Classification of Breast Cancer Cytological Images using Vision Transformers

MohammadReza JebeliHajiAbadi

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Computer Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

April 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **MohammadReza JebeliHajiAbadi**

Entitled: **Classification of Breast Cancer Cytological Images using Vision Transformers**

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

Master of Computer Science (Computer Science)

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

Signed by the Final Examining Committee:

_____ Chair
Dr. Joey Paquet

_____ Examiner
Dr. Sudhir P. Mudur

_____ Examiner
Dr. Ching Y. Suen

_____ Supervisor
Dr. Adam Krzyzak

Approved by _____
Dr. Leila Kosseim, Chair

_____ 2024

_____
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Classification of Breast Cancer Cytological Images using Vision Transformers

MohammadReza JebeliHajiAbadi

This thesis evaluates the effectiveness of Vision Transformers (ViT) and Swin Transformers for breast cancer classification, highlighting their advantages over traditional Convolutional Neural Networks (CNNs) in processing cytological images. Amid the critical need for better breast cancer diagnostics, these transformer-based models emerge as a promising solution, adept at capturing complex spatial and contextual data in medical images.

The research methodology involved collecting and preprocessing a dataset of cytological and histopathological breast cancer images. The performance of the vision transformers was assessed using metrics such as accuracy, precision, recall, and AUC-ROC, and compared against established CNN architectures. The results demonstrate that vision transformers excel at extracting complex patterns from images, significantly outperforming current methods. Specifically, the study reports a 3.06% improvement in classification accuracy over traditional approaches, achieving 95.01% accuracy on test sets and perfect accuracy in validation.

The thesis underscores the potential of ViT and Swin models to advance early detection and diagnosis of breast cancer. Their success in the study suggests a transformative shift towards utilizing advanced deep learning architectures in medical image analysis. This approach not only enhances diagnostic accuracy but also offers a data-efficient solution to the

challenges of breast cancer classification. The findings advocate for further exploration of transformer-based models, which could redefine the standards of computer-aided diagnosis and significantly impact the field of cancer classification.

Acknowledgments

First and foremost, I am deeply grateful to my parents for their unwavering love, support, and encouragement throughout my academic journey. Their belief in me has been a driving force behind my achievements.

I extend heartfelt thanks to my friends, whose constant encouragement and camaraderie provided me with the strength to overcome challenges and persevere. Your friendship has been a source of inspiration.

I would also like to express my sincere appreciation to my supervisor, Dr. Adam Krzyzak, for their invaluable guidance, expertise, and mentorship. Their insights and feedback played a crucial role in shaping the direction of this research.

I extend my sincere gratitude to the medical professionals and researchers, especially Dr. Roman Monczak from the University Hospital in Zielona Góra, who generously shared access to the cytological breast cancer images. Their invaluable support made this study possible.

Lastly, I acknowledge the contributions of the open-source community and the creators of the ViT and Swin models, as well as the researchers whose work has informed and enriched my own.

To everyone who has played a role in my journey, thank you for your integral contributions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we begin by briefly introducing the subject under investigation and providing an explanation for the motivation behind this thesis (Section 1.1). Subsequently, we summarize the primary objectives pursued in this research within Section 1.2. Moving forward, we present a list of contributions to this study in Section 1.3. Lastly, we outline the organized structure of this thesis in Section 1.4.

1.1 Overview

Breast cancer continues to be a global health concern, prompting the search for new methodologies for accurate and early detection. Among the list of emerging technologies, artificial intelligence (AI) and deep learning holds great promise in enhancing medical image analysis. While Convolutional Neural Networks (CNNs) have shown notable success, the advancement of Transformers offers an interesting approach to enhance breast cancer classification.

Breast cancer's prevalence, combined with the necessity for quick diagnosis, highlights the need for efficient analysis techniques. Fine-Needle Biopsy (FNB), a fundamental approach in obtaining cellular material, has traditionally relied on pathologists' microscopic

assessment to identify cancer cells. This process, though skill-dependent, is labor-intensive and susceptible to human error. Automating this procedure with AI can revolutionize efficiency, enabling large-scale deployment and potentially reducing diagnostic variability.

Recent advancements in medical imaging and AI have driven the development of computer-aided diagnostic systems, speeding up analysis. However, a universal solution for routine diagnostics is still out of reach. The complex nature of medical image analysis, involving segmentation, feature extraction, and classification, has driven the utilization of Machine Learning (ML) and Deep Learning (DL) techniques. Within this framework, transformers have shown remarkable promise in diverse domains. We refer to Vaswani et al. (2017) [39] and Dosovitskiy et al. (2020) [8] for detailed information on transformers and vision transformers.

Transformers, originally tailored for natural image classification, possess the ability to capture complex spatial relationships and contextual details within images. Leveraging their capabilities in medical image analysis, particularly for microscopic breast cancer images, presents an opportunity to elevate diagnostic accuracy. Despite their promise, there is a lack of research on Transformers applied to microscopic images, especially cytological images creating a unique field for exploration and innovation.

In this context, this thesis starts on a journey to explore the efficacy of vision transformers such as ViT and Swin in breast cancer classification using microscopic images (histopathological and cytological images). By bridging the gap between transformer-based architectures and medical image analysis, this research seeks to uncover the potential of vision transformers in achieving state-of-the-art performance, even with limited data availability. The Swin Transformer, introduced in the paper Liu et al. (2021) [26], is a novel architecture that adapts the Transformer model for computer vision tasks. It stands out for its efficiency and scalability, particularly for high-resolution images. This work aims to advance automated breast cancer diagnosis by focusing on an ensemble method combining

ViT and Swin Transformer models. The research utilized the BreakHis dataset, a widely recognized public resource for histopathological breast cancer images, along with authentic cytological breast cancer images provided by Dr. Roman Monczak from the Department of Pathomorphology at the University Hospital in Zielona Góra.

1.2 Main Goals

The primary aim of this research is to develop an effective medical decision framework using transformer models to classify microscopic breast cancer images efficiently. Our emphasis lies in creating a novel transfer learning (TL) approach that can address the shortcomings found in existing literature and attain satisfactory performance in the classification task. We evaluate the framework's performance using real medical data from anonymous patients. The central thesis of this study can be summarized as follows:

We demonstrate that automatic binary classification of breast cancer images with high accuracy is achievable, even with limited annotated data, through the use of transfer learning (TL) techniques.

This task is challenging because it involves addressing several smaller, complex tasks, such as:

- Collecting medical images from biopsy specimens and creating a database of cytological breast cancer images.
- Performing all necessary preprocessing procedures on the collected images.
- Working with three datasets to optimize fine-tuning outcomes.
- Exploring four different training scenarios for the cytological dataset.
- Using two different state-of-the-art transformer models and their ensemble method.

- Utilizing attention maps to highlight which parts of the input are most crucial for generating an output in the trained models, thereby aiding in understanding how the model functions.

To showcase the superiority of our proposed framework over previous studies, we conduct experiments and compare our results with previous studies conducted on the same datasets. The subsequent section outlines the primary contributions of this thesis.

1.3 Contributions

This thesis makes several significant contributions to the field of breast cancer classification using microscopic images through a comprehensive exploration of transfer learning methodologies and datasets. The primary contributions of this research are as follows:

1. Due to the limited number of annotated data in the datasets especially the cytological dataset, we will employ efficient transfer learning to achieve high accuracy even with a small dataset.

2. To achieve effective transfer learning results, we experiment with three different datasets: natural images (ImageNet), BreakHis, and the Cytology dataset.

3. In this study, we aim to achieve faster training times compared to the previous studies on the datasets while increasing accuracy.

4. Four different scenarios for transfer learning are thoroughly examined. These scenarios consider various combinations of source and target datasets.

5. Unlike previous studies on the same dataset that simplified the process by dividing images into patches, our research distinguishes itself by performing image-level classification. This approach is notably more challenging due to the reduced amount of training data and the diminished image quality resulting from resizing images to fit the model's input dimensions.

6. We are applying various versions of ViT and Swin Transformer models, which have different input dimensions and patch sizes developed by Google and Microsoft, to a histopathological and cytological dataset. We will also share the model weights publicly for broader use.

7. We have developed an innovative ensemble architecture that combines ViT and Swin Transformer, applying it to the dataset.

8. Highlight which parts of the input are most important for generating an output in the trained models, through the attention maps, helping us understand how the model works.

1.4 Structure

This thesis comprises five chapters, organized as follows:

Chapter 1: This chapter offered a brief introduction to the research subject and outlines the motivation driving this study. The key objectives and novel contributions of this research are summarized later in this chapter.

Chapter 2: This chapter furnishes essential background information about diagnosing breast cancer using computer-aided systems. It subsequently reviews recent works in this domain, highlighting the distinct contributions of each. Additionally, the application of transfer learning in solving medical image classification issues, particularly in the context of breast cancer images, is discussed, along with a review of relevant studies in this field.

Chapter 3: Dedicated to the comprehensive explanation of the proposed method, this chapter thoroughly describes each stage of the designed approach, presented individually.

Chapter 4: This chapter presents the experimental outcomes of all strategies implemented in this research. Subsequently, these results are analyzed and discussed in detail.

Chapter 5: Concluding the thesis, this chapter summarizes the findings of the research and outlines potential avenues for future investigation.

Appendix A: This appendix includes confusion matrices and ROC graphs for various

scenarios.

Chapter 2

Background and Literature Review

2.1 Overview

In this chapter, our goal is to give a thorough overview of the methods used in Medical Imaging, particularly in the classification of breast cancer images. We start by looking at common screening methods used to detect breast cancer. After that we discuss different architectures that have been used in this area. To provide some context, we do a detailed review of past research and what they found in this field. This thorough analysis helps us compare our own research with what has already been done, therefore we can see how significant and innovative our contributions are.

Cancer is caused by abnormal changes or mutations in the genes responsible for cell growth. These genes are in the cell's center (nucleus) and are constantly active. Cells have a specific lifespan where they keep dividing to replace old ones with new cells. If cell growth is not controlled, it can lead to issues. Sometimes, one cell or a small group of cells in the body starts growing uncontrollably, forming a tissue mass, known as a tumor. Not all tumors are cancer; they can be either benign (not cancerous) or malignant (cancerous). Malignant tumors spread through the lymphatic system and take nutrients from the body, while benign tumors do not invade nearby tissues and are manageable (Mahmood

et al. (2020) [27]). Breast cancer is a type of disease where cells in breast tissue divide uncontrollably, leading to a tumor.

One of the methods to evaluate breast masses is known as the triple test, which includes a physical examination (palpation), mammography, and fine-needle biopsy. The fine-needle biopsy without aspiration (Yang et al. (2010) [41]) involves collecting material directly from the breast tumor. This collected material is then examined under a microscope by a cytologist or a pathologist to identify cancer cells, a task that requires expertise. This is where a CAD system can assist specialists in making precise cancer cell diagnoses.

In general, malignant cells often have irregular shapes, enlarged nuclei, and variable nuclear-to-cytoplasmic ratios compared to benign cells. Malignant cells may also show hyperchromasia (darker staining), irregular nuclear contours, prominent nucleoli (small, round structures within the nucleus), and increased nuclear size. They can appear disorganized, with irregular spacing and overlapping, while benign cells tend to be more orderly. Additionally, malignant tumors often exhibit a higher cell density compared to benign lesions

2.2 Screening methods

2.2.1 Mamography

Mammography, a fundamental breast cancer screening method, employs low-energy X-rays to create detailed images of breast tissue. Its primary aim is early cancer detection, often before symptoms emerge. By identifying abnormalities like masses or microcalcifications, mammography facilitates timely intervention, improving treatment success rates.

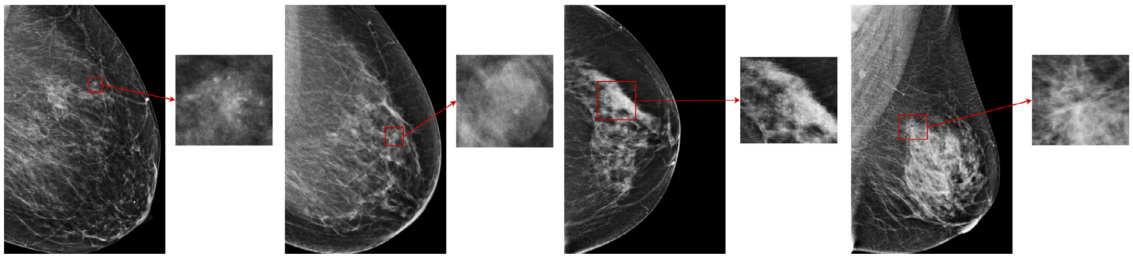


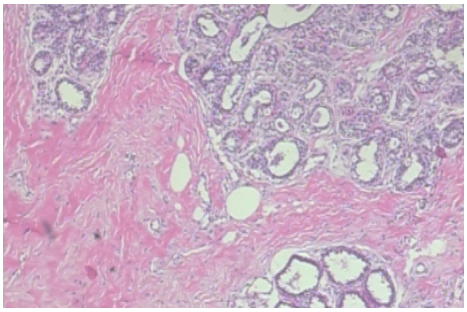
Figure 2.1: Four examples of breasts that were biopsied along with the annotated findings. The breasts (from left to right) were diagnosed with benign calcifications, a benign mass, malignant calcifications, and malignant architectural distortion. (Shen et al. (2021) [34])

Mammography involves compressing breast tissue for clear imaging, with digital technology enhancing image processing and analysis. Additionally, 3D mammography (tomosynthesis) offers improved lesion detection by providing three-dimensional views, reducing overlap issues.

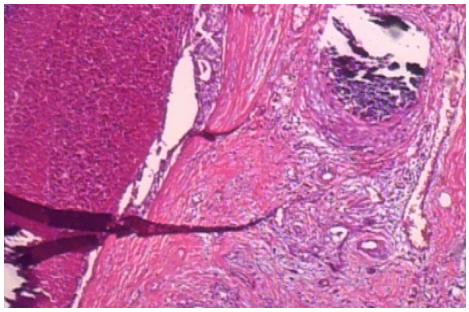
Routine mammography screenings for specific age groups and risk factors have significantly reduced breast cancer mortality rates. It is one of the earliest approaches to detect breast cancer, and in this thesis we will explore several studies conducted on mammography datasets.

2.2.2 Histopathology

Histopathology is a medical screening method that involves the microscopic examination of tissue samples to diagnose diseases and understand their underlying causes. This technique plays a critical role in the diagnosis and classification of various medical conditions, including cancer. In the context of breast cancer, histopathology involves the examination of breast tissue samples, typically obtained through biopsy or surgical procedures. These tissue samples are processed, embedded in paraffin wax, sliced into thin sections, and stained. A pathologist then examines these stained tissue sections under a microscope to identify and classify abnormalities, such as cancer cells. Histopathology provides valuable



(a) Adenosis benign



(b) Ductal carcinoma malignant

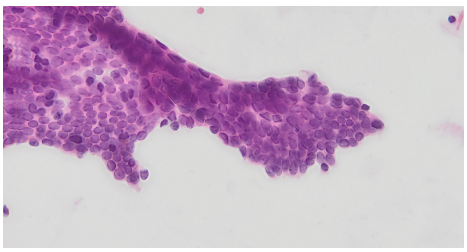
Figure 2.2: Two examples of histopathology breast images from BreakHis dataset

information about the type, stage, and grade of breast cancer, guiding treatment decisions and prognosis assessment. It remains a cornerstone of cancer diagnosis and plays a pivotal role in personalized medicine approaches for patients.

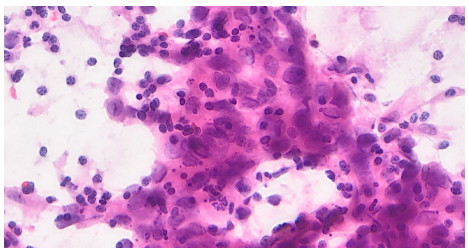
2.2.3 Cytology

Cytology, often referred to as cytopathology, is a medical screening method that involves the examination of individual cells to detect abnormalities and diagnose diseases. This technique is particularly valuable in cancer detection and diagnosis. In the context of breast cancer, cytology screening typically involves collecting cells from breast tissue through various methods, such as fine-needle aspiration or brushing, and then preparing these cells for microscopic examination. A cytotechnologist or pathologist examines the collected cells under a microscope to identify any abnormal or cancerous cells. Cytology screening is useful for early detection and diagnosis of breast cancer, providing information about cell characteristics, malignancy, and potential metastasis. It is a valuable tool for guiding treatment decisions and assessing patient prognosis.

