It computes the

fraction of the relative positive cases and is calculated as

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

- **F1-score**: Harmonic mean of precision and sensitivity calculated as

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

- **False Positive Rate (FPR)**: Measures the error rate when the model classifies a patch as malignant and is calculated as

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

- **False Negative Rate (FNR)**: Measures the error rate when the model classifies a patch as benign and is calculated as

$$FNR = \frac{FN}{FN + TP} \tag{8}$$

The value of TP mentioned in the equations above represents the number of image patches correctly classified as malignant in the test set. Conversely, FP is the number of image patches in the test set that were incorrectly identified by the model as malignant. Similar parameters are defined for the second class, i.e. benign. Thus, TN is the number of patches that were correctly diagnosed as benign. And FN is the number of test set patches, the model has incorrectly classified as benign. Given these parameters, we conclude that Equation 4, i.e. precision, represents the ratio of malignant tumor patches correctly allocated to all malignant tumor patches recognized by the model. Recall in Equation 5 expresses the ratio of the correctly assigned malignant tumor patches to all malignant tumor patches present in the test

set. The F1-score in Equation 6 is the harmonic mean of the precision and recall, for which the highest possible value is 1.0, indicating perfect precision and recall, and the lowest possible value is 0. Besides, FPR in Equation 7 refers to the error rate when the model recognizes a patch as a malignant tumor, and FNR in Equation 8 is the error rate when the model identifies a patch as a benign tumor.

Table 7 shows the complete results obtained by the first scenario. The results for each model are disaggregated based on the data set used, as well as how a model is fine-tuned. The difference between the model fine-tuning approaches is in the number of network layers that are updated. In fact, in partial fine-tuning, only the weights of the last few layers are updated, while in complete fine-tuning, the backpropagation operation is applied to all network layers. In general, the early layers of a CNN learn low-level image features, which are applicable to most CV tasks, but the late layers learn high-level features, which are specific to the application at hand. Therefore, fine-tuning the last few layers is usually sufficient for TL. However, if the distance between the source and target applications is significant, one may need to fine-tune the early layers as well. Hence, an effective fine-tuning approach is to start from the last layer and then incrementally include more layers in the update process until the desired performance is reached. Based on the results shown in Table 7, the difference between the accuracies obtained as a result of complete and partial fine-tuning is quite obvious, and all networks performed better when applying complete fine-tuning. However, the accuracy obtained as a result of using different data sets (with 30%, 50%, and 70% cellular coverage) are not significantly different, and for most networks, except VGG-16 and Inception-V3, using the 50%-data set has yielded slightly better performance. Furthermore, the best performing model in this scenario was the DenseNet-169, which achieved 98.73% accuracy for validation data set and 94.55% accuracy for test data set using the complete fine-tuning approach.

Classifier	Fine-tuning Approach	Cellular Coverage	Class type	Precision	Recall	F1-Score	Ave. Accuracy	False positive rate	False negative rate
DenseNet169	Complete Fine-tuning	30 %	B	0.979514	0.979958	0.972874	97.99%	2.00% ± 1.22%	2.00% ± 1.22%
			M	0.979961	0.979958	0.979959			
		50 %	B	0.985325	0.989474	0.987395	98.73%	1.47% ± 1.05%	1.05% ± 0.89%
			M	0.989429	0.985263	0.987342			
		70 %	B	0.969875	0.962782	0.969725	96.27%	3.72% ± 1.65%	3.72% ± 1.65%
			M	0.962133	0.962757	0.962445			
	Partial Fine-tuning	30 %	B	0.926317	0.925527	0.925951	92.59%	7.36% ± 2.28%	7.44% ± 2.30%
			M	0.936255	0.926317	0.931259			
		50 %	B	0.922242	0.922105	0.923054	93.29%	7.61% ± 2.32%	7.78% ± 2.34%
			M	0.945263	0.923868	0.934443			
		70 %	B	0.915485	0.920177	0.919522	92.01%	7.98% ± 2.37%	7.99% ± 2.37%
			M	0.920065	0.920168	0.920116			
VGG-16	Complete Fine-tuning	30 %	B	0.937123	0.935489	0.936357	93.63%	6.28% ± 2.12%	6.45% ± 2.15%
			M	0.961488	0.937123	0.949149			
		50 %	B	0.987872	0.943158	0.964997	96.57%	1.15% ± 0.93%	5.68% ± 2.02%
			M	0.945619	0.988421	0.966547			
		70 %	B	0.936687	0.936574	0.937529	93.74%	6.16% ± 2.10%	6.34% ± 2.13%
			M	0.965754	0.938372	0.951866			
	Partial Fine-tuning	30 %	B	0.918871	0.918764	0.919480	91.94%	7.99% ± 2.37%	8.12% ± 2.39%
			M	0.935362	0.920090	0.927663			
		50 %	B	0.925778	0.925648	0.926590	92.65%	7.25% ± 2.27%	7.43% ± 2.29%
			M	0.949862	0.927405	0.938499			
		70 %	B	0.934259	0.934224	0.934514	93.44%	6.52% ± 2.16%	6.57% ± 2.17%
			M	0.942581	0.934769	0.938658			
VGG-19	Complete Fine-tuning	30 %	B	0.962334	0.962332	0.962368	96.23%	3.75% ± 1.66%	3.76% ± 1.66%
			M	0.964217	0.962402	0.963309			
		50 %	B	0.989086	0.982105	0.985583	98.56%	1.16% ± 0.84%	1.08% ± 0.90%
			M	0.982105	0.989164	0.986305			
		70 %	B	0.966058	0.966057	0.966080	96.61%	3.38% ± 1.58%	3.39% ± 1.58%
			M	0.967402	0.966102	0.966751			
	Partial Fine-tuning	30 %	B	0.921292	0.921184	0.921928	92.18%	7.74% ± 2.34%	7.88% ± 2.36%
			M	0.939026	0.922565	0.930723			
		50 %	B	0.920647	0.921178	0.924275	92.40%	7.26% ± 2.27%	7.93% ± 2.36%
			M	0.945263	0.927394	0.934443			
		70 %	B	0.920396	0.920387	0.920451	92.04%	7.94% ± 2.37%	7.96% ± 2.37%
			M	0.921896	0.920506	0.921200			
InceptionResNet-V2	Complete Fine-tuning	30 %	B	0.981742	0.981763	0.981194	98.12%	1.93% ± 1.20%	1.82% ± 1.17%
			M	0.962293	0.980647	0.975377			
		50 %	B	0.976215	0.993684	0.984872	98.47%	2.42% ± 1.34%	0.63% ± 0.69%
			M	0.993569	0.975789	0.984599			
		70 %	B	0.983962	0.983982	0.983349	98.33%	1.72% ± 1.14%	1.60% ± 1.10%
			M	0.983982	0.982737	0.983359			
	Partial Fine-tuning	30 %	B	0.940008	0.939974	0.940282	94.02%	5.94% ± 2.08%	6.00% ± 2.08%
			M	0.949781	0.940556	0.945146			
		50 %	B	0.940059	0.940000	0.940526	94.04%	5.90% ± 2.06%	6.00% ± 2.08%
			M	0.956842	0.940993	0.948851			
		70 %	B	0.932638	0.932592	0.932961	93.29%	6.67% ± 2.18%	6.74% ± 2.19%
			M	0.942980	0.933285	0.938107			
Inception-V3	Complete Fine-tuning	30 %	B	0.979923	0.979927	0.979848	97.98%	2.02% ± 1.23%	2.00% ± 1.22%
			M	0.972411	0.979774	0.976079			
		50 %	B	0.979936	0.980000	0.978386	97.86%	2.31% ± 1.31%	2.00% ± 1.22%
			M	0.979937	0.976842	0.978387			
		70 %	B	0.970298	0.970291	0.970429	97.04%	2.94% ± 1.48%	2.97% ± 1.48%
			M	0.979458	0.970560	0.974989			
	Partial Fine-tuning	30 %	B	0.789875	0.789032	0.791456	79.10%	20.69% ± 3.55%	21.09% ± 3.57%
			M	0.808418	0.793044	0.800657			
		50 %	B	0.781057	0.780000	0.782940	78.24%	21.51% ± 3.60%	22.00% ± 3.63%
			M	0.802457	0.784832	0.793546			
		70 %	B	0.773544	0.772358	0.775566	77.49%	22.24% ± 3.67%	22.76% ± 3.64%
			M	0.795925	0.777599	0.786655			
ResNet101-V2	Complete Fine-tuning	30 %	B	0.983049	0.976853	0.982049	97.85%	1.96% ± 1.21%	2.31% ± 1.31%
			M	0.979078	0.980346	0.980941			
		50 %	B	0.989293	0.972632	0.980892	98.10%	1.05% ± 0.89%	2.73% ± 1.43%
			M	0.973085	0.989474	0.981211			
		70 %	B	0.979455	0.979455	0.979471	97.94%	2.05% ± 1.24%	2.05% ± 1.24%
			M	0.981104	0.979488	0.980295			
	Partial Fine-tuning	30 %	B	0.880967	0.880293	0.909983	88.31%	11.40% ± 2.78%	11.97% ± 2.84%
			M	0.903017	0.885963	0.883458			
		50 %	B	0.893117	0.892632	0.895142	89.49%	10.28% ± 2.66%	10.73% ± 2.71%
			M	0.896842	0.897177	0.906580			
		70 %	B	0.883698	0.883091	0.886001	88.57%	11.16% ± 2.76%	11.69% ± 2.81%
			M	0.899886	0.888317	0.898626			