In summary, cytology is a method primarily focused on the examination of individual cells and is often used for early detection, while histopathology involves the examination of intact tissue samples and provides a more comprehensive assessment of tissue structure



(a) Benign



(b) Malignant

Figure 2.3: Two examples of Cytopathology breast images from the University Hospital in Zielona Góra, Poland

and diseases.

2.3 Architectures

2.3.1 Neural Networks

In the realm of machine learning and artificial intelligence, neural networks stand as a pivotal paradigm for tackling complex tasks. These networks are designed to learn and model intricate relationships within data by mirroring the structure and function of the human brain. The fundamental premise of a neural network is to map input data to desired output, effectively capturing and understanding patterns, features, and associations present in the data.

Consider a dataset D, comprising pairs of input-output samples (x, y), sampled independently and identically from a target distribution Q(x, y). A neural network's primary objective is to learn a function f, often represented as f(x) = y, that takes an input x_n and transforms it into a corresponding output y_n. For example in handwritten digit recognition, This transformation process is akin to decoding handwritten digits from images, where x_n symbolizes an input image, and y_n signifies the digit depicted within the image.

The term "neural network" derives from the network's underlying architecture, which consists of interconnected layers. Each layer, denoted as f_l, encompasses a distinct set of

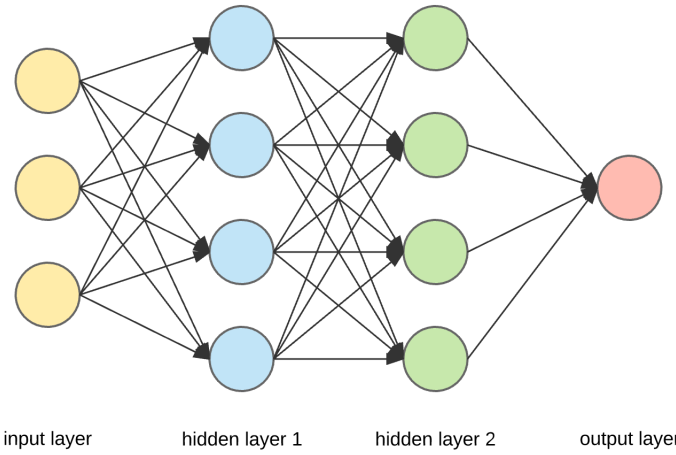


Figure 2.4: An example of a three layer neural network.

neurons or units, where each unit maps a vector input to a scalar output. These layers are parameterized by weight vectors θ_l and follow a specific activation function, typically non-linear except for the output layer. The composition of these layers results in a hierarchical function, for instance, f(x) = $f_3(f_2(f_1(x)))$, where f_1 is the first layer, f_2 is the second, and f_3 is the output layer. Layers other than the output layer are referred to as hidden layers due to the absence of target output for them.

Neural networks leverage the power of non-linear transformations, facilitated by activation functions like φ, to capture intricate patterns in data. By stacking multiple layers of neurons, these networks possess the remarkable ability to approximate a wide array of non-linear functions. In fact, when equipped with a sufficient number of neurons in hidden layers, neural networks can serve as universal approximators, capable of representing even highly complex relationships within the data.

The learning process of a neural network hinges on minimizing a loss function, which measures the discrepancy between the network's output and the desired target output. Given the non-linear nature of neural networks, these loss functions often become non-convex, making optimization a challenging task. Nevertheless, iterative gradient descent

methods, guided by backpropagation, enable the network to iteratively update its parameters to minimize the loss. These updates occur over batches of training data, enhancing convergence and robustness.

In practical applications, neural networks offer a versatile and powerful tool for solving complex problems such as image classification and segmentation. By distinguishing feature extraction from classification, neural networks transform the input space to enable more straightforward class discrimination. This distinction forms the basis for various applications explored in this thesis.

2.3.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have revolutionized computer vision, providing a powerful tool for tasks like image classification, object detection, and semantic segmentation. Their ability to extract complex patterns from visual data mirrors the functionality of the human visual cortex, making them adept at interpreting diverse visual content. This capability has positioned CNNs as foundational elements in various image-related applications.

At their core, CNNs utilize convolutional layers to process input images, leveraging learnable filters that scan the image to create feature maps. These maps capture visual characteristics, ranging from basic elements like edges to more complex features like parts of objects. One of the major strengths of CNNs is their hierarchical structure, where each layer builds upon the previous one, refining the extracted features.

A key advantage of CNNs is their efficiency in learning these features directly from the data, eliminating the need for manual feature extraction. This automated learning process is achieved through training on large datasets, where the network adjusts its filters to optimize feature detection. As a result, CNNs can automatically identify the most relevant features for a given task, a significant step forward in computer vision.

In the medical field, CNNs have shown remarkable potential, especially in breast cancer classification. They excel in analyzing complex medical images, identifying patterns that may be indicative of malignant or benign conditions. This has made CNNs indispensable in medical diagnostics, where they assist in accurate and early detection of diseases.

Furthermore, the integration of CNNs with other techniques like transfer learning has further enhanced their capabilities. Transfer learning allows these networks to apply knowledge gained from one task to another, improving their performance in new, unseen datasets.

Overall, CNNs represent a significant advancement in artificial intelligence, with wide-ranging applications that extend beyond image analysis to fields like healthcare and autonomous systems. Their ability to learn and adapt makes them a versatile tool in the rapidly evolving landscape of technology.

AlexNet

AlexNet, a pioneering convolutional neural network (CNN) model introduced by Krizhevsky et al. (2012) [23], brought several essential innovations to computer vision. Its architecture consisted of eight layers, including five convolutional and three fully connected layers, enabling deep feature extraction. AlexNet's use of the Rectified Linear Unit (ReLU) activation function facilitated faster training. The adoption of small 3x3 convolutional filters and max-pooling layers allowed it to capture intricate image features effectively. Local Response Normalization (LRN) layers enhanced neuron responses and generalization. To combat overfitting, AlexNet introduced dropout in fully connected layers. Trained on the ImageNet dataset, this network achieved a significant reduction in the top-5 error rate during the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012, setting a new standard in image classification tasks.

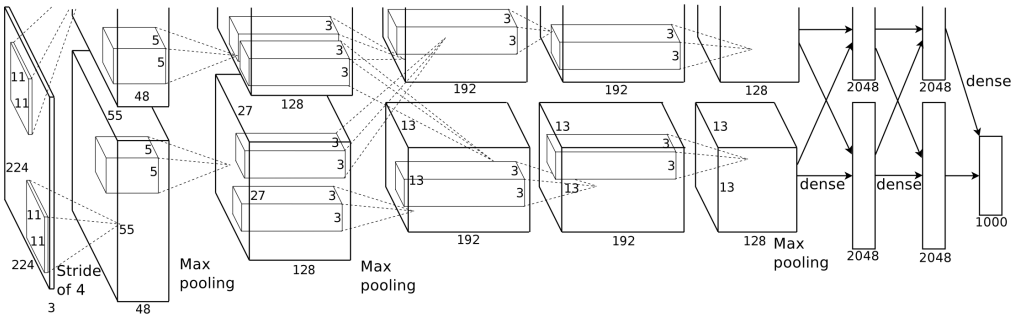


Figure 2.5: AlexNet architecture [23]

VGG-16

VGG-16, a convolutional neural network (CNN) architecture introduced by Simonyan and Zisserman (2014) [35] in 2014, represented an improvement over the well-known AlexNet model. This improvement was achieved by replacing large filters with smaller 3x3 filters, which reduced computational complexity. Additionally, the network's depth was increased to include 16-19 layers, resulting in significant performance enhancements. These modifications led to VGG-16 achieving top rankings in image classification and localization competitions. However, the network's depth and size, comprising over 138 million parameters, presented challenges such as complex deployment and the risk of overfitting. Despite its straightforward design, VGG-16 continues to serve as a benchmark in various computer vision tasks due to its strong performance.

VGG-19

VGG-19 by is a deep neural network architecture used in the ImageNet challenge for a 1000-class classification task. It is similar to VGG-16 but has three extra convolutional layers, making a total of 19 layers with 16 trainable convolutional layers and three fully connected layers.

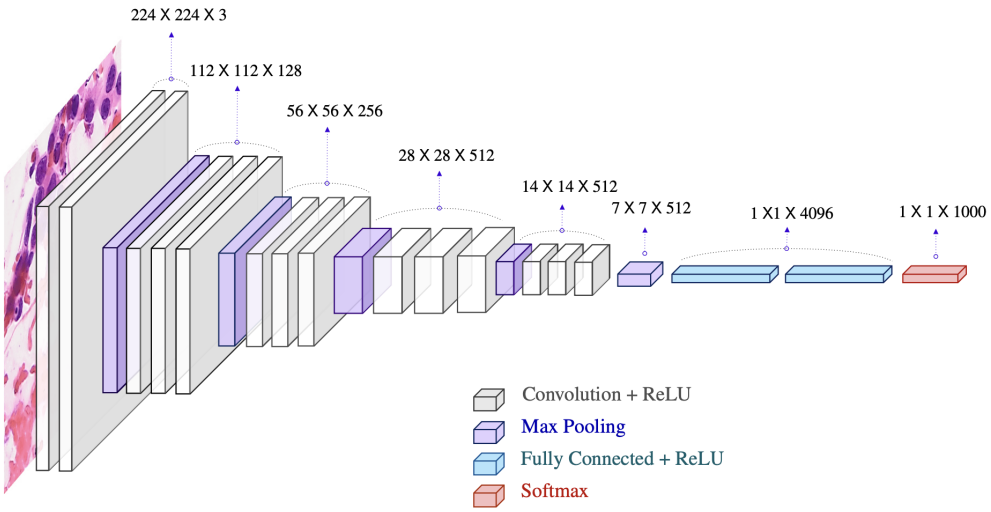


Figure 2.6: VGG-16 architecture

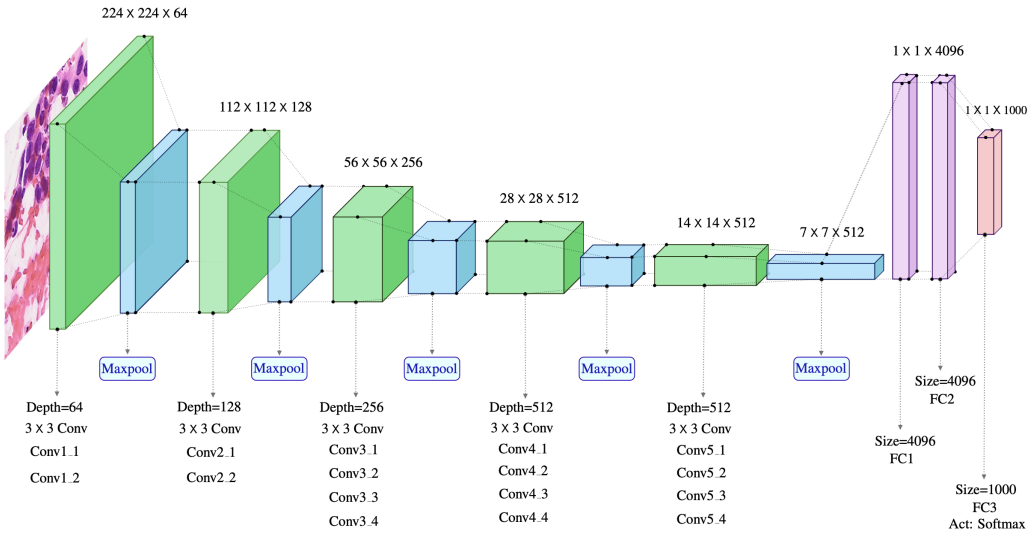


Figure 2.7: VGG-19 architecture

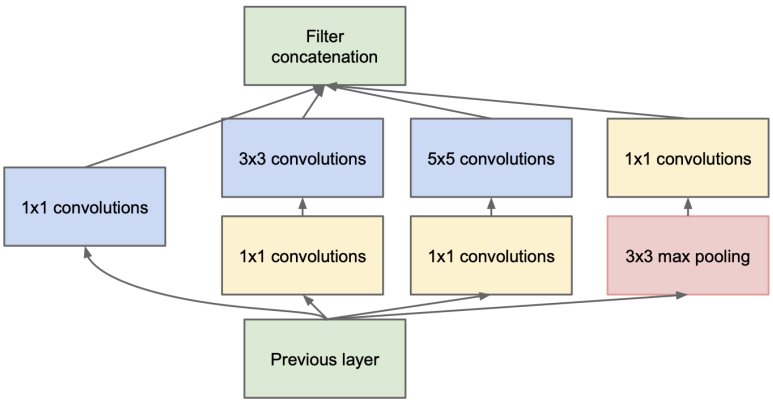


Figure 2.8: Inception module with dimensionality reduction architecture by Szegedy et al. (2015) [37]

GoogleNet

GoogleNet, also called Inception, is a well-known CNN architecture by Szegedy et al. (2015) [37] known for its deep and intricate design, which efficiently captures features at various levels. One of its important characteristics is the use of inception modules, which employ filters of different sizes within the same layer to capture both fine and large-scale features, enhancing the network's ability to represent data. GoogleNet's design is good at reducing computational complexity compared to earlier deep networks while maintaining high accuracy, thanks to its effective use of 1x1 convolutions to simplify data. Additionally, it introduced auxiliary classifiers, helping with gradient flow and addressing the vanishing gradient issue during training. These innovations led to its success in the ILSVRC 2014 competition. Inception v3 is also an improved version of the original GoogleNet. Inception v3 includes refinements and enhancements, making it more efficient and accurate for image classification tasks.

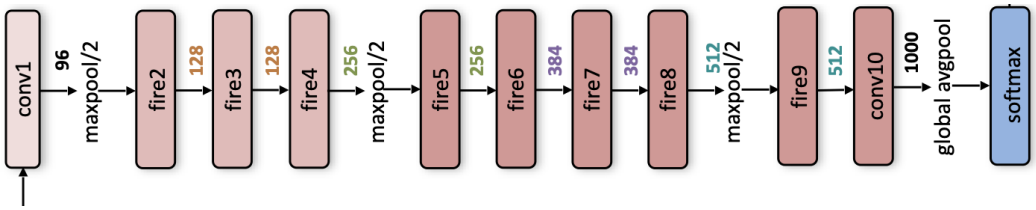


Figure 2.9: SqueezeNet architecture by Iandola et al. (2016) [15]

SqueezeNet

SqueezeNet is a notable CNN architecture by Iandola et al. (2016) [15] renowned for its emphasis on model size reduction and computational efficiency. It achieves this efficiency by employing a streamlined network structure with fire modules, which consist of both squeeze and expand layers. The squeeze layer uses 1x1 convolutions to reduce the number of input channels, effectively compressing the data. Subsequently, the expand layer employs 1x1 and 3x3 convolutions to extract and diversify features. SqueezeNet reduces the number of parameters, making it highly memory-efficient and suitable for resource-constrained environments.

ResNet

ResNet, also known as Residual Network, is a special kind of neural network developed by He et al. (2016) [13]. It performed very well in the 2015 ILSVRC image recognition and segmentation competitions. What makes ResNet special is its use of shortcuts, or skip connections, to solve problems like gradients disappearing and accuracy not improving with deeper networks. In general, deeper networks can handle more complicated tasks but become tough to train and might not get more accurate. ResNet's residual blocks allow information to pass through a shortcut, solving this problem. Each block has two 3x3 convolution layers, some normalization, and activation, along with a direct connection between the input and the addition operator. ResNet was inspired by VGG-19 but added skip

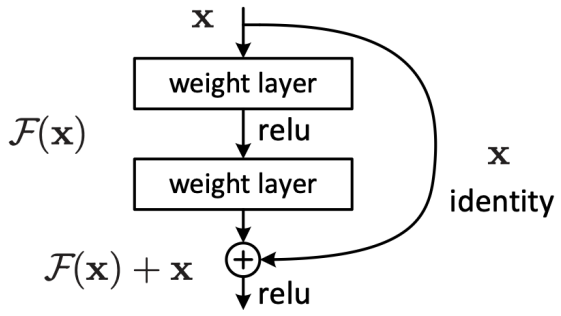


Figure 2.10: Residual learning: a building block [13]

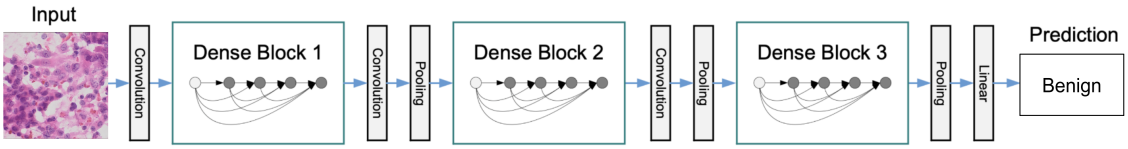


Figure 2.11: A deep DenseNet with three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature-map sizes via convolution and pooling. [14]

connections. There are different versions, like ResNet-34, ResNet-50, which only differ on the number of layers.

DenseNet

DenseNet, created by Huang et al. (2017) [14], is a type of CNN like ResNet. It helps with the problem of vanishing gradients. DenseNet's design has layers that connect to each other. Each layer takes input from the ones before and passes its own information to the ones after. This is done using dense connections and blocks inside the network. These blocks have several layers that do the same thing. This setup lets each layer access gradients and share important information, making data flow better. DenseNet also reduces how many learning parts the network needs, which makes training faster.

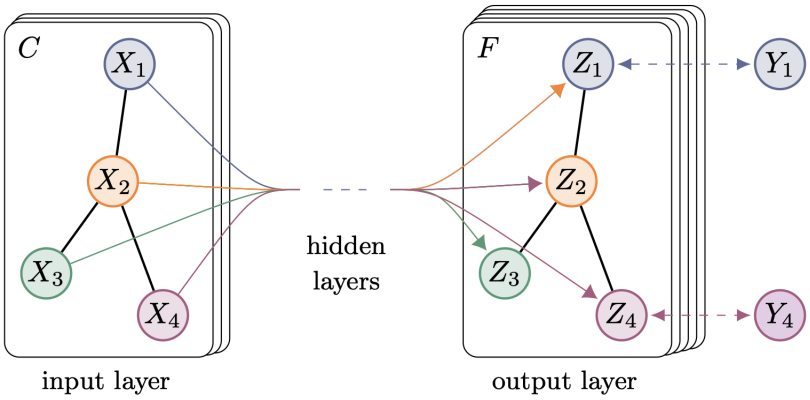


Figure 2.12: Graph Convolutional Network [17]

2.3.3 Graph Convolutional Neural Networks

Graph Convolutional Neural Networks (Graph CNNs) are a specialized class of neural networks designed for processing and analyzing data structured as graphs or networks. Unlike traditional neural networks that operate on grid-structured data like images, Graph CNNs are tailored for data with complex relationships and connections, such as social networks, recommendation systems, and molecular structures.

Graph CNNs extend the concept of convolutional layers to graph-structured data. They learn features by aggregating information from neighboring nodes within the graph, enabling them to capture local and global patterns effectively. This approach is particularly useful for tasks involving node classification, link prediction, community detection, and graph-based recommendation.

In essence, Graph CNNs leverage the inherent graph topology to generalize and make predictions about nodes or edges in the graph. They have found applications in diverse fields, including social network analysis, bioinformatics, natural language processing, and recommendation systems. Graph CNNs continue to drive innovation by providing powerful tools for extracting insights from interconnected data.

2.3.4 Transformers

In the field of natural language processing, Transformers changed the way machines understand and generate human language. They were first designed for language tasks but are now used in various areas like computer vision, speech recognition, and medical image analysis. In the Transformer architecture, there is a big shift away from traditional models like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). Transformers excel in tasks that demand a deep understanding of long-range relationships within data. The distinguishing feature of Transformers is their self-attention mechanism, which allows them to weigh the importance of different elements in a sequence when making predictions. This self-attention mechanism operates in parallel across all elements, enabling the model to attend to relevant information regardless of its position in the sequence. This is different from RNNs and CNNs, which process data sequentially or with fixed receptive fields.

Furthermore, Transformers are characterized by their deep architecture, composed of multiple layers of self-attention and feedforward neural networks. Each layer refines the model's understanding of the input data, gradually capturing higher-level abstractions and semantic representations. This depth and hierarchical processing enable Transformers to excel at capturing complex patterns and relationships, making them highly versatile and expressive models.

Transformers have demonstrated exceptional capabilities not only in text-related tasks but also in computer vision, where they have enabled groundbreaking advancements in image captioning, object detection, and segmentation. Their ability to capture contextual information and relationships across data modalities has unlocked new possibilities for multimodal applications, bridging the gap between language and vision. We will explain how the first published transformer, ViT, and Swin Transformer work in Chapter 3.

2.4 Literature Review

2.4.1 Overview

Breast cancer remains a significant challenge in healthcare, demanding innovative approaches for accurate diagnosis and classification, which in turn affect treatment and patient outcomes. In recent years, deep learning methods, including transformer-based models, have shown promise in improving breast cancer classification.

This review explores the evolving landscape of breast cancer classification with a focus on transformers. Transformers, originally developed for natural language processing, have expanded into computer vision and medical image analysis. They provide new possibilities for analyzing intricate patterns in breast microscopic images and revolutionizing breast cancer detection.

We delve into the history of breast cancer classification methods, from early foundations to recent transformer-based breakthroughs. The review covers historical context, important studies, challenges, and the impact of various techniques like CNNs, GCNs, and transformers in medical imaging and breast cancer classification.

Our journey reveals that transformers have the potential to enhance diagnostic accuracy, reduce false positives, and enable earlier interventions. Each study and innovation brings us closer to a future where breast cancer is not only detected but also understood in greater detail. This review highlights key milestones and sets the stage for the next chapter, where we examine the current state of breast cancer classification with transformers, including methodologies, findings, and research gaps.

2.4.2 Early Stages

Once medical images became scannable and loadable into computers, researchers began constructing automated analysis systems. In the early years, spanning from the 1970s

to the 1990s, medical image analysis primarily relied on a sequential application of low-level pixel processing techniques such as edge and line detection filters and region growing. Mathematical modeling techniques, like fitting lines, circles, and ellipses, were also employed to create rule-based systems tailored to specific tasks. These systems bore resemblance to expert systems that featured numerous if-then-else statements, a hallmark of artificial intelligence during that era. However, these expert systems, often referred to as ”GOFAI” (good old-fashioned artificial intelligence), were known for their brittleness, similar to rule-based image processing systems [25].

Towards the late 1990s, supervised techniques gained prominence in medical image analysis. These techniques involved the use of training data to develop systems. Notable examples included active shape models for segmentation, atlas methods that utilized fit atlases as training data, and the introduction of feature extraction and statistical classifiers for computer-aided detection and diagnosis. This approach, rooted in pattern recognition and machine learning, remains highly popular and underpins many successful commercially available medical image analysis systems. Consequently, we witnessed a transition from systems entirely crafted by humans to systems trained by computers using example data from which feature vectors are derived. Computer algorithms are tasked with determining the optimal decision boundaries within high-dimensional feature spaces. An integral aspect of designing such systems is the extraction of discriminant features from images, a process that still relies on human researchers, resulting in what are termed ”handcrafted features” systems [25].

2.4.3 Deep Networks

The emergence of neural networks and machine learning revolutionized the field of medical imaging. A significant portion of research efforts has been dedicated to the use of deep networks for the classification of breast images.

Cheng et al. (2016) [6] employed a Stacked Autoencoder (SAE) along with a denoising technique known as Stacked Denoising Autoencoder (SDAE) to distinguish between breast ultrasound lesions and lung CT nodules. Their approach involved resizing the regions of interest (ROIs) in the images to a standardized 28×28 pixel dimension. In this setup, all the pixels within each patch were considered as input for the SDAE. During the initial pretraining phase, the researchers introduced random noise to the input patches, thereby improving their model's ability to handle noisy data. Subsequently, in the fine-tuning stage, they incorporated the resized scale factors for both dimensions of the ROIs and the aspect ratios of the original ROIs to preserve the original data characteristics.

Shen et al. (2015) [33] devised a hierarchical learning framework incorporating a multiscale Convolutional Neural Network (CNN) designed to capture lung nodules of various sizes. Within this CNN architecture, they integrated three separate CNNs that took nodule patches from distinct scales as their input, all working in parallel. To mitigate the risk of overfitting, the researchers ensured that the parameters of these three CNNs were shared during the training process. The activations originating from the top hidden layers of the three CNNs, each corresponding to a specific scale, were combined to construct a unified feature vector. In the classification step, Shen et al. employed a Support Vector Machine (SVM) equipped with a radial basis function kernel and a random forest approach. These models were trained to minimize what they referred to as "companion objectives," which encompassed a combination of the overall hinge loss function and the sum of companion hinge loss functions [24].

Miselis et al. (2020) [29] investigated the effectiveness of various state-of-the-art deep neural network architectures in classifying breast cancer using cytological images from fine-needle biopsies (FNBs). They evaluated five different CNN architectures, including AlexNet, GoogleNet, SqueezeNet, DenseNet, and Inception-V3, for binary breast cancer classification. Among these, Inception-V3 performed the best, achieving an accuracy of

91.86% and an area under the ROC curve (AUC) of 0.97.

In contrast, Kowal et al. (2021) [21] proposed an alternative approach focused on classifying single-cell nuclei in breast cancer cytological images. They initially segmented the images using the U-Net neural network and marker-controlled watershed transform. Then, they classified individual cell nuclei as either benign or malignant based on a set of manually crafted features. For the two-class classification, the SVM classifier achieved an accuracy of 88.2%.

Shamshiri et al. (2023) [32] focuses on breast cancer classification using a transfer learning approach. The study addresses the challenge of limited annotated data by proposing a method that leverages compatible-domain transfer learning with CNNs.

In essence, the authors aim to improve breast cancer classification models when there is insufficient labeled data by transferring knowledge from a related domain with more data. The target dataset in this study is the cytological dataset. The author utilizes six state-of-the-art deep CNNs, including VGG-16, VGG-19, ResNet101-V2, DenseNet-169, Inception-V3, and InceptionResNet-V2. They also explore three distinct scenarios for training deep CNNs: initializing weights through pretraining on the BreakHis dataset, initializing through pretraining on ImageNet, or using random values. Additionally, two different fine-tuning approaches, partial fine-tuning and complete fine-tuning, are considered, and the results are compared to identify the optimal version.

The best-performing model in the study was DenseNet-169, achieving a test accuracy of 94.55%.

2.4.4 Graph Convolutional Neural Networks

An additional captivating field of study involves the application of Graph Convolutional Neural Networks (GCNNs). GCNNs are renowned for their capacity to account for the interconnections between various patches within the same whole slide image, a factor often

overlooked in previous architectures.

Zhang et al. (2021) [42] developed BDR-CNN-GCN to improve the detection of malignant breast lesions in mammograms, leveraging the advantages of both GCNs and CNNs simultaneously. They employed a standard 8-layer CNN and integrated batch normalization (BN) and dropout (DO). Moreover, traditional max pooling was replaced with rank-based stochastic pooling (RSP), leading to the creation of BDR-CNN, which combines CNN, BN, DO, and RSP. BDR-CNN was then combined with a two-layer GCN, resulting in the BDR-CNN-GCN model, utilized for the analysis of breast mammograms as a 14-way data augmentation technique. Finally, the algorithm was executed on the breast mini-MIAS dataset. The results demonstrated an accuracy of 96.10% with a variance of ±1.60%.

Gao et al. (2022) [11] focuses on the spatial correlation among different tissue components to be discovered by GCN. They use a CNN in order to get the high-level features and then feed them to GCN to consider the spatial correlation. Clique GCN (cGCN) is introduced to enhance graph representation with establishing forward and backward connections between any two graph convolution layers. Moreover, a group graph convolution is developed to improve feature representation and reduce redundancy. They evaluated the model on Databiox dataset and BioImaging 2015 challenge (BI) dataset acheiving 83.04±2.5% accuracy on Databiox dataset, and 94.4±2.21% accuracy on BI dataset.

Konda et al. (2020) [18] has an interesting approach. They consider each Whole-slide-Image (WSI) as a collection of patches with a single label. They first divide the WSI into patches, retaining only those containing tissue material, and then classify the entire WSI, not individual patches. In other words, they process the entire WSI, requiring only one WSI label without any patch-level annotations. This way, they don't lose any spatial relationships between patches. Two methods were employed to generate node feature representations for subsequent use in a graph network. The first method involved using

the raw image data as the feature representations. The second approach utilized a trainable CNN to extract essential features from each segmentation. They evaluated the model on a breast cancer histopathological dataset, achieving an accuracy of 87%, and also on a histopathological colon cancer dataset, achieving an accuracy of 85%.

Gao et al. (2021) [12] also uses a similar approach. However, they propopse a group quadratic graph convolutional network (GQ-GCN), which uses a CNN to extract features for further graph construction. The method is developed to implement both feature selection and compression of graph represenation. to enhance the representation ability of a single neuron. They experimented on Databiox dataset and BioImaging 2015 challenge (BI) dataset. They acheived 83.49±2.61% accuracy on Databiox dataset, and 94.38±0.86% accuracy on BI dataset.

2.4.5 Transformers

Vaswani et al. (2017) [39] introduces the Transformer architecture, a novel neural network model that revolutionized natural language processing and machine translation tasks. Unlike traditional models that rely heavily on recurrent or convolutional layers, the Transformer uses a self-attention mechanism to process input data in parallel, eliminating the need for sequential processing. This allows the Transformer to handle long-range dependencies and capture contextual information efficiently.

The core innovation of the Transformer is the multi-head self-attention mechanism, which allows the model to attend to different parts of the input sequence simultaneously. The authors also introduce positional encodings to account for the order of words in a sequence since the self-attention mechanism lacks inherent positional information.

The Transformer achieved state-of-the-art results on machine translation tasks and demonstrated the power of self-attention in sequence-to-sequence tasks. Its architecture has since become a foundation for various natural language processing tasks, including language

generation, text summarization, and question-answering systems. The paper's impact extends beyond NLP, as it has influenced developments in computer vision and other fields, highlighting the significance of the Transformer model in modern deep learning research.

Dosovitskiy et al. (2020) [8] explores the application of Transformers, originally designed for natural language processing, to the field of computer vision for image recognition tasks. The authors propose a novel model called the Vision Transformer (ViT) that adapts the Transformer architecture to process images.

In the ViT model, an input image is divided into non-overlapping patches, which are then linearly embedded into token sequences. These token sequences are treated as inputs to the Transformer model. Unlike traditional Convolutional Neural Networks (CNNs), the ViT model does not rely on convolutional layers but instead leverages self-attention mechanisms to capture long-range dependencies and contextual information within the image.

The authors demonstrate that the ViT model achieves state-of-the-art results on various image classification benchmarks when trained on large-scale datasets. They also highlight the importance of pre-training the model on a large corpus of images, similar to how language models are pre-trained on text data. The paper underscores the potential of Transformers in computer vision tasks and paves the way for further exploration of these models in the field of image recognition.

Matsoukas et al. (2021) [28] investigates whether vision transformers can outperform CNNs in all conceivable scenarios. Transformers excel over CNNs in capturing long-range relationships within an image. They enable adaptive modeling and also offer a built-in insight to the model. Training from scratch, ViTs perform less effectively than CNNs in scenarios with limited data. However, when employing transfer learning, the performance gap between CNNs and ViTs narrows, and their performance becomes comparable. The highest level of performance is achieved through self-supervised pre-training followed by fine-tuning, where ViTs hold a slight advantage over equivalent CNNs.

Dai et al. (2021) [7] introduces The TransMed model for multi-modal medical image classification, leveraging the strengths of both CNN and transformer approaches. It efficiently extracts low-level image features and establishes connections between different data sources. The model was tested on two datasets, parotid gland tumors, and knee injuries, achieving significant improvements in accuracy compared to other CNN-based models, 10.1% for the former and 1.9% for the latter.

Touvron et al. (2021) [38] addresses the challenge of training large-scale image recognition models with limited labeled data. The authors propose a method to improve the data efficiency of image Transformers while maintaining their performance.

The key innovation in this paper is the introduction of a novel attention mechanism called ”Data-Efficient Attention” (DEA). DEA allows the model to focus on relevant parts of the input data, reducing the computational cost and memory requirements compared to standard self-attention mechanisms. This attention mechanism is designed to be more data-efficient, making it possible to train image Transformers effectively even with smaller datasets.

Additionally, the paper explores the concept of ”knowledge distillation through attention.” This involves transferring knowledge from a large, pre-trained model (teacher) to a smaller, data-efficient model (student) by distilling the attention maps learned by the teacher model. This process helps the student model benefit from the knowledge learned by the teacher model while maintaining a smaller footprint.

The results demonstrate that the proposed data-efficient attention mechanism and knowledge distillation techniques significantly improve the efficiency of image Transformers, allowing them to achieve competitive performance with less labeled data. This work contributes to the development of more data-efficient and computationally economical image recognition models.

Chapter 3

Methodology

This chapter provides a comprehensive explanation of the methodology employed in this thesis. To categorize breast cancer microscopic images as either benign or malignant, several steps are required. Each step will be detailed individually in this chapter.