Table 7: Classification results obtained from the models pre-trained on the histopathological images (BreakHis). The presented results are related to the validation set containing patches of 128 × 128 pixels.

4.2.2 Results for the Second Scenario

We now turn to the results of the second scenario, which relates to utilizing TL using pre-trained networks on the ImageNet data set. Here we used networks that had already been trained on ImageNet and solved the cytological image classification using complete and partial fine-tuning approaches. The results of this scenario are summarized in Table 8. The table contains the results for using the 50%-data set containing 128 × 128 pixel patches. The accuracy obtained by using this data set was better than those got from other data sets, although it was negligible. As in the previous scenario, the difference between the accuracies is quite clear when using different fine-tuning approaches. The best performance was achieved by ResNet101-V2 using complete fine-tuning, with 95.70% and 91.50% accuracies on the validation and test sets, respectively. However, in partial fine-tuning mode, the best results were acquired by DenseNet-169 with 90.98% accuracy and Inception-V3 with 87.69% accuracy on the validation and test data sets, respectively. Comparing the results of the most efficient models in this scenario (i.e., ResNet101-V2 and DenseNet-169) with what was obtained by the DenseNet-169 and InceptionResNet-V2 networks in the previous scenario, demonstrates the effectiveness of employing a compatible database in classification performance.

	Classifier	Accuracy	FPR	FNR
Validation	**Complete Fine-tuning**			
	ResNet101-V2	**95.70%**	8.60% ± 2.45%	**0**
	DenseNet169	92.80%	14.40% ± 3.07%	0
	Inception-V3	92.30%	15.40% ± 3.16%	0
	VGG-16	90.70%	18.60% ± 3.41%	0
	VGG-19	88.40%	23.20% ± 3.69%	0
	InceptionResNet-V2	88.20%	21.60% ± 3.60%	2.00% ± 1.22%
	Partial Fine-tuning			
	DenseNet169	**90.98%**	16.60% ± 3.26%	1.60% ± 2.45%
	VGG-19	90.50%	19.00% ± 3.43%	0
	InceptionResNet-V2	88.88%	20.00% ± 3.50%	2.40% ± 1.34%
	ResNet101-V2	88.30%	21.00% ± 3.57%	2.40% ± 1.34%
	Inception-V3	87.90%	22.00% ± 3.63%	2.20% ± 1.28%
	VGG-16	85.60%	28.80% ± 3.96%	0

	Classifier	Accuracy	FPR	FNR
Test	**Complete Fine-tuning**			
	ResNet101-V2	**91.50%**	16.95% ± 3.28%	0.05% ± 0.19%
	Inception-V3	90.95%	17.80% ± 3.35%	0.30% ± 0.47%
	DenseNet169	90.57%	18.85% ± 3.42%	0
	InceptionResNet-V2	86.87%	24.70% ± 3.78%	1.55% ± 1.08%
	VGG-19	85.90%	28.20% ± 3.94%	0
	VGG-16	85.30%	29.40% ± 3.99%	0
	Partial Fine-tuning			
	Inception-V3	**87.69%**	22.92% ± 3.68%	1.69% ± 1.13%
	InceptionResNet-V2	87.07%	22.76% ± 3.67%	3.07% ± 1.51%
	VGG-19	86.57%	26.85% ± 3.88%	0
	DenseNet169	85.61%	27.85% ± 3.92%	0.92% ± 0.84%
	ResNet101-V2	83.46%	31.23% ± 4.06%	1.84% ± 1.17%
	VGG-16	82.88%	34.24% ± 4.15%	0

Table 8: Classification results of fine-tuning deep CNNs pre-trained on ImageNet. The presented results are related to the data set containing patches of 128 × 128 pixels having 50% cellular coverage. The results are sorted by the accuracy value.

According to the results, the use of histopathological images of breast cancer instead of natural images during the pre-training phase has improved the classification accuracy by 3.03% when applying complete fine-tuning and by 3.6% when using partial fine-tuning.

4.2.3 Results for the Third Scenario

The third scenario was devoted to training deep CNNs from scratch. The results showed that, as in the previous scenarios, there was no discernible difference in models' performance when using patches with different cellular coverage. Also, the accuracy obtained as a result of using patches of different sizes was almost the same. What we have presented in Table 9 is the results of models' training using data set No. 5, i.e. 128×128 patches with 50% cellular coverage, with which the best result was obtained. The experiments performed, and the results obtained, are based on the validation and test data sets. As can be seen in the results, the InceptionResNet-V2 outperformed other models, with an accuracy of 95.30% for the validation data set and 93.77% for the test data set. However, in terms of FPR, DenseNet-169 had the lowest error rates among networks.

We also provide ROC curves for the best models in training scenarios. Figure 25

Classifier	Accuracy	FPR	FNR
Validation			
InceptionResNet-V2	**95.30%**	6.80% ± 2.20%	2.60% ± 1.39%
DenseNet169	93.50%	4.40% ± 1.79%	8.60% ± 2.45%
VGG-16	92.50%	4.40% ± 1.79%	10.60% ± 2.69%
VGG-19	92.00%	9.00% ± 2.50%	7.00% ± 2.23%
Inception-V3	91.80%	11.60% ± 2.80%	4.80% ± 1.87%
ResNet101-V2	91.60%	8.20% ± 2.40%	8.60% ± 2.45%
Test			
InceptionResNet-V2	**93.77%**	10.60% ± 2.69%	1.85% ± 1.18%
DenseNet169	92.72%	7.90% ± 2.36%	6.65% ± 2.18%
VGG-16	91.80%	13.95% ± 3.03%	2.45% ± 1.35%
VGG-19	91.47%	12.00% ± 2.84%	5.05% ± 1.91%
ResNet101-V2	90.82%	13.00% ± 2.94%	5.35% ± 1.97%
Inception-V3	90.15%	16.90% ± 3.28%	2.80% ± 1.44%

Table 9: Classification results of training deep CNNs from scratch using cytological images of the target data set. The results are sorted by the accuracy value.

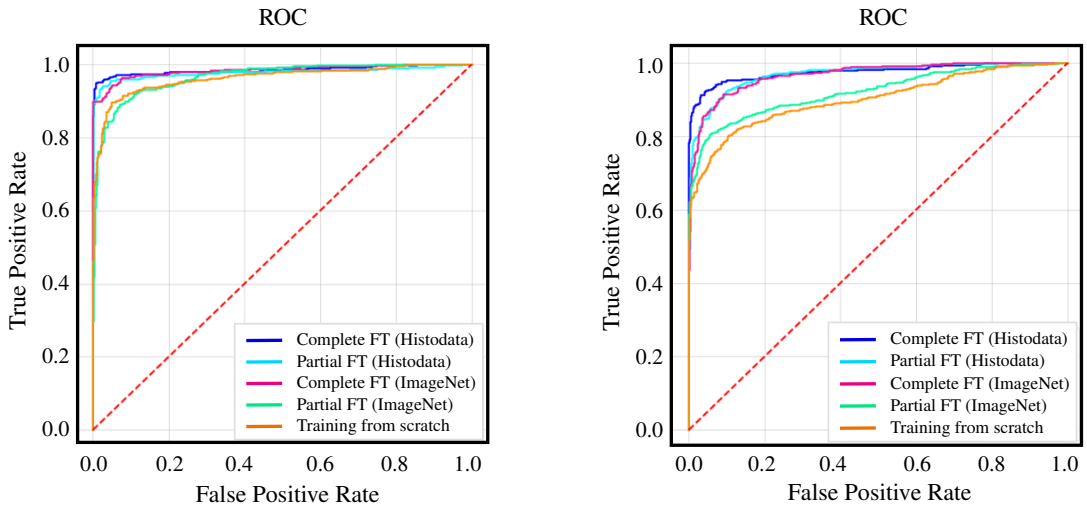


Figure 25: ROC results obtained from the validation (left) and test (right) sets.

Training approach	AUC (Validation)	AUC (Test)
Complete FT (Histodata)	$98.50\% \pm 0.75\%$	$97.68\% \pm 0.93\%$
Partial FT (Histodata)	$97.82\% \pm 0.91\%$	$97.04\% \pm 1.06\%$
Complete FT (ImageNet)	$98.44\% \pm 0.76\%$	$96.83\% \pm 1.08\%$
Partial FT (ImageNet)	$96.64\% \pm 1.12\%$	$92.27\% \pm 1.65\%$
Training from scratch	$96.33\% \pm 1.17\%$	$90.24\% \pm 1.84\%$

Table 10: AUC results

shows the ROCs obtained by the best models in each scenario using the validation and test data sets. For both data sets, the best result in terms of AUC was achieved by the model pre-trained on BreakHis and completely fine-tuned (see Table 10 for AUC results).

4.3 Discussion

Training deep CNNs from scratch to classify medical images is usually troublesome due to the small number of training samples. Most of the approaches published so far tackle this problem by using models pre-trained on the ImageNet data set and then fine-tuning the top layers of the model with target images. However, other works indicate that the low-level features learned on the ImageNet data set do not provide

optimal classification results if the target images are from a medical domain. In this work, we are stating a thesis that TL using a compatible data set will allow us to build a model for the classification of cytological images that will be more accurate than the similar model pre-trained on the ImageNet data set. Comparing the different approaches whose results are shown in Table 12, it can be seen that our method has clearly performed better than other benchmark approaches. This suggests that high-precision binary classification of cytological images is possible, even when limited annotated images are available, if the TL technique is employed with a compatible data set (e.g. BreakHis). This, in fact, accentuates the importance of an efficient weight initialization approach in achieving high system performance.

What can be discussed here is that applying partial fine-tuning, whether when using BreakHis or ImageNet data set, has not worked as well as the complete fine-tuning approach and has resulted in less accuracy. Considering ImageNet, complete fine-tuned models have achieved an accuracy range of 85-92%, while following the partial fine-tuning approach, the models' accuracies have been in the range of 83-88%. Similarly, according to the results, models pre-trained on BreakeHis and completely fine-tuned to the target data set achieved better accuracy than those that were partially fine-tuned. It can be argued that although the network's early layers have already discovered low-level image features that can potentially be applied to any task, updating the weights in such layers is necessary to achieve higher accuracy in our classification challenge.

It is also worth noting that among the fine-tuning approaches, the one based on ImageNet was the least efficient and even less accurate than when the model was trained from scratch on the target data set. The most efficient models in the classification task were those pre-trained on BreakHis and completely fine-tuned with the target data (93-95%), followed by models trained from scratch on the cytological images of breast cancer with accuracies in the range of 90-94%. These results indicate

that using the ImageNet for pre-training DL models, especially in medical applications, does not have to be the optimal approach. Therefore, in scenarios where the target data set contains a limited number of images and there is a risk of overfitting, the proposed solution is to pre-train the model on a larger set of images compatible with the target data. To compare the effectiveness of different deep neural network architectures in our image classification challenge, we should focus on the results obtained from complete fine-tuning on the BreakHis data set. They reveal no significant differences between network architectures because accuracies oscillated nearby 94% and only the VGG19 model was significantly worse, with an accuracy value of 92%. Aiming to identify the exact area in the cytological image that is most responsible for predicting cancer malignancy on deep CNNs, we used Gradient-weighted class activation mapping (Grad-CAM) [71, 98], a visual explanation algorithm, to visualize a class-specific heatmap based on the input image (See Figure 26). Heatmaps are usually obtained using the last convolution layer of the network. Here we used a VGG-16 network pre-trained on the BreakHis data set for this purpose.

In addition to utilizing the last convolution layer, we also generated heatmaps employing an intermediate convolution layer. Theoretically, the heat map for the last

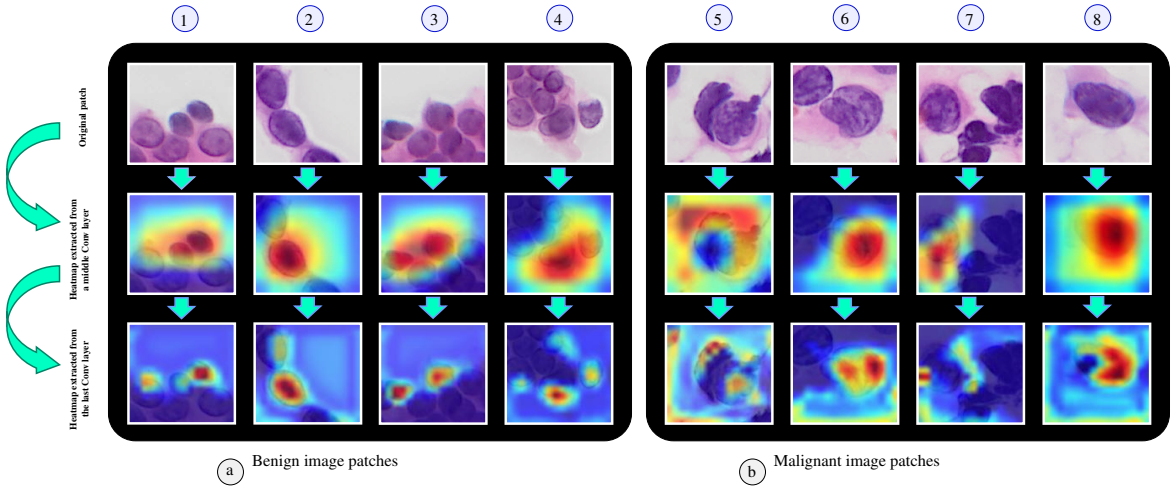


Figure 26: Heatmap visualizations using Grad-CAM. A VGG-16 neural network (pre-trained on BreakHis) was employed to generate the heatmaps.