In summary, the initial step involves acquiring medical images from biopsy specimens, as described in Section 3.1. To facilitate the training of transformers for the classification task, the images are then divided into subsets for training, validation, and testing, and pre-processing is performed (Section 3.3). Finally, in Section 3.4, we discuss how transformers work in details and how the task of classifying the images into benign and malignant categories is executed using. Section 3.5 will examine all the fine-tuning scenarios.

3.1 Image acquisition

3.1.1 Histopathological Dataset (BreakHis)

The BreaKHis dataset introduced by Spanhol et al. (2015) [36] consists of microscopic biopsy images of both benign and malignant breast tumors. These samples are derived from breast tissue biopsy slides, which are stained with hematoxylin and eosin (HE). The

specimens are collected by SOB, prepared for histological examination, and then labeled by pathologists. To prepare them for slide mounting, sections of 3 m are obtained using a microtome. After staining, glass coverslips are applied to these sections. Subsequently, anatomopathologists visually identify the tumoral areas in each slide by examining tissue sections under a microscope. The final diagnosis for each case is established by experienced pathologists. The digitization of breast tissue slides is achieved using an Olympus BX-50 system microscope equipped with a relay lens offering 3.3× magnification. This microscope is coupled with a Samsung digital color camera SCC-131AN, which utilizes a 1/3" Sony Super-HAD (Hole-Accumulation Diode) interline transfer charge-coupled device. The camera has a pixel size of 6.5 m × 6.25 m and a total pixel count of 752 × 582. Images are acquired in a three-channel red–green–blue (RGB) TrueColor format, featuring 24-bit color depth with 8 bits per color channel. Different magnification levels, namely 40×, 100×, 200×, and 400×, corresponding to objective lenses of 4×, 10×, 20×, and 40×, are used for image capture. The camera is set for automatic exposure, and manual focusing is carried out through the microscope while viewing the digital image on a computer screen.

The original images include black borders on both the left and right sides and text annotations in the upper left corner. These undesired areas are cropped from the resulting images, which are then saved in three-channel RGB format. Each channel has an 8-bit depth, and the images are stored in the portable network graphics format without compression. The dimensions of the resulting images are 700 × 460 pixels.

Initially, the pathologist identifies the tumor and defines a region of interest (ROI). To cover the entire ROI, multiple images are captured using the lowest magnification, i.e., 40×. The pathologist preferably selects images with a single type of tumor, although some images may contain transitional tissue, such as normal-pathological regions. On average, a total of 24 images per patient are captured from each slide using the lowest magnification.

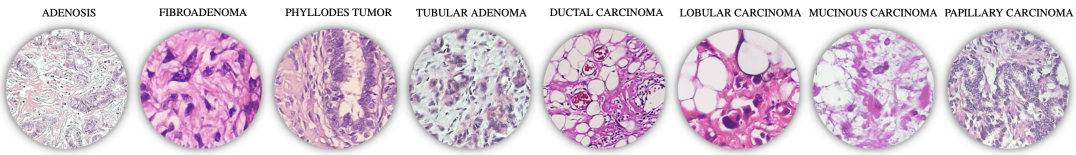


Figure 3.1: Sample images belong to 8 subgroups in BreakHis. From left to right, the first 4 images belong to the Adenosis (A), Fibroadenoma (F), Phyllodes Tumor (PT), and Tubular Adenoma (TA) subtypes, all of which come from the benign category. And the next 4 images extracted from the malignant category belong to the Ductal Carcinoma (DC), Lobular Carcinoma (LC), Mucinous Carcinoma (MC), and Papillary Carcinoma (PC) subclasses, respectively. (Image is taken from [31])

Subsequently, the magnification is manually increased to 100×, and a similar number of images are captured within the initial ROI. This process is repeated for 200× and 400× magnifications, respectively. Out-of-focus images are eliminated through a final visual (manual) inspection.

The dataset consists of 7,909 images, categorized into benign and malignant tumors. Both benign and malignant breast tumors are further classified into different types based on the appearance of tumoral cells under the microscope. The dataset consists of four histologically distinct types of benign breast tumors: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA). It also includes four malignant tumors (breast cancer): ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC).

Figure 3.1 provides examples of images from each subtype found within BreakHis.

3.1.2 Cytological Dataset

The dataset includes 500 cytology images acquired through Fine Needle Biopsy (FNB). This material came from 50 patients at the University Hospital in Zielona Góra, Poland. Biopsies were done without aspiration, guided by an ultrasonograph and a 0.5-mm-diameter needle. The smears from this material were fixed using by spray (Cellfix by Shandon) and

stained with hematoxylin and eosin [16]. 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. The time between preparing the smears and fixing them never exceeded three seconds. Afterward, these cytological preparations were turned into virtual slides using the Olympus VS120 Virtual Microscopy System [4]. It has a 2/3-inch charge-coupled device (CCD) camera and a 40X objective lens, resulting in a resolution of 0.172 μm/pixel. On average, the slides are around 200000 x 100000 pixels. Extended Focal Imaging (EFI) was used in scanning, where the preparation was scanned multiple times with different focus plane positions along the axis. These frames were merged, keeping only the sharply focused regions from each frame. This allowed for an extended focal depth impossible to achieve through optics alone. After this, a pathologist manually chose 11 distinct areas on each slide. These areas were converted into 8-bit/channel RGB TIFF files, each sized at 1583 x 828 pixels and compressed using the lossless LZW algorithm. It is important to note that the selection of areas was based on the quantity of cytological material, not medical information. The pathologists at the hospital recommended the number of areas per patient to ensure accurate diagnosis. The database comprises 25 benign cases (275 images) and 25 malignant cases (275 images), all of which were confirmed through histology. Patients with benign conditions either underwent biopsy or were monitored for a year. Figure 3.2 provides an overview of the acquisition process, while Figure 3.3 presents sample images of the 40X enlargement and digitally obtained 10X enlargement [10].

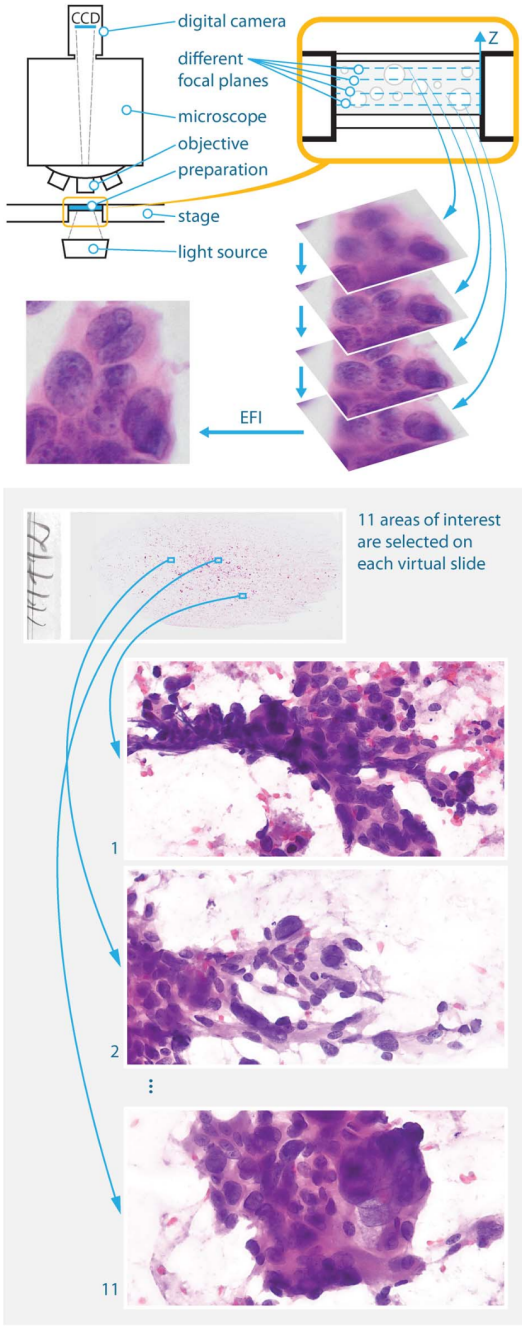


Figure 3.2: Scanning process using extended focal imaging (EFI) (top, figure not to the scale) and sample virtual slide with areas of interest selection (bottom) (image is taken from [10]).

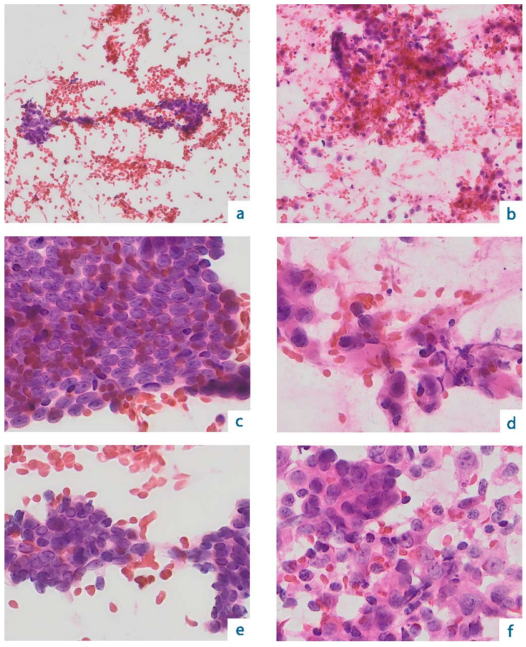


Figure 3.3: Sample images: digitally obtained 10X enlargement (a), (b) and 40X enlargement (c)–(f). Images a, c and e are from a benign case, and images b, d, and f are from a malignant case (image is taken from [10]).

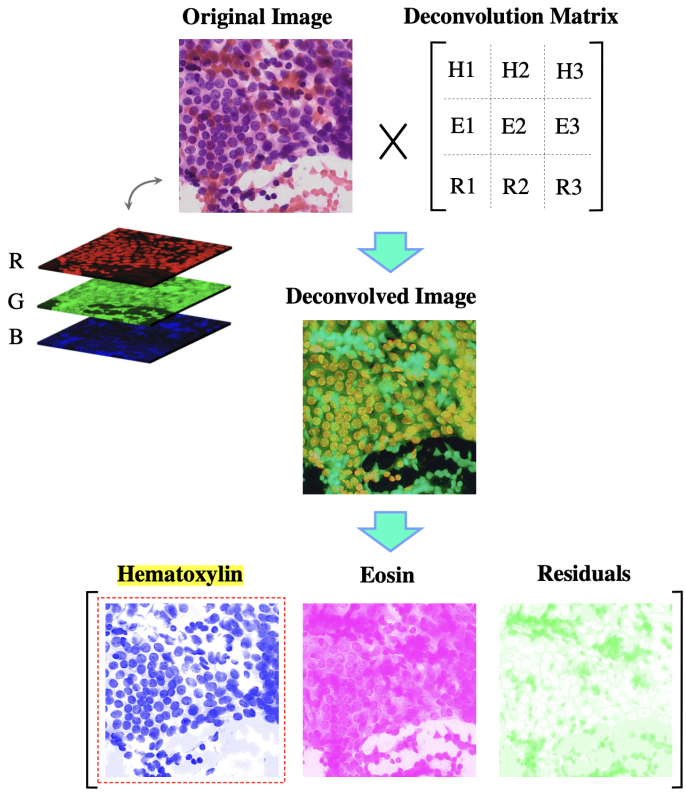


Figure 3.4: Representation of images obtained after performing H&E color deconvolution operation. The deconvolution matrix is first applied to the original image (RGB color channel). The color channels of the consequent image are then separated, resulting in Hematoxylin, Eosin, and Residuals images.

3.2 Segmentation

Shamshiri (2022) [31] recommended the use of minimum thresholding approach on cytological images before their analysis with computational models, highlighting its superior jaccard similarity index compared to other thresholding methods. Motivated by their approach, We included minimum thresholding into our research to assess its effectiveness in enhancing the performance of vision transformer models. This process entails deconvolving the original cytological image into three channels — hematoxylin, eosin, and residuals — using a specific deconvolution matrix. The focus is directed towards the hematoxylin-stained channel, which more prominently displays cell nuclei.

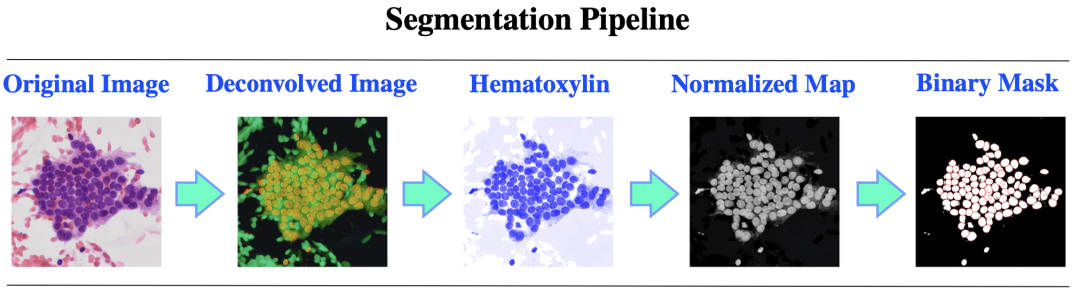


Figure 3.5: Segmentation pipeline (The image is taken from [31])

Following this segmentation, normalization is specifically applied to the Hematoxylin channel. Post-normalization, minimum thresholding is performed, preparing the data for model input and subsequent processing.

However, upon evaluating the model's metrics post-segmentation, a notable decrease in accuracy was observed. This led to the hypothesis that for complex models like vision transformers, extensive segmentation and preprocessing might be counterproductive, potentially eliminating crucial information. Hence, it may be more beneficial to allow these models to learn from unprocessed data, free from any biases. Nonetheless, we encourage other researchers to explore a variety of segmentation and preprocessing techniques to determine their potential benefits for similar models.

3.3 Building dataset and preprocessing

To make effective use of the transfer learning technique for addressing image analysis tasks, it is necessary to undergo two distinct phases for model training, each requiring its own dataset. The initial phase, referred to as pre-training, typically involves a large dataset and serves to initialize the model at a specific point within the parameter space. This initialization enhances the efficiency of the model optimization process [9]. Subsequently, in the following phase, the pre-trained model undergoes fine-tuning using the target dataset,

allowing it to adapt its features to the new data, potentially achieving significant improvements. The success of using a pre-trained network largely depends on the size of the pre-train dataset and the commonalities between the tasks or datasets involved. In our research, we utilized two distinct datasets for pretraining purposes: the ImageNet dataset and the BreakHis dataset. Each dataset offers unique benefits. The BreakHis dataset's texture closely resembles that of our cytological images, providing a more relevant training context. On the other hand, ImageNet is significantly larger and has been extensively trained by leading technology companies such as Google and Microsoft, leveraging substantial computing resources over an extended period. An additional advantage of ImageNet is the availability of pre-trained weights, eliminating the need for training from scratch. To fully leverage the strengths of both datasets, our strategy involved initial pretraining on ImageNet, followed by further pretraining using the BreakHis dataset.

3.3.1 Fine Tuning

Pre-train Dataset

Unlike many medical image analysis studies that utilize large-scale annotated natural images for the initial model pre-training phase, we will consider the use of breast cancer histopathological images as well. Additionally, we leverage the ViT and Swin Transformer, which is already pre-trained on the ImageNet dataset, to compare these two approaches and assess the impact of employing domain-compatible data on system performance. Generally, the choice of dataset for pre-training can come from related or unrelated domains. However, previous research has indicated that the greater the similarity between the two datasets, the more effective the pre-training ([3], [40]). However, due to the similarities in nature and structure between histopathological and cytological images [31] showed that the features extracted from them will be compatible when training on CNNs. To determine whether utilizing histopathological images as the pre-trained dataset, instead of Imagenet,

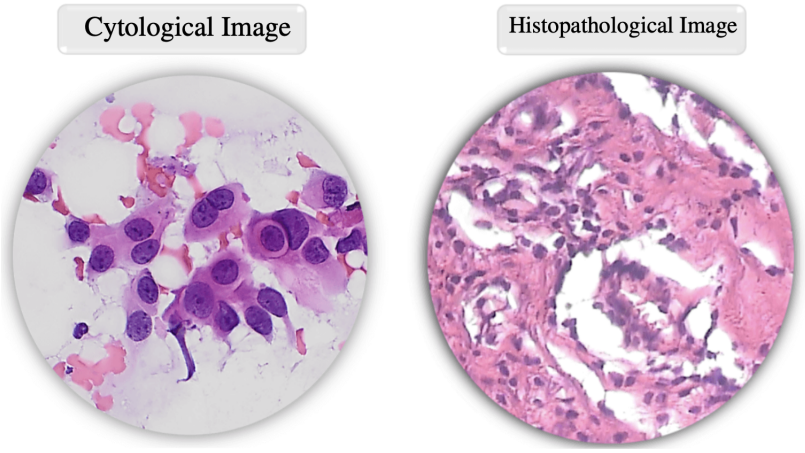


Figure 3.6: Sample of cytological and histopathological images. While cell nuclei in histopathological images usually do not have clear boundaries, cell clusters are clearly separated in cytological images. (Image is taken from [31])

will yield benefits when applying to transformers, we have no alternative but to experiment with it. Cytological image analysis is generally easier compared to examining histopathological specimens because of their unique characteristics. For instance, cytological images usually show well-separated cell clusters, and complex structures like glands are rarely observed in this image category. Conversely, histopathological images offer a more detailed view of diseases and their impact on tissues because they preserve the tissue's underlying structure during preparation. You can find sample images from both types in Figure 3.6. Thanks to the availability of publicly accessible histopathological images, acquiring the necessary images for our pre-training phase presents no difficulties. In our research, in order to classify cytological images we utilize one of the most widely recognized collections of histopathological breast cancer images, known as BreakHis [36], for our pre-training.