layer should reveal the most accurate visual explanation of the object classified by the model. This is consistent with our results because network attention worsens at shallow layers, but the deeper layer captures more semantic concepts. We can see from the heatmap obtained for the last convolution layer that the network mainly focuses on cell nuclei features. This outcome is consistent with the medical knowledge in this field. Thanks to the TL using a compatible data set, the neural network extracted valuable knowledge about diagnosing breast cancer from a relatively small number of samples.

Article	Segmentation Method	Classification Method	Evaluation Approach	Best Results (ACC.)
Kowal *et al.* (2014) [44]	Multilevel Image Thresholding (Honey-Bees Mating Optimization (HBMO) Algorithm)	Naive Bayes, Decision Trees, SVM, KNN (based on morphometric features of cell nuclei)	Leave-one-out patient	77.64%
Kowal *et al.* (2016) [43]	Image Thresholding + Fast Marching (unsupervised method)	Naive Bayes (based on morphometric features of cell nuclei)	Leave-one-out patient	87.64%
Kowal *et al.* (2018) [46]	U-Net + Marker-Controlled Watershed	Naive Bayes, Decision Trees, SVM, KNN (based on morphometric features of cell nuclei)	K-fold CV	83.13%
Miselis *et al.* (2019) [51]	No Segmentation	Deep CNNs: AlexNet, GoogleNet, SqueezeNet, DenseNet-121, Inception-V3	Training-Validation	91.86%
Kowal *et al.* (2021) [45]	U-Net + Marker-Controlled Watershed	LDA, QDA, SVM, Naive Bayes, Random Forest, KNN, RPART	Hold-out, K-fold CV, Leave-one-out CV	88.20%
Proposed method	Image Thresholding Algorithms: ISODATA, LI, LOCAL, MEAN, MINIMUM, OTSU, SAUVOLA, TRIANGLE, YEN	Deep CNNs: DenseNet-169, ResNet-101, InceptionResNet-V2, Inception-V3, VGG-16, VGG-19	Training-Validation-Test	**Val.: 98.73%** **Test: 94.55%**

Table 11: Detailed information for ML/DL-based methods used in classification of breast cancer cytological images. The same data set has been used in all previous articles. The data set includes a total of 275 images of benign patients and 275 images of malignant ones.

Eventually, to demonstrate the superiority and effectiveness of the proposed method over other approaches presented so far applied to the same data set, we compared our results with those of five SOA studies in the literature and presented the results in Table 11. As noted, previous studies have used a variety of approaches to evaluate their models. In this research, as discussed earlier, we considered validation and test sets to estimate the performance of our models. The results of the validation set reveal an improvement of about 7% in terms of accuracy compared to the best result obtained previously, and for the test set, we had an improvement of approximately 3%.

Weight Initialization Approach	Classifier	Class type	Precision	Recall	F1-Score	Accuracy	FNR	FPR
Validation								
Random (Training from scratch)	ResNet101-V2	B	0.917671	0.914000	0.915832	91.60%	8.20% ± 2.40%	8.60% ± 2.45%
		M	0.914343	0.918000	0.916168			
	VGG-16	B	0.957872	0.913158	0.934997	93.57%	8.68% ± 1.79%	4.15% ± 1.26%
		M	0.915619	0.958421	0.936547			
	VGG-19	B	0.952105	0.952105	0.952105	95.21%	4.78% ± 1.35%	4.78% ± 1.35%
		M	0.952105	0.952105	0.952105			
	DenseNet169	B	0.954071	0.914000	0.933606	93.50%	4.40% ± 1.79%	8.60% ± 2.45%
		M	0.917466	0.956000	0.936337			
	Inception-V3	B	0.891386	0.952000	0.920696	91.80%	11.60% ± 2.80%	4.80% ± 1.87%
		M	0.948498	0.884000	0.915114			
	InceptionResNet-V2	B	0.934741	0.974000	0.953967	95.30%	6.80% ± 2.20%	2.60% ± 1.39%
		M	0.972860	0.932000	0.951992			
ImageNet	ResNet101-V2	B	0.920810	1	0.958773	95.70%	8.60% ± 2.45%	0.0
		M	1	0.914000	0.955068			
	VGG-16	B	0.843170	1	0.914913	90.70%	18.60% ± 3.41%	0.0
		M	1	0.814000	0.897464			
	VGG-19	B	0.811688	1	0.896057	88.40%	23.20% ± 3.69%	0.0
		M	1	0.768000	0.868778			
	DenseNet169	B	0.874126	1	0.932836	92.80%	14.40% ± 3.07%	0.0
		M	1	0.856000	0.922414			
	Inception-V3	B	0.866551	1	0.928505	92.30%	15.40% ± 3.16%	0.0
		M	1	0.846000	0.916576			
	InceptionResNet-V2	B	0.819398	0.980000	0.892532	88.20%	21.60% ± 3.60%	2.00% ± 1.22%
		M	0.975124	0.784000	0.86918			
HistoData (Proposed approach)	ResNet101-V2	B	0.989293	0.972632	0.980892	98.10%	1.05% ± 0.89%	2.73% ± 1.43%
		M	0.973085	0.989474	0.981211			
	VGG-16	B	0.987872	0.943158	0.964997	96.57%	1.15% ± 0.93%	5.68% ± 2.02%
		M	0.945619	0.988421	0.966547			
	VGG-19	B	0.989086	0.982105	0.985583	98.56%	1.16% ± 0.84%	1.80% ± 0.90%
		M	0.982105	0.989164	0.986305			
	DenseNet169	B	0.985325	0.989474	0.987395	98.73%	1.47% ± 1.05%	1.05% ± 0.89%
		M	0.989429	0.985263	0.987342			
	Inception-V3	B	0.979923	0.979927	0.979848	97.98%	2.02% ± 1.23%	2.00% ± 1.22%
		M	0.972411	0.979774	0.976079			
	InceptionResNet-V2	B	0.976215	0.993684	0.984872	98.47%	2.42% ± 1.34%	0.63% ± 0.69%
		M	0.993569	0.975789	0.984599			
Test								
Random (Training from scratch)	ResNet101-V2	B	0.879238	0.946500	0.9116300	90.82%	13.00% ± 2.94%	5.35% ± 1.97%
		M	0.942068	0.870000	0.904601			
	VGG-16	B	0.874888	0.975500	0.922459	91.80%	13.95% ± 3.03%	2.45% ± 1.35%
		M	0.972316	0.860500	0.912997			
	VGG-19	B	0.887798	0.949500	0.917613	91.47%	12.00% ± 2.84%	5.05% ± 1.91%
		M	0.945728	0.880000	0.911681			
	DenseNet169	B	0.921975	0.933500	0.927702	92.72%	7.90% ± 2.36%	6.65% ± 2.18%
		M	0.932658	0.921000	0.926792			
	Inception-V3	B	0.851884	0.972000	0.907987	90.15%	16.90% ± 3.28%	2.80% ± 1.44%
		M	0.967404	0.831000	0.894029			
	InceptionResNet-V2	B	0.902529	0.981500	0.940359	93.77%	10.60% ± 2.69%	1.85% ± 1.18%
		M	0.979726	0.894000	0.934902			
ImageNet	ResNet101-V2	B	0.855004	0.9995	0.921623	91.50%	16.95% ± 3.28%	0.05% ± 0.19%
		M	0.999398	0.8305	0.907155			
	VGG-16	B	0.772798	1	0.871840	85.30%	29.40% ± 3.99%	0
		M	1	0.706000	0.827667			
	VGG-19	B	0.780031	1	0.876424	85.90%	28.20% ± 3.94%	0
		M	1	0.718000	0.835856			
	DenseNet169	B	0.841397	1	0.913868	90.57%	18.85% ± 3.42%	0
		M	1	0.811500	0.895943			
	Inception-V3	B	0.848510	0.997000	0.916781	90.95%	17.80% ± 3.35%	0.30% ± 0.47%
		M	0.996363	0.822000	0.900821			
	InceptionResNet-V2	B	0.799432	0.984500	0.882366	86.87%	24.70% ± 3.78%	1.55% ± 1.08%
		M	0.979831	0.753000	0.851569			
HistoData (Proposed approach)	ResNet101-V2	B	0.907493	0.981000	0.942816	94.05%	10.00% ± 2.69%	1.90% ± 1.19%
		M	0.979325	0.90000	0.937989			
	VGG-16	B	0.912882	0.985000	0.947571	94.55%	9.40% ± 2.55%	1.50% ± 1.06%
		M	0.983713	0.906000	0.943259			
	VGG-19	B	0.934917	0.905000	0.919715	92.10%	6.30% ± 2.12%	9.50% ± 2.57%
		M	0.907946	0.937000	0.922244			
	DenseNet169	B	0.907594	0.992000	0.947922	94.55%	10.10% ± 2.64%	0.80% ± 0.78%
		M	0.991180	0.899000	0.942842			
	Inception-V3	B	0.897517	0.994000	0.943298	94.02%	11.35% ± 2.78%	0.60% ± 0.67%
		M	0.993277	0.886500	0.936856			
	InceptionResNet-V2	B	0.882197	0.996000	0.935651	93.15%	13.30% ± 2.97%	0.40% ± 0.55%
		M	0.995408	0.867000	0.926777			