Target Dataset

In this research, we focused on a target dataset consisting of digital cytological images related to breast cancer. These images are sourced from patient samples collected at the University Hospital in Zielona Gora, Poland. Through the data acquisition process, we

assembled a dataset comprising 550 regions of interest (ROIs) from a total of 50 patients. Among these ROIs, 275 were associated with benign cases, and the remaining 275 were linked to malignant cases.

As mentioned previously, our target dataset comprises cytological breast cancer images from 50 patients, evenly distributed between benign and malignant cases (25 each). Additionally, for each patient, there are a total of 11 images included in the dataset. To assess the performance of our models, we divided the images from the 25 patients with benign cancer into three distinct sets for training, validation, and testing. Specifically, 15 patients' images (60%) were allocated to the training set, while the validation and test sets each contained images from 5 patients (20%). We followed a similar procedure for the images from the 25 malignant patients.

It is important to note that we have divided the dataset in such a way that each patient is only included in one subset. For instance, if an image from a patient is used in the training dataset, we cannot use other images of the same patient in the test or validation datasets. This is to avoid overfitting the model to a specific patient, which would not truly represent its ability to detect cancer

Image Augmentation

Since the number of images available for model training was limited, we chose to use image augmentation as a common technique to expand our image sets and address overfitting issues. To achieve this, we considered two approaches:

The first was to increase the number of images by applying augmentation techniques. For this purpose, we adopted a creative approach that combined various augmentation methods to maximize the benefits of generalizing the images. We considered eight augmentation techniques, including vertical flip, horizontal flip, vertical shift, horizontal shift, random scaling, random rotation, increased contrast, and shearing. Since our images are

microscopic images of cells, flipping the images vertically or horizontally does not alter the label of the image. Similarly, rotation preserves the information within the images. Small horizontal or vertical shifts, as well as slight zooming in or out, do not compromise the definition of the images and can be beneficial as augmentation techniques. Instead of applying these augmentations individually to a single original image, we have employed customized augmentation to fully enhance image generalization. We applied two augmentation techniques for each image in a way that it ensures we have used each technique at least once. The image augmentation procedure is detailed in Algorithm 1.

Algorithm 1 Image Augmentation

procedure AUGMENTATION(dataset)
    Initialize an empty list $augmented\_dataset$
    for each img in $dataset$ **do**
        Append img to $augmented\_dataset$
        Append $VFlip(RandomAugmentation(img))$ to $augmented\_dataset$
        Append $HFlip(RandomAugmentation(img))$ to $augmented\_dataset$
        Append $VShift(RandomAugmentation(img))$ to $augmented\_dataset$
        Append $HShift(RandomAugmentation(img))$ to $augmented\_dataset$
        Append $Scaling(RandomAugmentation(img))$ to $augmented\_dataset$
        Append $Rotation(RandomAugmentation(img))$ to $augmented\_dataset$
        Append $Contrast(RandomAugmentation(img))$ to $augmented\_dataset$
        Append $Shearing(RandomAugmentation(img))$ to $augmented\_dataset$
    end for
end procedure

By applying these augmentations, we expanded our datasets significantly. The cytological dataset consists of 4,950 images after augmentation, with 2,475 cases each for malignant and benign class.

In the figure below, you can observe the various augmented images generated from a single original image.

The second augmentation approach involved utilizing built-in PyTorch transforms and the T.Compose option. This method does not increase the total number of images. However, at the beginning of every epoch, it applies augmentations with a certain likelihood.

As a result, the images used to train the model in each epoch differ from those used in the previous one. The augmentations implemented included random horizontal and vertical flips with a 50% probability, rotations of up to 180 degrees, shifts up to 10% of the image's length and width, and zoom adjustments ranging from 0.8 to 1.2 times the original size. Furthermore, we explored two options to adapt the input image dimensions to suit the network requirements:

- Resizing Directly to the network input dimension: This method might distort the aspect ratio, which could lead to warped features in the images. For example, circular cells could become elliptical, which might affect the model's ability to learn accurate representations. However, it can be faster and simpler and maintains all the information of the image.

- Resizing to Maintain Aspect Ratio and Then Cropping: Maintaining the aspect ratio preserves the spatial relationships within the images. After resizing to the closest dimension, we could crop the center or use random cropping as an augmentation technique. This approach is generally preferred for medical images where the shape and structure of features (like cells or tissues) are important for diagnosis. Cropping can also be used to focus on the most informative parts of the image, which might be beneficial if the images contain a lot of background or irrelevant information.

We concluded that the second method of resizing, which prioritizes maintaining the aspect ratio before cropping, demonstrated superior performance compared to the direct resizing approach. This conclusion was drawn from consistent and enhanced results yielded in various experimental setups.

Furthermore, our analysis revealed that the second augmentation strategy significantly surpassed the first in terms of efficiency. This advantage is attributed to its faster processing, owing to the reduced number of images, while still providing a diverse range of

input images for each training epoch. Consequently, we adopted this second augmentation method for our thesis research.

Additionally, it is necessary to normalize the images using the mean and standard deviation of the dataset. The dataset mean for the BreakHis dataset is (0.79, 0.66, 0.76), and for the Cytological dataset, it is (0.78, 0.66, 0.77).

These values indicate that the two datasets are very similar, exhibiting nearly identical textures and colors. Moreover, when comparing these values to the ImageNet mean (approximately 0.5 in every channel), we realize that larger datasets tend to have means closer to the midpoint (0.5), as ImageNet is a vast dataset containing a diverse range of natural images. However, the datasets we are using consist only of microscopic images of cells and are smaller in size, resulting in means that are more biased and further from the midpoint. This discrepancy makes it more challenging to prevent overfitting and to train the model effectively.

The standard deviation for the BreakHis dataset is (0.14, 0.17, 0.12), and for the Cytological dataset, it is (0.14, 0.25, 0.13). Comparing these values to the standard deviation for ImageNet (approximately 0.25 in every channel), we find that our datasets have smaller standard deviations, indicating that the images are more homogeneous. This similarity among images makes training an effective model more difficult.

The image augmentation procedure is detailed in Algorithm 2.

Algorithm 2 Image Augmentation

```
1: procedure AUGMENTATION
2:     transforms.RandomHorizontalFlip(),
3:     transforms.RandomVerticalFlip(),
4:     transforms.RandomAffine(degrees=180, translate=(.1, .1), scale=(.8, 1.2)),
5:     transforms.Resize(RESHAPE_SIZE),
6:     transforms.RandomCrop(IMG_SIZE),
7:     transforms.ToTensor(),
8:     transforms.Normalize(DATASET_MEAN, DATASET_STD),
9: end procedure
```

Since the number of malignant and benign cases in the cytological dataset are similar, we do not need to perform downsampling [30].

3.4 Structures

3.4.1 Vision Transfomer (ViT)

Attention is all you need

In the context of a machine translation application, a transformer takes a sentence in one language and produces its translation in another. It consists of two primary components: an encoder and a decoder.

The encoding component comprises a stack of encoders, while the decoding component is a stack of decoders, both with the same number (six as per the original paper).

All encoders and decoders share an identical structure, though their weights are distinct. Each one consists of two and three sub-layers, respectively.

Now, we discuss how self-attention and encoder-decoder attention work.

In addition to the encoders and decoders, we need the word embeddings. The embedding only happens in the bottom-most encoder and decoder. The convention is that the encoders receive a list of vectors each of the size of the projection dimension – In the bottom encoder(decoder) that would be the word embeddings, however in other encoders(decoder), it would be the output of the component which is directly below.

Self-attention enables the model to examine other positions (words) in the input sequence to gather information for better encoding the current word.

First, it calculates the Query, Key, and Value matrices for self-attention. This involves arranging our embeddings into a matrix X and multiplying it by the trained weight matrices

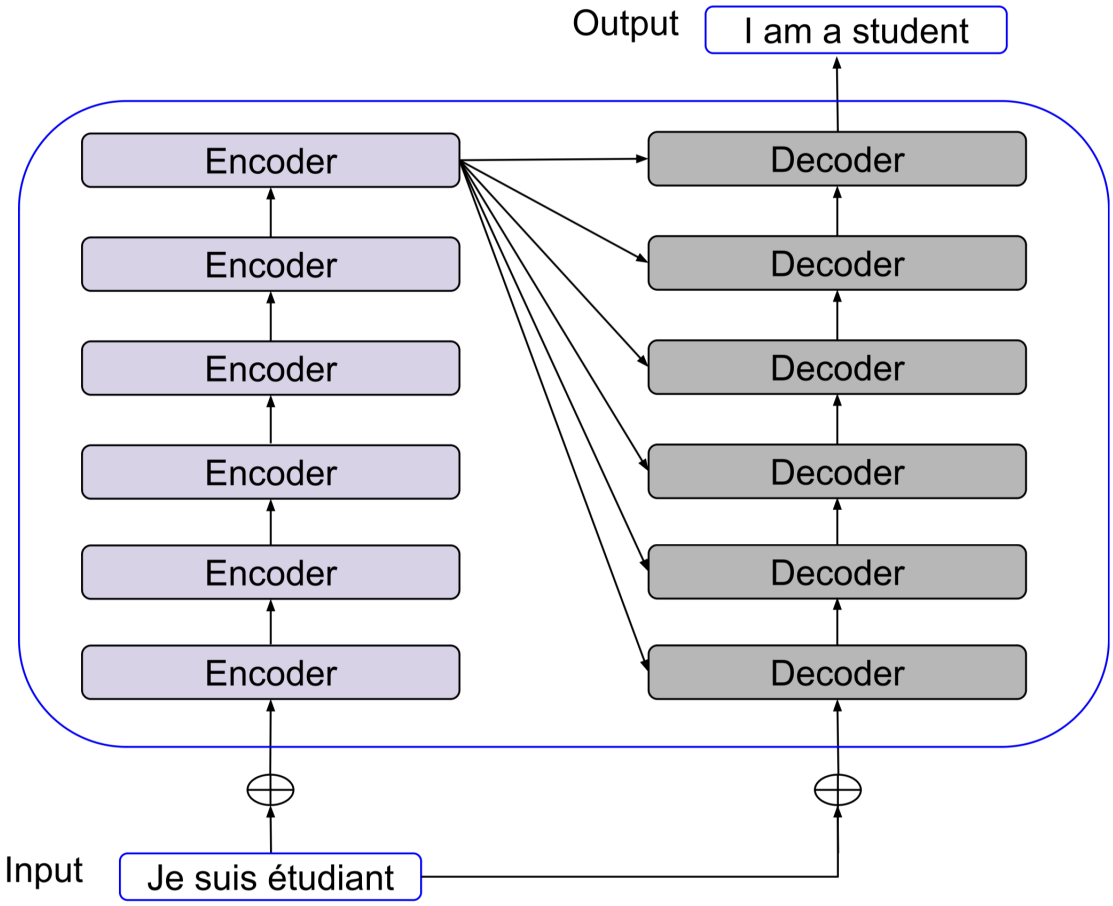


Figure 3.7

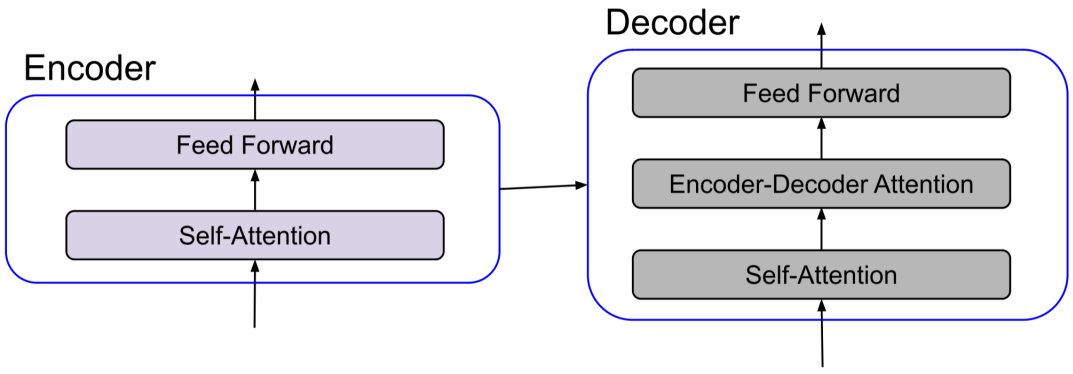


Figure 3.8

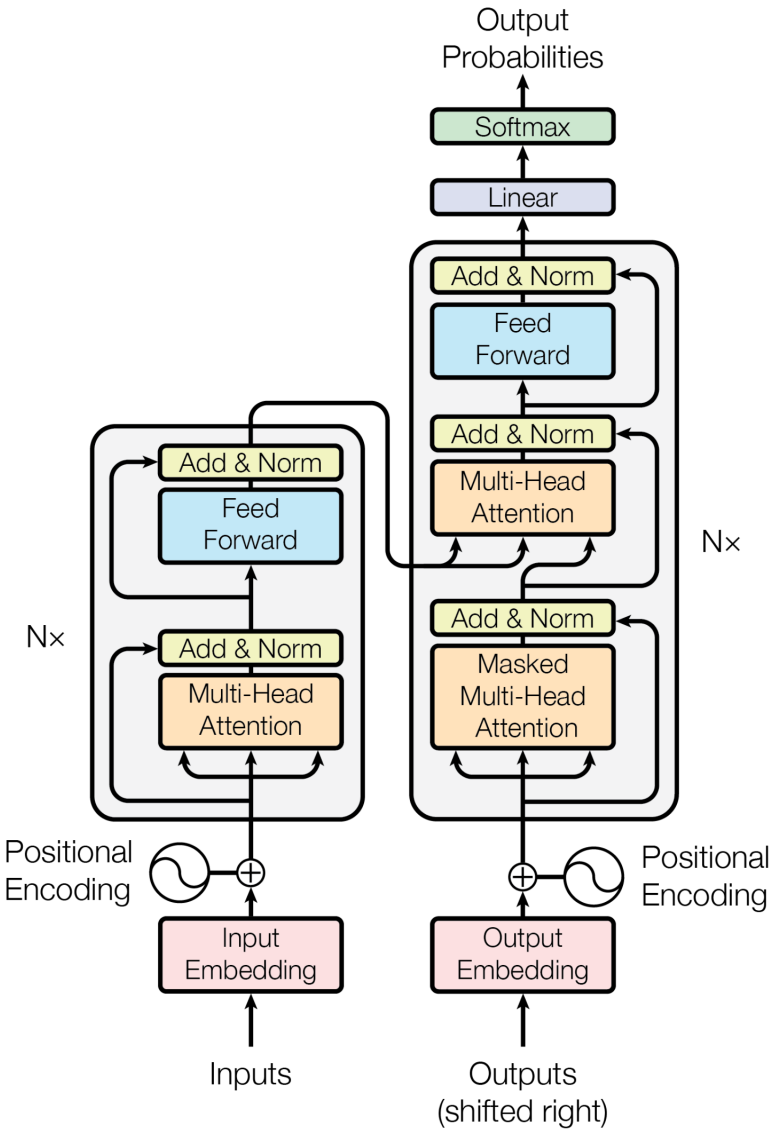


Figure 3.9: The transformer architecture (Image is taken from [39])

(W^Q, W^K, W^V).

$$Q = X\,W^Q$$
$$K = X\,W^K \tag{1}$$
$$V = X\,W^V$$

With the Q, K, and V matrices for a given input X (which may include three words in this example), the self-attention output will be computed using this formula:

$$Z = softmax(\frac{Q\,K^T}{\sqrt{d_k}})\,V \tag{2}$$

d_k is the dimension of the key vectors which is 64 for the original paper. The "multi-headed" attention repeats this process several times. The original paper employs 8 self-attentions in each multi-head attention step. This necessitates eight sets of matrices for W^Q, W^K, W^V, resulting in eight different Z matrices for the input sequence. These Z matrices are combined by concatenating all attention heads (Z_i) and then multiplying them with a jointly trained weight matrix W^O.

$$Z = [Z0 \quad Z1 \quad Z2 \quad Z3 \quad Z4 \quad Z5 \quad Z6 \quad Z7]\,W^O \tag{3}$$

The result is a Z matrix, which goes through an addition and normalization step, followed by a feedforward network, and another addition and normalization step. After these steps for each encoder, the output is ready to be fed into the decoder.

Now, we discuss how the decoder component operates.

Once again, we provide the decoder with the word embeddings. However, during training, we need to ensure that the decoder does not know the next words before predicting them. To achieve this, we apply masking along with the attention mechanism. A mask matrix is employed, which is an upper triangular matrix with entries above the main diagonal set to $-\infty$. This mask is added to $Q\,K^T$, known as the score, to ensure that the softmax

assigns a 0 probability to forbidden words. As a result, the attention weight for these words becomes 0.

The outcome of this masked multi-head attention is then forwarded to an addition and normalization component. Subsequently, the output is passed to another multi-head attention, this time using it as the Value matrix (V). This is where the output from previous encoders becomes relevant. We obtain the Q and K matrices from the last encoder by creating two copies of the last encoder's output through linear transformation. The outcome of this attention step undergoes addition and normalization, followed by a feedforward network, and another attention step. Finally, a softmax operation yields word probabilities, and the word with the highest probability is selected.

The encoder part of the Transformer architecture is executed once, while the decoders work to predict all the words in the target sentence until an "eos" token indicating the end of the sentence is generated. At this point, the model has predicted the target sentence [1].

An image is worth 16x16 words

The Vision Transformer (ViT) utilizes the encoding component of the transformer model. In this approach, an image is divided into non-overlapping patches, each of size $Patch\_size \times Patch\_size$. These patches are then flattened and linearly transformed using a trainable matrix E.

$$patch\_embedding = flattened\_patch * E \tag{4}$$

In this thesis, two distinct patch sizes were employed, namely $Patch\_size = 16$ and $Patch\_size = 32$. The selection of patch size is a trade-off: larger patch sizes increase the computational demand for generating patch embeddings but result in a reduced number of patches, thereby diminishing comparative analysis between patches. Ultimately, the choice of patch size depends on empirical insights and the specific requirements of the task at

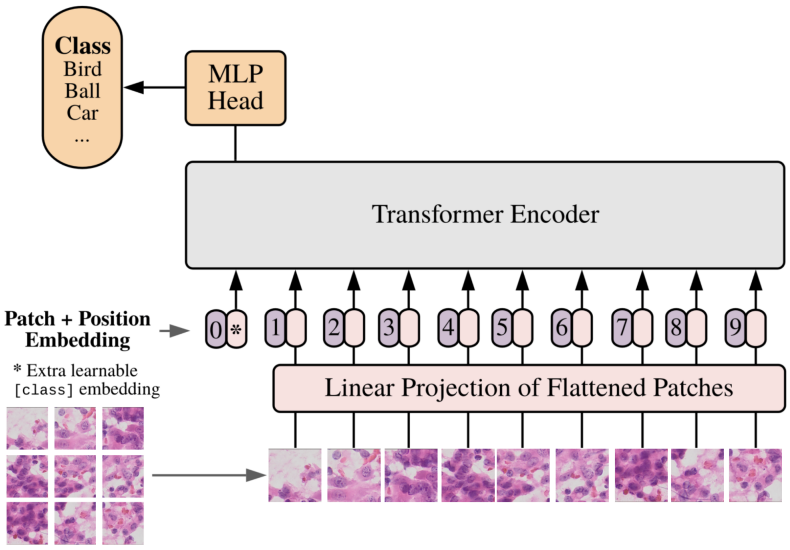


Figure 3.10: An example of dividing an image into 9 patches and feeding into ViT.

hand. The performance of the model can vary significantly based on these configurations, emphasizing the importance of experimental tuning to optimize for the desired application.

Subsequently, positional information is added to the patch embedding. Each patch is associated with a positional encoding. We add the corresponding patch embedding with its positional encoding and then feed it into the transformer encoder.

$$embedding\_vector = patch\_embedding + positional\_encoding \tag{5}$$

Additionally, instead of employing a traditional start token in natural language processing tasks, we utilize a learnable class token denoted as 'patch0'.

Finally, the output from the final encoder of the Vision Transformer includes information of all initial patches. Notably, the first patch, referred to as 'patch0', holds particular significance. This is due to its role as a representative summary of the entire image. Consequently, we focus exclusively on 'patch0'. This specific patch is passed through a final Multi-Layer Perceptron (MLP) head, which facilitates the classification into two distinct categories: benign and malignant.

3.4.2 Swin Transformer

The Swin Transformer is an innovative architecture in the field of computer vision, first introduced in [26]. The Swin Transformer's design addresses some of the limitations of applying standard Transformers to images, particularly regarding computational efficiency and scalability.

Key Concepts of Swin Transformer:

(1) Hierarchical Structure: Unlike standard Vision Transformers (ViTs) that flatten an image into a sequence of patches, Swin Transformers maintain a hierarchical structure similar to CNNs. This hierarchy allows for multi-scale processing, where the model can learn representations at various resolutions.

(2) Shifted Window-Based Self-Attention: The core idea is to compute self-attention within local windows (sub-regions of the image). However, unlike standard window-based approaches, Swin Transformers introduce a novel technique called "shifted window partitioning." In consecutive layers, the windows are shifted, ensuring cross-window connections and enabling the model to capture a broader context over layers.

(3) Efficient Computation: The local window-based approach reduces the computational complexity significantly compared to global self-attention used in standard Transformers. This makes Swin Transformers more scalable and efficient, especially for high-resolution images.

(4) Linear Complexity with Image Size: The computational complexity of Swin Transformers grows linearly with the image size, as opposed to the quadratic growth seen in standard Transformers. This is a significant advantage for processing large images.

Architecture and Layers: The model is illustrated in figure 3.11.

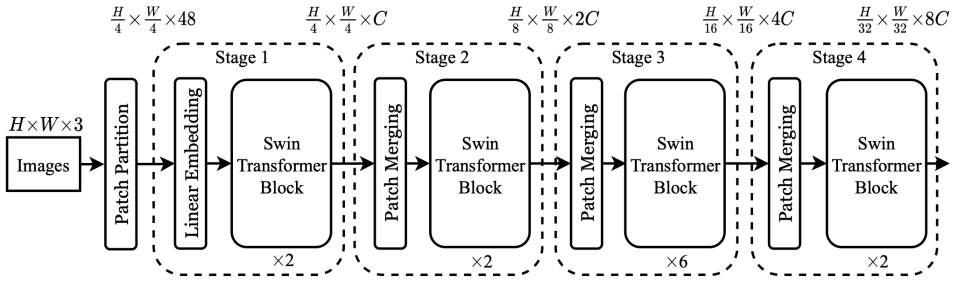


Figure 3.11: The architecture of a Swin Transformer (Swin-T) (Image is taken from [26]).

(1) Patch Partition: The input image with dimensions H x W x 3 is divided into non-overlapping patches. Each patch has a fixed size (e.g., 4x4 pixels), therefore every patch consist of 48 = 4 x 4 x 3 numbers. The patches are then flattened and linearly embedded to create a sequence of tokens with a dimension of C (number of channels).

(2) Linear Embedding: The flattened patches go through a linear embedding layer, which projects them into a higher-dimensional space suitable for the Transformer model. This is where the initial C features per patch are created.

(3) Swin Transformer Block: These blocks are the core of the Swin Transformer, where self-attention and MLP (multi-layer perceptron) operations are performed. Each block consists of a LayerNorm (LN), Multi-Head Self-Attention (MSA) module specific to Swin called ”Window MSA”, another LN, and an MLP with GELU non-linearity.

The architecture of a swin transformer block is illustrated in figure 3.12

- Layer Normalization (LN): Before any self-attention or MLP operation, Layer Normalization is applied. It normalizes the inputs across the features, stabilizing the learning process and accelerating convergence.
- Window-based Multi-head Self-Attention (W-MSA) / Shifted Window Multi-head Self-Attention (SW-MSA): The W-MSA module computes self-attention

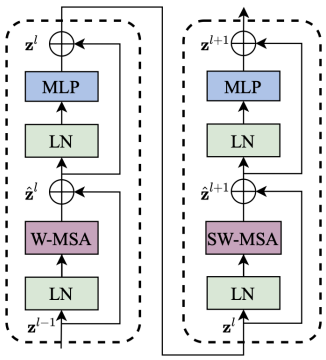


Figure 3.12: two successive Swin Transformer Blocks. W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively. (Image is taken from [26]).

within non-overlapping local windows (subregions of the image). The SW-MSA module is similar to W-MSA but operates on windows that are shifted by a certain number of pixels from the original windows. This shift allows for cross-window connection and broader receptive fields across the layers. The shifting operation alternates with each block, enabling the model to integrate local and global information efficiently.

- Residual Connections: After the self-attention and MLP operations, the results are added back to the input of each module (a residual connection). This skip-connection strategy helps in preventing the vanishing gradient problem and allows for deeper architectures.
- MLP (Multi-Layer Perceptron): Following each self-attention module, an MLP with two layers and a GELU non-linearity is used. It further processes the output of the self-attention mechanism. MLPs work on each token (patch representation) independently and can be seen as providing a channel-wise transformation.

(4) Stage 1 to Stage 4: The architecture processes the tokens through multiple stages,

each with a different number of Swin Transformer blocks. Each stage corresponds to a different resolution level in the hierarchy. As you move through the stages, the resolution decreases, while the feature dimensionality increases.

(5) Patch Merging: Between stages, "Patch Merging" layers are used to reduce the spatial dimensions and increase the number of feature channels. This is similar to downsampling in CNNs. It merges adjacent patches (2x2) into one patch, effectively reducing the number of patches by a factor of four (since each merge takes 4 adjacent patches) and doubling the feature dimensionality (since the features from the merged patches are concatenated).

3.4.3 Custom ViT

The implementation of Vision Transformers (ViTs) in computer vision tasks has demonstrated significant advancements in performance by leveraging self-attention mechanisms. Standard ViT architectures often require input images to be resized to a lower resolution, such as 224x224 pixels, to align with pre-trained model specifications. However, this resizing can result in a substantial loss of information, particularly for datasets with high-resolution images. In response to this limitation, a custom ViT architecture was developed specifically tailored to the unique requirements of the dataset used in this thesis. The proposed architecture deviates from the conventional ViT-base model by altering the input image size and patch resolution. Images are resized to 480x928 pixels, a dimension that preserves more information while maintaining an aspect ratio close to the original high-resolution images (1583x828 pixels). To further accommodate the increased image size, the patch size was expanded from the standard 16x16 to 32x32 pixels. This alteration allows the model to capture more detailed features at the expense of increased computational complexity. Keeping more details in the images might make the model work better, even though it uses more resources. A significant challenge introduced by these modifications

is the inability to utilize pre-trained weights from models trained on ImageNet. Since the model is custom-made and totally different from standard Vision Transformer models, we need to train it from scratch. Starting without pre-trained weights presents a considerable challenge, as it requires extensive computational resources and time. However, this approach is justified by the potential for the customized model to achieve superior performance on the specialized dataset, given that it is designed to process higher-resolution input without the information loss incurred by aggressive downscaling.

3.4.4 ViT-Swin Ensemble Model

To enhance the predictive accuracy of our system, as shown in figure 3.13 we are introducing an ensemble model that combines the capabilities of two advanced structures: ViT and the Swin Transformer. This ensemble model is engineered to outperform the individual models by harnessing their unique abilities to interpret different features of the dataset [5].

Each transformer model is trained separately to optimize its understanding of the data. Following training, the ensemble model uses a method called 'soft voting' to consider the predictions from both models. This approach allows the ensemble to evaluate which model is more certain about its prediction for each image.

The soft voting technique is a method that averages the probability predictions from each model for all possible categories. When the ensemble receives an image, both the ViT and Swin models calculate the probability of the image belonging to each category. These probabilities are then added together and averaged. The category with the highest average probability is chosen as the final prediction. This process ensures that the decision made by the ensemble reflects the most likely outcome as agreed upon by both models.

$$\hat{y} = \underset{i}{\operatorname{argmax}} \left\{ \frac{1}{N} \sum_{j=1}^{n} p_{ij} \right\} \tag{6}$$

where N is the number of classifiers, and p_{ij} is the probability value of j^{th} classifier for

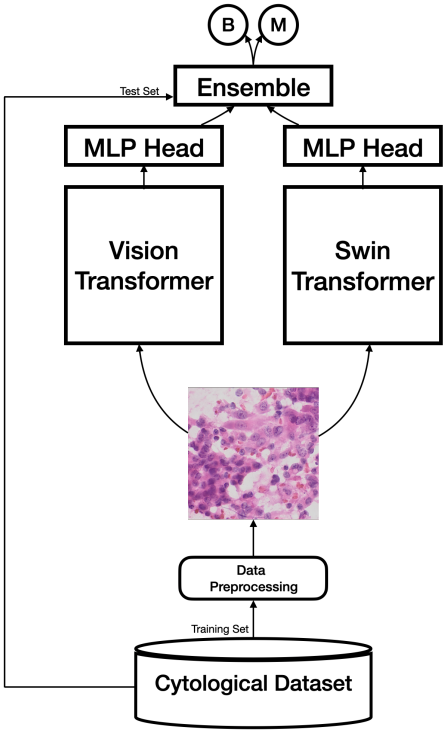


Figure 3.13: Structure of the proposed ViT–Swin ensemble model

the i^{th} category.

3.5 Fine tuning

To train the model, transformers require a large number of samples. To address this problem, a transfer learning technique is employed. With a few images, a pre-trained model derived from training on large datasets can be fine-tuned, thereby greatly reducing the training time. In addition to improving the convergence speed and generalization ability of the model, the transfer learning also reduces the risk of overfitting [2].

In this study, we will explore four different scenarios for training on cytological images.

In the first scenario, we train the model from scratch using cytological images without any fine-tuning. In the second scenario, we initialize the model with pre-trained weights from ImageNet and then fine-tune it on the cytological images.

For the third scenario, we start with training the model from scratch on the BreakHis dataset, which is a well-known histopathological breast cancer dataset. Afterward, we fine-tune it on the cytological dataset.

In the last scenario, we use the pre-trained model from ImageNet, fine-tune it on the BreakHis dataset, and then further fine-tune it on the cytological dataset.

Different frameworks and libraries have been utilized in this study, including but not limited to PyTorch, Keras, TensorFlow, and Hugging Face.

We have imported the pretrained Swin model by Microsoft and the ViT model by Google.

The final model is saved for the further use for other researches and also the code is publicly available here: github.com/rezajebeli97.

3.6 Summary

This chapter provided a comprehensive overview of the methodology employed in this thesis. The primary objective addressed is the classification of breast cancer microscopic images, which is achieved through a novel transfer learning (TL) framework. The proposed method envoloves several sequential steps, including image acquisition, dataset formation, network pre-training, weight initialization, and image classification. Each of these steps was extensively discussed within this chapter.

To assess the framework's effectiveness, real medical images obtained from biopsy specimens were employed. The process of preparing histopathological and cytological images, involving both traditional microscopy and a virtual microscopy system, is detailed. Furthermore, the chapter outlined the classification task involving categorizing these images into benign and malignant categories.

The subsequent chapter will present the experimental results obtained from exploring various training scenarios and fine-tuning strategies for the networks. These results will be

analyzed to determine the most effective approach.

Chapter 4

Experimental Results

4.1 Introduction

In this chapter, we present the experimental outcomes of our study to classify breast cancer microscopic images. These results highlight the effectiveness of our proposed method for this task. Our experiments primarily focus on accuracy, which is consistent with the evaluation criterion used in a previous state of the art study on the cytological dataset. However, we will provide other evaluation criteria in addition to accuracy. As mentioned earlier, this study explores different training scenarios for each dataset. We now discuss the results obtained from each scenario separately.

We also explain how we fine tune the models. Initially, fine-tuning was commonly applied to CNNs, where the initial layers learned fundamental image features useful for many computer vision tasks, while the later layers learned more task-specific features.