Table 12: The best classification results obtained from different scenarios; 1) Complete fine-tuning of pre-trained models on BreakHis, 2) Complete fine-tuning of pre-trained models on ImageNet, and 3) Training models from scratch on the target data set.

4.4 Summary

This chapter presented the experimental results obtained from different scenarios explored in this thesis. In order to demonstrate the effectiveness of the proposed TL approach, we examined three different scenarios for training deep CNNs and applied two different approaches to fine-tuning the models. The first scenario was to classify cytological images using pre-trained networks on the BreakHis data set. The second scenario involved binary classification using pre-trained networks on the ImageNet data set. And the third scenario was to train deep CNNs from scratch. We also performed *complete* and *partial* fine-tuning approaches over the networks in each of the first and second scenarios, the former meaning the weights of all layers are updated, and the latter meaning that the backpropagation operation is performed only over the last few layers of the network. The results showed that the use of breast cancer histopathological images for the pre-training phase of the networks can significantly improve the classification accuracy of cytological biopsy specimens. It was also found that applying partial fine-tuning, either when using the BreakHis or ImageNet data sets, does not work as well as the complete fine-tuning approach, resulting in less accuracy.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this thesis, we proposed a new TL approach that could efficiently classify cytological biopsy specimens of breast cancer using an auxiliary data source comprising histopathological images. A significant advantage of the proposed method is its high accuracy in the classification task despite having a limited number of annotated images. The task of classifying cytological images was performed employing six SOA deep CNNs, and finally, by comparing the results from different aspects, we introduced the most efficient network in terms of accuracy. To evaluate the effectiveness of the proposed method, we explored three different training scenarios as well as two different approaches for fine-tuning networks. The experimental results revealed that the use of histopathological images can improve the system accuracy by more than 3% compared to when using natural images (e.g., ImageNet) for the pre-training phase of the models. In addition, it was uncovered that the proposed method is 7% superior to training deep CNNs from scratch in terms of the AUC value. The suggested method was finally compared with five SOA research previously conducted on the same data set, and the results showed that the classification accuracy was improved by 6% to

17% compared to studies using traditional ML methods. Also, compared to research that has utilized DL techniques, the proposed method has improved the classification accuracy by almost 7%.

5.2 Future Work

The contributions of this research have provided incentives to address other problems related to TL in medical image analysis. In future research, we intend to consider other TL approaches such as Multi-source Domain Adaptation (He, *et al.* (2021) [31]) to solve the breast cancer image classification challenge. This technique is primarily designed to minimize the impact of domain shift between the source and target domains, and potentially will allow us to employ our proposed method not only to solve the classification of breast cancer images but also to address other scenarios designed for different diseases. Designing and operating lightweight networks to perform the classification task is another idea that will be pursued in future research. This will help to assess the extent to which deep CNNs with a large number of parameters have actually contributed to achieving high performance, and whether such performance can be achieved with restricted computational resources. We also plan to explore image-level classification in future research, meaning that we will examine the model performance in diagnosing cancer malignancy in each patient's image. This gives a more comprehensive estimate of system capabilities than when only patch-level classification is considered. Lastly, we encourage future researchers to apply the proposed method for different image modalities to address classification challenges, as we believe that the compatible-domain TL technique is not limited to microscopic (histopathological or cytological) images and can be generalized to other types of medical images in different scenarios. We have also released our pre-trained models in a public GitHub repository so that it can be utilized by all researchers.

Appendix A

Confusion Matrices

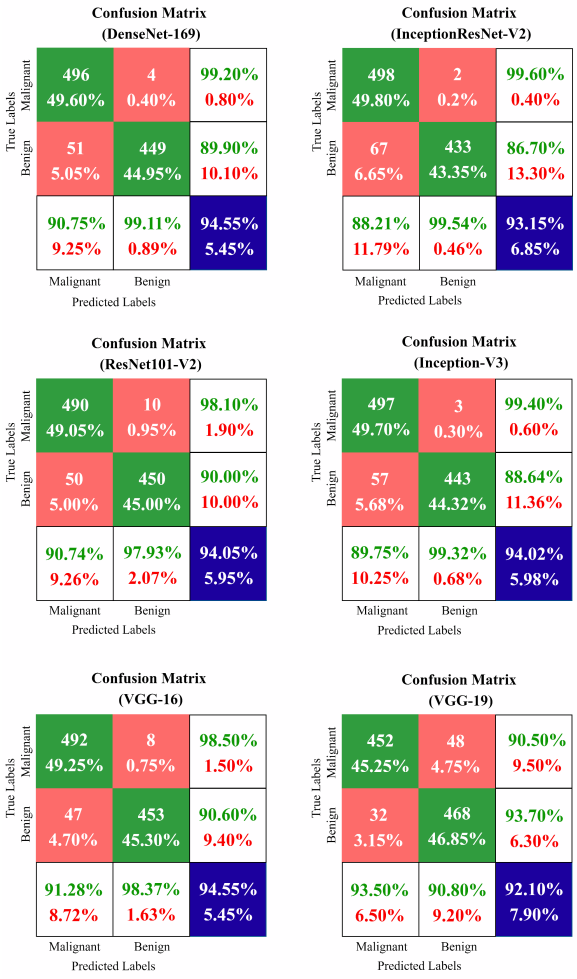


Figure A.1: Confusion matrices obtained from the first scenario (complete fine-tuning of pre-trained models on BreakHis).