Typically, fine-tuning the last few layers suffices for transfer learning, and this approach is often used with transformers due to its computational efficiency. However, if the source and target tasks are significantly different, it may be necessary to fine-tune the earlier layers as well. Therefore, an effective fine-tuning strategy could be to start with the last layer and gradually include more layers in the updating process until the desired performance is

achieved.

4.2 Overview

In this study, various versions of ViT-base, Swin-base, ViT-small, and Swin-small were evaluated on datasets featuring diverse input dimensions and patch sizes. As detailed in table 4.1 (sorted by the number of parameters). It is evident that these models have significant differences in terms of the number of parameters. We also included DenseNet-169 as the last best performed model on the cytological dataset by Shamshiri (2022) [31], highlighting the importance of this study. Generally, transformers have less human bias, allowing the model to train without prior assumptions, enabling them to consider long-range dependencies. Consequently, this often leads to higher number of parameters.

Model	**DenseNet-169**	**ViT-Base**	**Swin-Base**	**ViT-Small**	**Swin-Small**
Parameters	14.3M	86M	88M	22.2M	50M

Table 4.1: Architectures' Details

Typically, a model with a higher number of learning parameters can extract more information from input data, but this also increases the risk of overfitting, particularly when the dataset is limited in size. To address this concern, we implemented data augmentation techniques to increase the dataset size. Additionally, we increased the drop_out rate and closely monitored both training and validation losses in each epoch, as well as other relevant metrics.

In the case of classifiction on cytological images, we considered two pre-training datasets: ImageNet and BreakHis, a histopathology image dataset. The ViT and Swin models were originally trained on the ImageNet dataset, which is a publicly available collection comprising over 14 million labeled images spanning 1000 different classes. The weights of these pre-trained models, obtained after training on ImageNet, are publicly accessible.

As previously mentioned, the BreakHis dataset includes two main categories: Benign and Malignant, each containing four subcategories. In our study, we selected Fibroadenoma from the benign category and Lobular Carcinoma from the malignant category due to their textural similarity to our cytological images. Furthermore, the BreakHis images are available in four magnification scales: 40x, 100x, 200x, and 400x. We chose to exclude the 40x magnification as it significantly differed from our images, retaining the 100x, 200x, and 400x scales. This resulted in a dataset of 761 benign images and 470 malignant images.

To ensure a balanced dataset, we performed downsampling, resulting in a collection of 940 images that included an equal number of both benign and malignant samples.

We applied the same augmentation techniques to both histological and cytological images. The primary difference lies in the dimension ratios between the two datasets. To accommodate this, we employed a distinct reshape size for each dataset, ensuring the maintenance of their aspect ratio.

For the cytology dataset, the data distribution included 20% for testing, 20% for validation, and 60% for training. In contrast, for the BreakHis dataset, we decided not to allocate a test set, allowing us to use more images for the model's pretraining phase. Table 4.2 details the exact number of images utilized for training across both datasets.

Dataset	**Dataset**	**Benign**	**Malignant**	**Total**
Cytological dataset	**Train**	165	165	330
	Test	55	55	110
	Validation	55	55	110
BreakHis	**Train**	376	376	752
	Test	-	-	-
	Validation	94	94	188

Table 4.2: Number of images of each dataset after preprocessing

The base transformers use an input size of either 224 × 224 or 384 × 384 pixels, following the standard for training on the ImageNet dataset. Therefore, the images have to be resized to these dimensions in the preprocessing step. Training was conducted for 100

epochs, and the embedding vector size for the ViT-base model was 768. After three months of hyperparameter tuning, we ended up selecting the hyperparameters shown in table 4.3. Details of experimental settings and parameters are also presented in table 4.3. All models were trained with the same hyperparameters.

Hyperparameter	Value
Number of epochs	100
Learning rate	10^{-6}
Train and Test batch size	32
Weight decay	0.001
Hidden activation function	GELU
Num attention heads	12
Num channels	3
Drop out Rate	0.3
Optimizer	Adam
BETA1	0.9
BETA2	0.999

Table 4.3: Hyperparameters

After evaluating how the model trains using the train and validation loss graph per epoch, we set 100 epochs as the most optimal value for the number of epochs of our model.

We experimented with different learning rates ranging from 10^{-3} to 10^{-9}, ultimately selecting 10^{-6} as the best choice.

Weight decay, also known as L2 regularization, is a technique used to prevent overfitting in machine learning. It adds a term to the loss function as shown in Equation 7 during training that encourages smaller weight values by penalizing large ones, where w_i are the weights of the model, and λ is the weight decay or regularization strength. This helps the model generalize better and reduces the risk of fitting training data too closely.

$$loss = loss + \lambda \sum_{i} (w_i^2) \tag{7}$$

The dropout rate refers to the proportion of neurons randomly turned off (i.e., set to zero) during the training process of a neural network. For example, a dropout rate of 0.3

means that there is a 30% chance that any given neuron will be turned off during a particular update cycle of the training phase. In our case, since the dataset is limited, the model leans toward overfitting, so it is important to choose a high value for weight decay to prevent overfitting; 0.3 was chosen at the end, which is considered a high value for weight decay.

The Adam optimizer is a popular gradient-based optimization algorithm used in training deep learning models. It combines aspects from AdaGrad and RMSProp optimizers, maintaining separate adaptive learning rates for each parameter. This is achieved through calculating exponential moving averages of the gradients and squared gradients.

Beta1 and Beta2 are hyperparameters for this optimizer. Beta1 controls the decay rate of these averages, and Beta2 adjusts the decay rate for the squared gradients, stabilizing the learning rate by prioritizing a longer history of gradients. We used the default values for these hyperparameters.

Encode strides refer to the number of pixels by which the patching operation moves when processing an image. For example, with an input size of 224x224 and an encoder stride of 16, patching will result in 196 (14x14) non-overlapping patches of 16x16 images. The hidden activation function specifies the activation function used in the model, which is GELU (Gaussian Error Linear Unit) in this case.

The setup ran on an A100 GPU High RAM in Google Colab Pro Plus.

4.3 Classification of Cytological Dataset

4.3.1 First Scenario

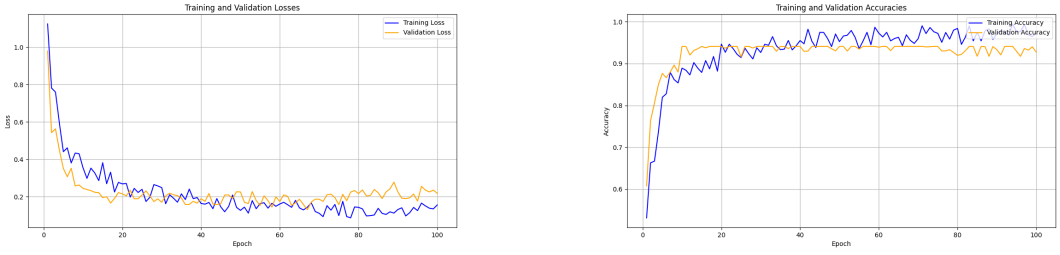
The first scenario was devoted to training transformers from scratch. The experiments performed, and the results obtained, are based on the test data set. The performance of the models are shown in the table 4.4.

Model	Input Dimension	Patch Size	Accuracy		Precision		Recall		F1-Score		ROC	
			Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test
Vit Base	**224 x 224**	16	91.81	87.69	92.38	85.35	92.81	84.43	92.3	85.87	92.16	87.29
		32	90.33	86.97	92.25	85.15	91.47	84.6	92.97	84.08	89.51	86.43
	384 x 384	16	93.28	88.07	92.6	88.88	93.67	**89.38**	93.36	88.41	93.84	87.35
		32	91.77	88.38	90.46	86.59	93.16	86.95	90.41	87.5	92.02	85.53
Swin Base	**224 x 224**	16	91.76	86.1	88.44	84.06	89.77	84.94	88.77	84.19	88.02	85.72
		32	88.9	85.81	88.53	83.83	89.96	83.02	88.67	85.89	91.48	83.12
	384 x 384	16	94.18	85.74	93.67	86.55	91.13	87.26	91.92	85.32	91.48	87.51
		32	91.38	86.13	92.26	84.72	92.44	88.25	91.87	85.21	90.76	87.36
ViT-Swin Base Ensemble	**224 x 224**	16	90.51	88.13	91.11	87.81	89.8	88.01	90.2	85.83	93.12	87.36
		32	93.19	84.88	92.52	87.47	90.85	87.91	92.41	86.15	92.29	84.51
	384 x 384	16	**95.39**	**89.56**	**95.41**	**89.23**	92.69	87.72	**94.55**	88.19	92.21	86.3
		32	91.47	86.1	93.85	89.37	92.06	88.7	91.01	**88.94**	**94.22**	87.68
Vit Small	**224 x 224**	16	88.08	86.16	90.47	83.87	90.45	83.63	91.6	84.54	90.75	86.4
		32	91.79	84.63	88.2	84.74	90.95	83.7	91.98	83.36	91.37	83.74
	384 x 384	16	92.45	84.61	91.1	86.32	92.93	84.76	93.25	87.19	93.02	87.65
		32	92.99	86.19	92.3	85.13	89.56	84.99	89.94	86.15	90.77	84.58
Swin Small	**224 x 224**	16	88.02	85.76	89.41	82.89	87.31	83.78	87.44	85.82	90.78	85.83
		32	90.36	82.81	87.71	82.49	89.53	84.35	88.73	84.87	88.85	82.52
	384 x 384	16	91.73	87.47	90.85	87.34	91.51	87.23	92.32	85.21	89.73	87.24
		32	92.27	84.14	91.06	85.62	89.3	84.17	90.22	86.12	90.65	85.97
ViT-Swin Small Ensemble	**224 x 224**	16	90.26	84.94	90.14	84.13	91.98	83.55	91.82	84.8	88.85	87.01
		32	89.86	84.45	89.37	85.09	88.7	83.2	89.77	83.81	91.62	85.82
	384 x 384	16	93.81	87.15	92.57	86.67	**94.26**	88.94	91.33	86.51	92.15	87.52
		32	90.45	84.78	90.92	87.87	92.41	88.49	91.1	86.52	90.22	87.97
Custom ViT	**480x912**	32	91.37	86.36	94.16	88.52	93.07	86.54	93.27	88.19	93.47	**89.84**

Table 4.4: Scenario1's Performance (Bold numbers represent the highest value for the column).

4.3.2 Second Scenario

Now, we look at the outcomes of the second scenario, where we used pre-trained networks from ImageNet. These networks had prior training on ImageNet, and we applied them to perform cytological image classification. You can find a summary of the results in the table 4.5. Comparing the results to the previous scenario, you can observe the significant benefits of fine-tuning for the model.



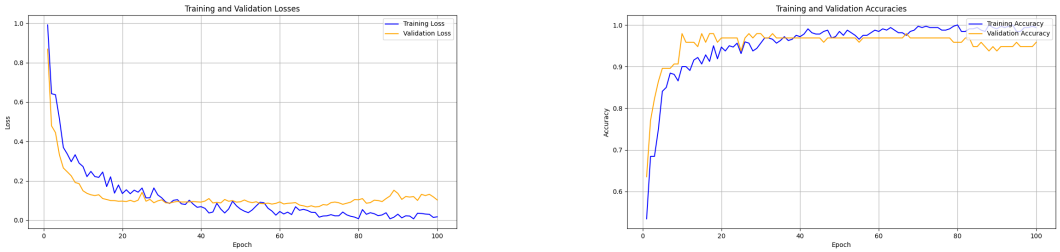
(a) Training and Validation Loss Over Epochs

(b) Training and Validation Accuracy Over Epochs

Figure 4.1: Scenario1: Graphs for Swin-base model with 384x384 input dimension and a 16x16 patch size

Model	Input Dimension	Patch Size	Accuracy		Precision		Recall		F1-Score		ROC	
			Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test
Vit Base	**224 x 224**	16	95.16	91.28	96.22	91.53	95.31	91.51	95.3	91.53	95.96	90.23
		32	95.83	90.64	95.15	90.55	96.39	91.18	95.4	90.02	95.8	91.11
	384 x 384	16	97.33	91.94	97.58	**92.89**	**98.92**	93.07	97.05	92.73	97.93	93.38
		32	97.33	91.89	96.22	91.8	96.73	92.37	97.23	92.86	97.08	92.39
Swin Base	**224 x 224**	16	95.26	91.11	95.68	89.64	94.75	90.95	94.92	89.62	95.84	89.86
		32	95.89	89.77	94.52	90.23	95.07	90.73	95.62	89.91	95.85	90.01
	384 x 384	16	97.3	91.91	96.97	91.11	96.97	91.56	97.56	92.08	98.03	91.65
		32	97.1	91.81	96.55	92.07	95.99	92.27	95.91	91.93	95.67	91.04
ViT-Swin Base Ensemble	**224 x 224**	16	96.16	91.48	95.77	91.76	96.54	91.63	96.01	91.95	95.56	90.98
		32	95.89	91.9	95.59	90.65	95.91	91.25	96.95	91.53	96.22	91.84
	384 x 384	16	**98.36**	92.57	**98.77**	92.3	97.67	**93.99**	**98.66**	**93.27**	**98.61**	**93.89**
		32	96.76	**92.7**	97.94	92.44	98.13	93.37	96.68	92.31	98.42	92.59
Vit Small	**224 x 224**	16	94.17	89.96	94.76	90.94	94.71	89.42	95.26	90.68	94.6	89.1
		32	94.57	89.95	94.91	88.98	94.89	89.65	94.84	89.71	94.87	88.56
	384 x 384	16	97.86	92.5	97.35	91.44	96.05	91.92	97.75	91.18	96.63	91.39
		32	96.89	91.22	95.44	90.94	95.87	91.26	95.71	90.95	95.11	90.93
Swin Small	**224 x 224**	16	93.63	89.5	93.87	88.78	94.7	90.18	93.63	90.08	94.66	90.02
		32	94.23	88.16	94.05	88.56	93.66	89.49	94.8	88.83	94.81	88.66
	384 x 384	16	96.92	91.05	95.74	91.9	96.26	90.71	97.38	90.82	96.93	90.99
		32	96.35	90.73	94.99	89.79	95.44	90.58	94.81	89.71	94.9	91.45
ViT-Swin Small Ensemble	**224 x 224**	16	94.75	90.8	95.58	90.77	94.84	90.96	94.58	91.25	95.66	91.13
		32	94.77	90.33	95.98	89.49	95.71	89.08	96.25	89.14	94.72	90.14
	384 x 384	16	97.57	92.5	96.97	91.12	97.34	91.33	97.72	91.91	96.64	91.59
		32	96.28	91.43	95.81	91.61	96.68	92.21	96.06	91.51	97.13	91.14

Table 4.5: Scenario2's Performance



(a) Training and Validation Loss Over Epochs

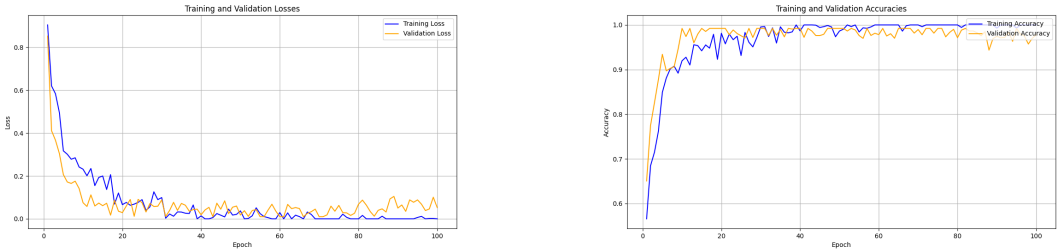
(b) Training and Validation Accuracy Over Epochs

Figure 4.2: Scenario2: Graphs for ViT-base model with 384x384 input dimension and a 16x16 patch size

4.3.3 Third Scenario

The third scenario aims to investigate how well transformers perform when using transfer learning on a dataset of related medical images called BreakHis.

We train the model using histo from scratch. After pre-training, we stored the model weights for initializing the networks in the next phase. Then the training is performed on the cytological images from data augmentation resulting in 4,950 images. We do not need to perform downsampling since the number of images for both classes are similar. Then, we have divided into training, validation, and test sets for binary classification. The models'



(a) Training and Validation Loss Over Epochs

(b) Training and Validation Accuracy Over Epochs

Figure 4.3: Scenario3: Graphs for Swin-base model with 384x384 input dimension and a 16x16 patch size

success rates in each scenario were compared.