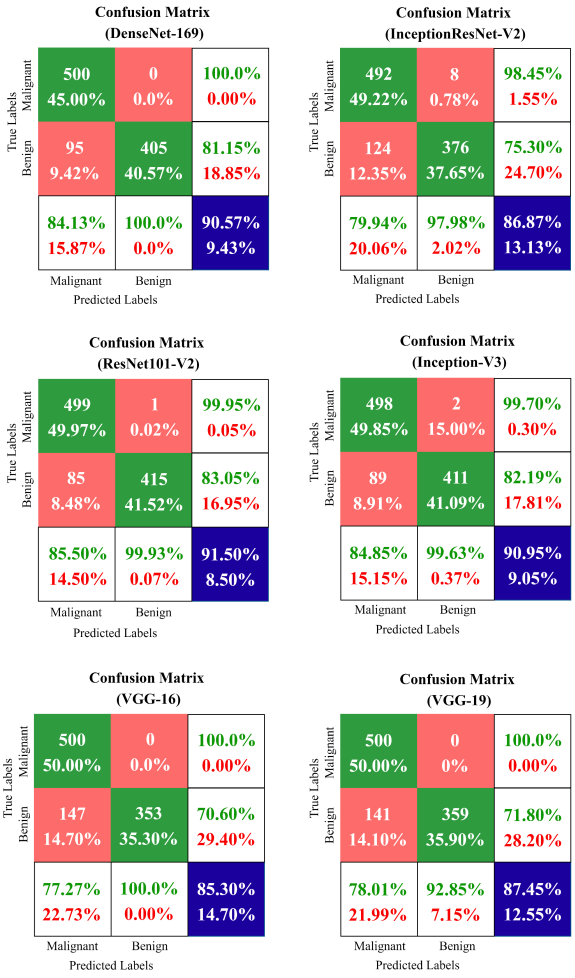


Figure A.2: Confusion matrices obtained from the second scenario (complete fine-tuning of pre-trained models on ImageNet).

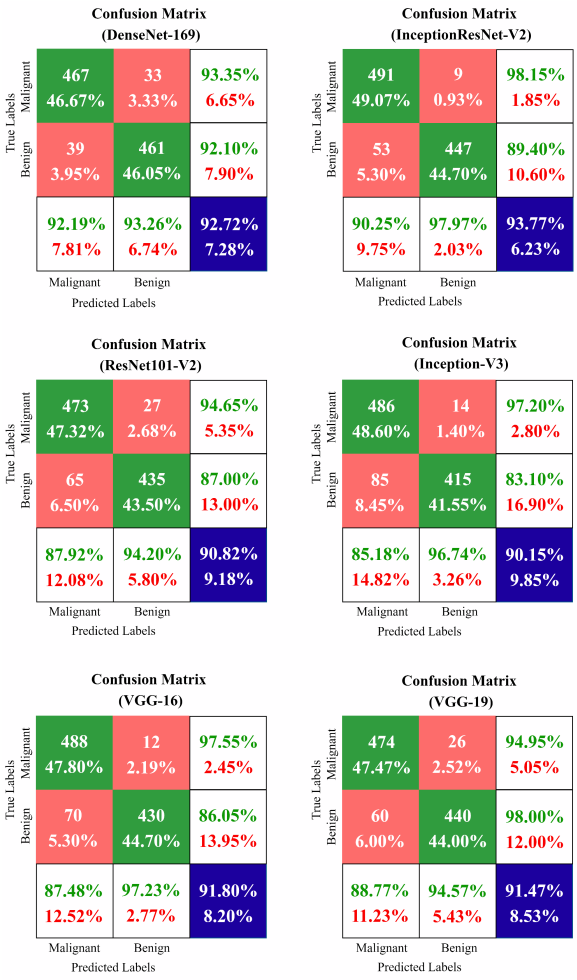


Figure A.3: Confusion matrices obtained from the third scenario (training deep CNNs from scratch).

Bibliography

[1] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1533–1545, 2014.

[2] C. Affonso, A. L. D. Rossi, F. H. A. Vieira, and A. C. P. de Leon Ferreira de Carvalho. Deep learning for biological image classification. *Expert Systems with Applications*, 85:114–122, 2017.

[3] A.V. Alvarenga, W.C.A. Pereira, A.F.C. Infantosi, and C.M. Azevedo. Complexity curve and grey level co-occurrence matrix in the texture evaluation of breast tumor on ultrasound images. *Medical Physics*, 34(2):379–387, 2007.

[4] L. Alzubaidi, O. Al-Shamma, M. Fadhel, L. Farhan, J. Zhang, and Y. Duan. Optimizing the performance of breast cancer classification by employing the same domain transfer learning from hybrid deep convolutional neural network model. *Electronics*, 9(3):445, 2020.

[5] L. Alzubaidi, M.A. Fadhel, O. Al-Shamma, J. Zhang, and Y. Duan. Deep learning models for classification of red blood cells in microscopy images to aid in sickle cell anemia diagnosis. *Electronics*, 9:427, 2020.

[6] Laith Alzubaidi, Muthana Al-Amidie, Ahmed Al-Asadi, Amjad J. Humaidi, Omran Al-Shamma, Mohammed A. Fadhel, Jinglan Zhang, J. Santamaría, and

Ye Duan. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers*, 13, Mar. 2021.

[7] Laith Alzubaidi, Mohammed A. Fadhel, Omran Al-Shamma, Jinglan Zhang, J. Santamaría, Ye Duan, and Sameer R. Oleiwi. Towards a better understanding of transfer learning for medical imaging: A case study. *Applied Sciences*, 10(13):4523, 2020.

[8] Breast Anatomy and Larger Version. What is breast cancer? https://www.breastcancer.org/symptoms/understand_bc/what_is_bc, 2018.

[9] J. Arevalo, F.A. González, R. Ramos-Pollán, J.L. Oliveira, G. Lopez, and M. Angel. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer Methods and Programs in Biomedicine*, 127:248–257, Aug. 2016.

[10] U. B. Baloglu, M. Talo, O. Yildirim, R. S. Tan, and U. R Acharya. Classification of myocardial infarction with multi-lead ecg signals and deep cnn. *Pattern Recognition Letters*, 122:23–30, 2019.

[11] E.P. Bankes. *Cervical Cancer Research Trends*. Nova Science Publishers, 2007.

[12] P. Bankhead, M. B. Loughrey, J. A. Fernández, Y. Dombrowski, D. G. McArt, P. D. Dunne, S. McQuaid, R. T. Gray, L. J. Murray, H. G. Coleman, J. A. James, M. Salto-Tellez, and P. W. Hamilton. Qupath: Open source software for digital pathology image analysis. *Scientific Reports*, 7:16878, 2017.

[13] B. Behboodi, H. Rasaee, K. Tehrani, and H. Rivaz. Deep classification of breast cancer in ultrasound images: more classes, better results with multi-task learning. In *Medical Imaging 2021: Ultrasonic Imaging and Tomography*, volume 11602, pages 170–175, Feb. 2021.

[14] T. Bel, M. Hermsen, B. Smeets, L. Hilbrands, J. Laak, and G. Litjens. Automatic segmentation of histopathological slides of renal tissue using deep learning. In *Medical Imaging 2018: Digital Pathology*, volume 10581, pages 285–290, 2018.

[15] N. Bouteldja, B. Klinkhammer, R. Bülow, P. Droste, S. Otten, S. Stillfried, J. Moellmann, S. Sheehan, R. Korstanje, S. Menzel, P. Bankhead, M. Mietsch, C. Drummer, M. Lehrke, R. Kramann, J. Floege, P. Boor, and D. Merhof. Deep learning-based segmentation and quantification in experimental kidney histopathology. *Journal of the American Society of Nephrology*, 32(1):52–68, 2021.

[16] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P.A. Heng. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. *IEEE Journal of Biomedical and Health Informatics*, 19(5):1627–1636, 2015.

[17] S. Chen, K. Ma, and Y. Zheng. Med3d: Transfer learning for 3d medical image analysis. *ArXiv*, abs/1904.00625:1–12, 2019.

[18] C.K. Chow and T. Kaneko. Automatic boundary detection of the left ventricle from cineangiograms. *Computers and Biomedical Research*, 5(4):388–410, 1972.

[19] A.A. Cruz-Roa, J.E.A Ovalle, A. Madabhushi, and F.A.G Osorio. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. *International Conference on Medical Image Computing and Computer Assisted Intervention*, 16:403–410, 2013.

[20] C. Dang, T. Mapayi, S. Viriri, and J. Tapamo. Comparative study of retinal vessel segmentation based on global thresholding techniques. *Computational and Mathematical Methods in Medicine*, 2015:895267, 2015.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A arge-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, page 248–255, June 2009.

[22] D. Erhan, Y. Bengio, A. Courville, P. Manzagol, P. Vincent, and S. Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11:625–660, Mar. 2010.

[23] Y. Feng, H. Zhao, X. Li, X. Zhang, and H. Li. A multi-scale 3d otsu thresholding algorithm for medical image segmentation. *Digital Signal Processing*, 60:186–199, 2017.

[24] P. Filipczuk, T. Fevens, A. Krzyżak, and R. Monczak. Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies. *IEEE Transactions on Medical Imaging*, 32(12):2169–2178, 2013.

[25] W.F. Flores, W.C. Pereira, and A.F.C. Infantosi. Improving classification performance of breast lesions on ultrasonography. *Pattern Recognition*, 48(4):1125–1136, 2015.

[26] B. Ginneken, A.A.A. Setio, C. Jacobs, and F. Ciompi. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 286–289, Apr. 2015.

[27] C.A. Glasbey. An analysis of histogram-based thresholding algorithms. *CVGIP: Graphical Models and Image Processing*, 55(6):532–537, 1993.

[28] Q. Guan, Y. Wang, B. Ping, D. Li, J. Du, Y. Qin, H. Lu, X. Wan, and J. Xiang. Deep convolutional neural network vgg-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study. *Journal of Cancer*, 10:4876–4882, 2019.

[29] A.M.J. Hanson, A. Joy, and J. Francis. Plant leaf disease detection using deep learning and convolutional neural network. *International Journal of Engineering Science*, 7:5324–5328, 2017.

[30] P. Haub and T. Meckel. A model based survey of colour deconvolution in diagnostic brightfield microscopy: Error estimation and spectral consideration. *Scientific Reports*, 5:12096, 2015.

[31] J. He, X. Jia, Sh. Chen, and J. Liu. Multi-source domain adaptation with collaborative learning for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11008–11017, June 2021.

[32] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[33] M.H. Hesamian, W. Jia, X. He, and P. Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of Digital Imaging*, 32:582–596, 2019.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[35] O. Iizuka, F. Kanavati, K. Kato, M. Rambeau, K. Arihiro, and M. Tsuneki. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Scientific Reports*, 10(1):1504–1515, 2020.

[36] Z. Jadoon, S.H. Ahmad, M.A. Khan Jadoon, A. Imtiaz, N. Muhammad, and Z. Mahmood. Retinal blood vessels segmentation using isodata and high boost

filter. *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pages 1–6, 2020.

[37] P. Ji, Y. Gong, M.L. Jin, X. Hu, G.H. Di, and Z.M. Shao. The burden and trends of breast cancer from 1990 to 2017 at the global, regional, and national levels: Results from the global burden of disease study 2017. *Frontiers in Oncology*, 10:1–13, 2020.

[38] W. Jiang and Z. Yin. Human activity recognition using wearable sensors by deep convolutional neural networks. In *Proceedings of the 23rd ACM International Conference on Multimedia*, page 1307–1310, 2015.

[39] T. Kaur and T.K. Gandhi. Automated brain image classification based on vgg-16 and transfer learning. In *2019 International Conference on Information Technology (ICIT)*, pages 94–98, 2019.

[40] A. Khan, A. Sohail, U. Zahoora, and A.S. Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, pages 1–62, 2020.

[41] D.F. King and L.A. King. A brief historical note on staining by hematoxylin and eosin. *The American Journal of Dermatopathology*, 8(2):168, Apr. 1986.

[42] M. Kiran, I. Ahmed, N. Khan, and A. G. Reddy. Chest x-ray segmentation using sauvola thresholding and gaussian derivatives responses. *Journal of Ambient Intelligence and Humanized Computing*, 10:4179–4195, 2019.

[43] M. Kowal, P. Jacewicz, and J. Korbicz. Combining image thresholding and fast marching for nuclei extraction in microscopic images. *Image Processing and Communications Challenges 8, IP&C 2016*, pages 195–202, 2016.

[44] M. Kowal, A. Marciniak, R. Monczak, and A. Obuchowicz. Discovering important regions of cytological slides using classification tree. *Image Processing and Communications Challenges 6, IPC 2014*, pages 67–74, 2014.

[45] M. Kowal, M. Skobel, A. Gramacki, and J. Korbicz. Breast cancer nuclei segmentation and classification based on a deep learning approach. *International Journal of Applied Mathematics and Computer Science*, 31(1):85–106, 2021.

[46] M. Kowal, M. Skobel, and N. Nowicki. The feature selection problem in computer–assisted cytology. *International Journal of Applied Mathematics and Computer Science*, 28(4):759–770, 2018.

[47] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*, page 3361. MIT Press, 1995.

[48] C.H. Li and C.K. Lee. Minimum cross entropy thresholding. *Pattern Recognition*, 26(4):617–625, 1993.

[49] T. Mahmood, J. Li, Y. Pei, F. Akhtar, A. Imran, and K. Rehman. A brief survey on breast cancer diagnostic with deep learning schemes using multi-image modalities. *IEEE Access*, 8:165779–165809, 2020.

[50] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, Feb. 2021.

[51] B. Miselis, T. Fevens, A. Krzyzak, M. Kowal, and R. Monczak. Deep neural networks for breast cancer diagnosis: fine needle biopsy scenario. *Proceedings of Polish Conference on Biocybernetic and Biomedical Engineering (PCBBE), Zielona Góra, Poland*, pages 131–142, Sept. 2019.

[52] P. Mitra, S.and Dey. Fine-needle aspiration and core biopsy in the diagnosis of breast lesions: A comparison and review of the literature. *Cytojournal*, 13, Aug. 2016.

[53] S.P. Mohanty, D.P. Hughes, and M. Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7:1419, 2016.

[54] C. D. L. Nascimento, S. D. D. S. Silva, T. A. D. Silva, W. C. D. A. Pereira, M. G. F. Costa, and C. F. F. Costa Filho. Breast tumor classification in ultrasound images using support vector machines and neural networks. *Research on Biomedical Engineering*, 32(3):283–292, 2016.

[55] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[56] Sourav Pramanik, Debotosh Bhattacharjee, and Mita Nasipuri. Wavelet based thermogram analysis for breast cancer detection. In *2015 International Symposium on Advanced Computing and Communication (ISACC)*, pages 205–212, 2015.

[57] J.M.S. Prewitt and M.L. Mendelsohn. The analysis of cell images. *Annals of the New York Academy of Sciences*, 128:1035–1053, 1966.

[58] Q. Qi, Y. Li, J. Wang, H. Zheng, Y. Huang, X. Ding, and G. K. Rohde. Breast cancer classification using deep learning approaches and histopathology image: A comparison study. *IEEE J. Biomedical and Health Informatics*, 23(5):2108–2116, 2019.

[59] Y. Qiu, S. Yan, R.R. Gundreddy, Y. Wang, S. Cheng, H. Liu, , and B. Zheng. A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology. *Journal of X-Ray Science and Technology*, 25(5):751–763, 2017.

[60] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Proceedings of 33rd International Conference on Neural Information Processing Systems*, volume abs/1902.07208, page 3347–3357, 2019.

[61] R. Rasti, M. Teshnehlab, and S. L. Phung. Breast cancer diagnosis in dce-mri using mixture ensemble of convolutional neural networks. *Pattern Recognition*, 72:381–390, 2017.

[62] A.S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 512–519, 2014.

[63] T.W. Ridler and S. Calvard. Picture thresholding using an iterative selection method. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(8):630–632, 1978.

[64] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241, 2015.

[65] K. Roy, D. Banik, D. Bhattacharjee, and M. Nasipuri. Patch-based system for classification of breast histology images using deep learning. *Computerized Medical Imaging and Graphics*, 71:90–103, 2019.

[66] Y. Ruan, A.E. Poirier, J. Pader, K. Asakawa, C. Lu, S. Memon, A.B. Miller, S.D. Walter, P.J. Villeneuve, W.D. King, K.D. Volesky, L. Smith, P. De, C.M. Friedenreich, and D.R. Brenner. Estimating the future cancer management costs attributable to modifiable risk factors in canada. *Canadian Journal of Public Health*, 112:1083–1092, 2021.

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

[68] R. K. Samala, H. P. Chan, L. M. Hadjiiski, M. A. Helvie, K. H. Cha, and C. B. Richter. Multi-task transfer learning deep convolutional neural network: Application to computer-aided diagnosis of breast cancer. *Physics in Medicine & Biology*, 62(23):8894–8908, 2017.

[69] J. Sauvola and M. Pietikäinen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000.

[70] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9:682, 2012.

[71] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[72] F. Shahidi, S. M. Daud, H. Abas, N. A. Ahamd, and N. Maarop. Breast cancer classification using deep learning approaches and histopathology image: A comparison study. *IEEE Access*, 8:187531–187552, 2020.

[73] Y. Shen, N. Wu, J. Phang, J. Park, K. Liu, S. Tyagi, L. Heacock, S.G. Kim, L. Moy, K. Cho, and K.J. Geras. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical Image Analysis*, 68:101908, 2021.

[74] J. Shi, A. Vakanski, M. Xian, J. Ding, and C. Ning. Emt-net: Efficient multitask network for computer-aided diagnosis of breast cancer. *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, abs/2201.04795:1–5, 2022.

[75] D.A. Shumway, A. Sabolch, and R. Jagsi. Breast cancer. *Medical Radiology*, pages 1–43, 2020.

[76] R.L. Siegel, K.D. Miller, H.E. Fuchs, and A. Jemal. Cancer statistics, 2021. *CA: A Cancer Journal for Clinicians*, 71:7–33, Jan. 2021.

[77] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[78] F.A. Spanhol, L.S. Oliveira, C. Petitjean, and L. Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016.

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

[80] C. Szegedy, S. Ioffe, V. Vanhoucke, and A.A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the 31 AAAI Conference on Artificial Intelligence*, page 4278–4284, 2017.

[81] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[82] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[83] L. Tabár, P.B. Dean, T.H.H. Chen, A.M.F. Yen, S.L.S. Chen, J.C.Y. Fann, S.Y.H. Chiu, M.M.S. Ku, W.Y.Y. Wu, C.Y. Hsu, Y.C. Chen, K. Beckmann, R.A. Smith, and S.W. Duffy. The incidence of fatal breast cancer measures the increased effectiveness of therapy in women participating in mammography screening. *Cancer*, 125:515–523, 2019.

[84] M. Talo, O. Yildirim, U. B. Baloglu, G. Aydin, and U. R. Acharya. Convolutional neural networks for multi-class brain disease detection using mri images. *Computerized Medical Imaging and Graphics*, 78:101673, 2019.

[85] T.F. Ursuleanu, A.R. Luca, L. Gheorghe, R. Grigorovici, S. Iancu, M. Hlusneac, C. Preda, and A. Grigorovici. Deep learning application for analyzing of constituents and their correlations in the interpretations of medical images. *Diagnostics*, 11(8):1–48, 2021.

[86] T. Wan, S. Xu, C. Sang, Y. Jin, and Z. Qin. Accurate segmentation of overlapping cells in cervical cytology with deep convolutional neural networks. *Neurocomputing*, 365:157–170, 2019.

[87] G. Wang, W. Li, M.A. Zuluaga, R. Pratt, P.A. Patel, M. Aertsen, T. Doel, A.L. David, J. Deprest, S. Ourselin, and T. Vercauteren. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Transactions on Medical Imaging*, 37:1562–1573, 2018.

[88] Y. Wen, Chen L., Y. Deng, and C. Zhou. Rethinking pre-training on medical imaging. *Journal of Visual Communication and Image Representation*, 78:103145, July 2021.

[89] B. Xu, X. Liu, J. anf Hou, B. Liu, J. Garibaldi, I. O. Ellis, A. Green, L. Shen, and G. Qiu. Attention by selection: A deep selective attention approach to breast cancer classification. *IEEE Transactions on Medical Imaging*, 39(6):1930–1941, 2020.

[90] S. Xu, Y. Liu, T. Zhang, J. Zheng, W. Lin, J. Cai, J. Zou, Y. Chen, Y. Xie, Y. Chen, and Z. Li. The global, regional, and national burden and trends of breast cancer from 1990 to 2019: Results from the global burden of disease study 2019. *Frontiers in Oncology*, 11:1–13, 2021.

[91] C.S. Yang, M.C. Chang, Y.J. Jan, and J. Wang. Fine needle aspiration of breast myofibroblastoma. *Acta Cytologica*, 54(3):356–358, 2010.

[92] J.C. Yen, F.J. Chang, and S. Chang. A new criterion for automatic multilevel thresholding. *IEEE Transactions on Image Processing*, 4(3):370–378, 1995.

[93] O. Yersal and S. Barutca. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World Journal of Clinical Oncology*, 5(3):412–424, 2014.

[94] T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.

[95] G.W. Zack, W.E. Rogers, and S.A. Latt. Automatic measurement of sister chromatid exchange frequency. *Journal of Histochemistry & Cytochemistry*, 25(7):741–753, 1977.

[96] N. M. Zaitoun and M. J. Aqel. Survey on image segmentation techniques. *Procedia Computer Science*, 65:797–806, 2015.

[97] D. Zhao, D. Zhu, J. Lu, Y. Luo, and G. Zhang. Synthetic medical images using f&bgan for improved lung nodules classification by multi-scale vgg-16. *Symmetry*, 10(10), 2018.

[98] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929. IEEE Computer Society, June 2016.