Model	Input Dimension	Patch Size	Accuracy		Precision		Recall		F1-Score		ROC	
			Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test
Vit Base	**224 x 224**	16	97.21	91.5	95.7	89.12	97.47	92.87	97.64	89.0	96.28	92.89
		32	95.79	89.09	95.37	89.04	97.08	92.23	94.97	89.54	95.67	91.33
	384 x 384	16	96.05	91.34	**99.27**	92.85	98.42	**93.21**	**99.61**	92.7	96.3	90.99
		32	98.47	92.25	96.04	91.82	97.96	90.63	97.08	92.55	98.94	91.69
Swin Base	**224 x 224**	16	94.67	89.79	95.1	90.64	93.16	91.63	93.75	88.68	94.73	92.47
		32	96.59	88.1	95.32	91.82	96.06	89.22	94.25	88.78	97.39	88.22
	384 x 384	16	**99.22**	93.11	98.11	**93.34**	**98.48**	91.81	99.27	92.53	97.29	90.37
		32	94.83	90.69	97.97	92.27	95.33	90.39	95.86	92.49	98.2	91.26
ViT-Swin Base Ensemble	**224 x 224**	16	95.62	91.33	95.8	90.81	97.91	90.45	95.14	**93.42**	97.91	91.03
		32	95.94	89.01	98.46	92.31	95.35	90.85	94.5	91.27	98.2	91.11
	384 x 384	16	97.15	92.89	98.6	91.11	96.56	92.83	98.62	92.38	98.1	91.69
		32	96.0	**93.83**	96.06	92.11	95.93	90.82	95.88	93.22	96.3	92.89
Vit Small	**224 x 224**	16	95.27	88.23	96.95	89.1	95.27	91.44	93.2	90.43	95.76	89.66
		32	95.75	88.71	96.78	89.81	96.65	89.67	94.15	89.39	94.11	91.11
	384 x 384	16	95.39	92.23	97.27	90.85	96.89	90.59	95.78	92.46	97.59	**93.47**
		32	96.82	89.22	96.33	92.41	97.02	89.08	95.24	90.67	95.66	90.98
Swin Small	**224 x 224**	16	93.32	89.1	95.1	87.83	92.18	90.65	93.44	88.02	95.65	90.65
		32	92.5	89.48	95.76	89.04	94.31	90.9	96.08	89.43	93.59	88.2
	384 x 384	16	98.11	92.58	94.81	92.41	94.67	92.14	97.17	90.49	94.82	89.93
		32	94.33	89.64	95.35	90.56	97.36	90.83	94.02	88.98	97.34	92.46
ViT-Swin Small Ensemble	**224 x 224**	16	96.16	88.91	94.47	90.65	95.27	91.45	95.17	91.72	95.39	88.79
		32	95.3	89.84	96.44	90.24	96.6	91.67	96.89	91.92	94.3	89.25
	384 x 384	16	95.95	91.92	96.03	90.4	95.78	90.91	96.34	90.58	**99.47**	93.16
		32	96.51	91.31	97.89	92.95	94.79	90.53	97.08	92.13	97.85	92.78
Custom ViT	**480x912**	32	97.84	89.08	95.07	90.71	97.49	92.11	95.11	91.76	96.93	90.62

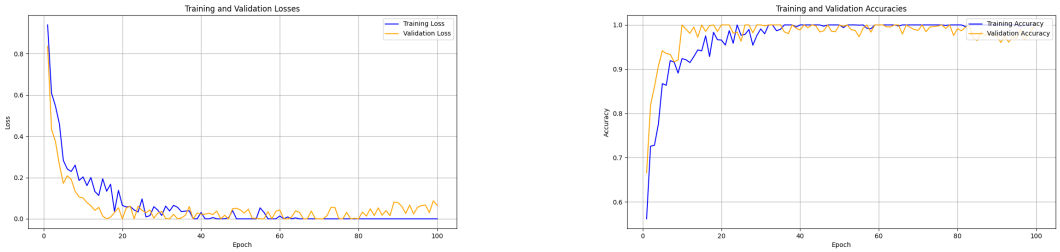
Table 4.6: Scenario3's Performance

4.3.4 Fourth Scenario

The fourth scenario was devoted to using pretrained transformers on Imagenet to train on breakHis dataset, and then fine tune the models on Cytology dataset. The experiments performed, and the results obtained, are based on the test data sets.

Model	Input Dimension	Patch Size	Accuracy		Precision		Recall		F1-Score		ROC	
			Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test
Vit Base	**224 x 224**	16	98.03	91.48	97.87	92.3	97.36	91.81	97.16	92.52	98.63	93.66
		32	97.77	93.0	96.19	92.57	96.37	91.29	98.4	91.27	98.65	92.19
	384 x 384	16	99.9	94.92	98.32	94.1	**100**	93.81	**100**	92.62	98.83	92.78
		32	97.62	93.44	99.69	93.95	99.56	92.95	97.99	92.09	97.8	92.27
Swin Base	**224 x 224**	16	96.01	91.72	96.31	93.08	95.79	93.38	95.55	91.6	96.06	91.13
		32	96.37	90.38	95.86	91.91	97.5	92.21	98.15	92.63	97.51	90.07
	384 x 384	16	**100**	95.04	99.33	95.04	98.86	94.84	**100**	94.76	97.61	**94.73**
		32	97.37	92.63	97.01	93.5	97.04	93.49	97.18	93.92	99.21	93.39
ViT-Swin Base Ensemble	**224 x 224**	16	98.84	94.16	97.33	91.58	98.2	91.75	96.85	92.83	96.94	93.48
		32	97.09	92.41	98.36	93.3	98.46	93.68	98.33	93.6	98.63	92.65
	384 x 384	16	**100**	**95.37**	98.8	**96.04**	98.7	**95.05**	98.56	**95.75**	**100**	93.66
		32	98.75	94.73	**99.91**	94.19	99.0	94.16	99.88	93.68	**100**	93.56
Vit Small	**224 x 224**	16	95.96	91.0	96.0	90.08	96.07	90.04	96.27	92.44	97.19	91.35
		32	95.38	91.69	95.53	89.81	96.87	89.74	96.56	92.49	97.51	92.16
	384 x 384	16	97.21	92.28	99.23	94.1	98.95	92.94	99.0	92.54	99.69	94.48
		32	97.9	91.64	97.07	92.01	98.04	91.54	98.26	92.64	97.18	91.41
Swin Small	**224 x 224**	16	94.41	89.56	95.03	89.58	94.57	89.82	94.24	92.1	96.11	92.37
		32	95.21	90.43	95.84	91.76	97.1	90.05	97.49	89.58	96.37	90.46
	384 x 384	16	98.8	92.9	97.87	92.05	97.87	91.16	98.49	91.2	97.41	91.93
		32	97.4	91.05	96.03	93.39	96.69	92.27	96.04	90.53	96.22	91.1
ViT-Swin Small Ensemble	**224 x 224**	16	97.45	92.19	97.97	93.01	97.61	90.82	95.65	91.69	96.08	92.42
		32	97.64	91.09	97.96	90.2	96.11	90.95	96.46	92.9	98.38	91.2
	384 x 384	16	98.29	92.96	98.23	93.76	99.2	93.82	99.26	94.47	97.66	94.54
		32	99.34	93.1	98.36	91.75	98.77	93.73	97.16	93.61	97.72	92.04

Table 4.7: Scenario4's Performance



(a) Training and Validation Loss Over Epochs

(b) Training and Validation Accuracy Over Epochs

Figure 4.4: Scenario4: Graphs for Swin-base model with 384x384 input dimension and a 16x16 patch size

4.4 Cross Validation

Given the extensive variety of models explored in this study—including 98 distinct configurations with varying patch sizes and input dimensions—and considering the substantial computational time required per model, undertaking k-fold cross-validation across all models proved impractical. Consequently, the decision was made to apply 5-fold cross-validation exclusively to the highest-performing model, specifically the ensemble ViT-Swin base model with a patch size of 16 and an input dimension of 384.

Figure 4.5: 5 fold cross validation

4.4.1 5 fold cross validation including test and validation sets

5-fold cross-validation involves dividing the dataset into five equal parts, or ”folds”. In each round of validation, one fold is reserved for the test set, and the remaining four folds act as the training and validation sets. In each round, one fold out of the remaining four will be randomly chosen as the validation fold. We ensure that each fold is used only once as the validation set. Therefore, once a fold has been used as the validation set, it will be removed from the options for potential validation sets in subsequent rounds. This process is repeated five times, with each fold being the validation set once. The final performance metric is obtained by averaging the results from all five rounds, providing a comprehensive evaluation of the model's performance across the entire dataset. Figure 4.5 illustrates how our approach to 5-fold cross-validation works.

Therefore, to ensure comprehensive evaluation, the dataset was partitioned into five equal segments, each segment including exactly 10 patients, 5 benign and 5 malignant, with care taken to allocate images from the same patient exclusively to a single segment. This approach allows for each segment to serve once as the validation set while the remaining are utilized for training, thereby rotating through all possible training-validation set combinations. This methodology affords a robust measure of the model's generalizability

across diverse subsets of the data, as opposed to relying on a singular predetermined training set. You can observe the model's performance across different folds and the average in Table 4.8.

Model		**ViT Base**	**Swin Base**	**ViT-Swin Base Ensemble**
Input Dimension		384 x 384	384 x 384	384 x 384
Patch Size		16	16	16
Accuracy	Fold1	94.92	93.61	95.87
	Fold2	90.60	94.93	94.70
	Fold3	98.12	94.35	96.11
	Fold4	95.23	91.97	93.43
	Fold5	96.71	92.42	94.95
	Average	95.11	93.45	95.01
Precision	Fold1	94.1	93.34	96.04
	Fold2	91.74	90.83	95.74
	Fold3	92.58	91.02	98.43
	Fold4	95.29	93.90	92.66
	Fold5	94.79	95.40	95.16
	Average	93.7	92.89	95.60
Recall	Fold1	94.31	92.31	95.55
	Fold2	90.8	95.15	91.27
	Fold3	91.4	94.34	99.32
	Fold4	96.8	95.11	95.07
	Fold5	97.8	96.11	96.07
	Average	94.22	94.60	95.45
F1-Score	Fold1	93.2	93.03	96.25
	Fold2	88.42	89.33	91.45
	Fold3	95.46	96.16	99.66
	Fold4	89.09	89.39	95.25
	Fold5	91.62	92.32	94.61
	Average	91.55	92.04	95.44
AUC	Fold1	91.49	91.87	94.16
	Fold2	94.07	89.43	96.43
	Fold3	89.69	92.52	95.73
	Fold4	91.84	86.74	91.21
	Fold4	92.11	90.17	93.91
	Average	91.84	90.14	94.28

Table 4.8: Model's Performance on test set using 5 fold cross validation

4.5 Attention Map Visualizations Across Model Layers

In the domain of deep learning, understanding the focus of a neural network, particularly in vision transformers, is crucial to interpreting the model's decision-making process. Attention maps provide a window into the inner workings of these models, revealing which regions of the input data are prioritized during the learning and inference stages. This section delves into the attention map visualizations of our top-performing vision transformer model, ViT-Base with a 384x384 input dimension and a 32x32 patch size, as outlined in the fourth scenario. We examine how the model's attention evolves and refines across its various layers.

4.5.1 Introduction

Vision transformers leverage self-attention mechanisms, a concept borrowed from natural language processing. Unlike Convolutional Neural Networks (CNNs), ViTs divide an image into patches and learn to focus on the most informative parts of these patches to make predictions. Attention maps serve as a powerful tool for model interpretability and diagnosis. They not only offer insights into the decision-making process but also aid in validating the model's focus against human intuition. For instance, in medical imaging, ensuring that a model's attention aligns with clinically relevant features is paramount. The visualization across layers further enhances our understanding by showing how attention evolves with depth, enabling us to fine-tune or debug the model with precision.

4.5.2 Understanding Attention Maps

An attention map is a visual representation that highlights the areas within an image that a model pays attention to when making predictions. These maps are particularly enlightening for vision transformers that consist of multiple self-attention layers. Each layer

can be thought of as an observer with a different lens, focusing on various features of the image—from edges and textures in early layers to more abstract and complex patterns in deeper ones.

4.5.3 Visualization Technique

The visualization of attention maps is achieved through overlaying heatmaps onto the original image. The intensity of colors in the heatmap corresponds to the magnitude of attention, with warmer colors indicating areas of higher focus. This method allows us to interpret the model's behavior visually.

4.5.4 Layer-wise Analysis

A vision transformer model comprises several layers, each contributing to the feature extraction process. By visualizing the attention maps of each layer, we can trace the model's attention trajectory as follows:

- Early Layers (1-4): The initial layers tend to focus on low-level features such as edges, corners, and basic textures. Attention in these layers is often scattered and lacks specificity, as the model is in the initial stages of understanding the image structure.

- Middle Layers (5-8): As we progress through the layers, the attention maps begin to coalesce around more defined areas of interest. These layers start interpreting shapes and object parts, transitioning from raw pixel data to a more semantic understanding.

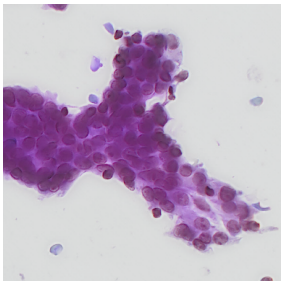
- Later Layers (9-12): In the final layers, attention maps often display a highly focused pattern. The model has learned to hone in on the most critical regions for classification or regression tasks. It is here that the attention maps resonate closely with the salient features of the target objects or areas within the image.

In Figure 4.6, we present the attention map visualizations of the twelve distinct layers of our top-performing vision transformer model, ViT-Base with a 384x384 input dimension and a 32x32 patch size, as outlined in the fourth scenario, on a random image.

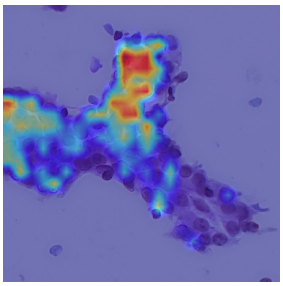
The layer-by-layer visualization of attention maps provides a narrative of the model's learning progression. Early layers resemble the curiosity of a child, observing everything with equal interest. As maturity sets in, akin to a learned scholar, the model discerns what is and isn't worthy of attention, thereby refining its predictions. The attention maps revealed that the network primarily focused on cell nuclei features, aligning with medical knowledge.

4.6 Discussion

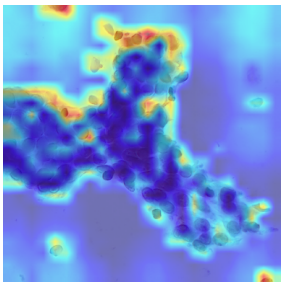
Vision transformers are a recent development in computer vision. Training them usually requires a lot of data, which can be challenging, especially for medical images. Past studies on the same datasets have mainly used CNNs and other machine learning methods due to this limitation or not availability of vision transformers by the time. However, given the recent advancements of transformers in outperforming CNNs, we chose to use vision transformers despite the challenges. The use of transfer learning made it possible to train transformers even with limited data. We also explored various methods to improve the performance of vision transformers in this context. Training deep networks from scratch for medical image classification is challenging due to limited samples. This is the first study on vision transformers on the cytological dataset, and we have demonstrated that, through transfer learning, we can achieve the best performance, outperforming all the previous models. Additionally, We considered pretraining on the BreakHis dataset, a compatible dataset, to train the model on the cytology dataset as Shamshiri (2022) [31] found it useful. We also could improve the model's performance by pretraining on breakhis after being pretrained on imagenet, otherwise only pretraining on breakhis from scratch worsen



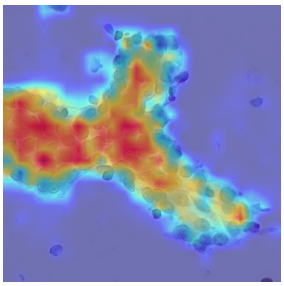
(a) Original image



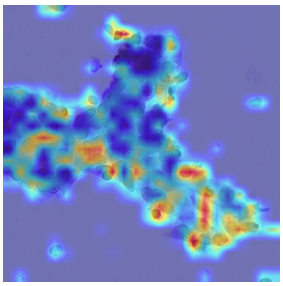
(b) Layer 1



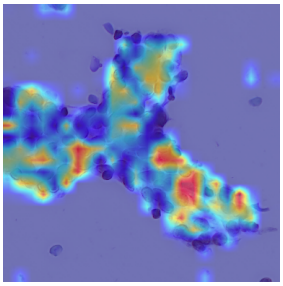
(c) Layer 2



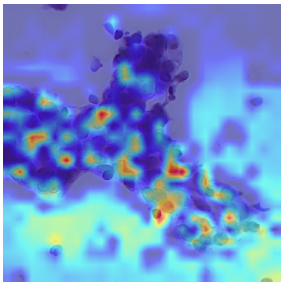
(d) Layer 3



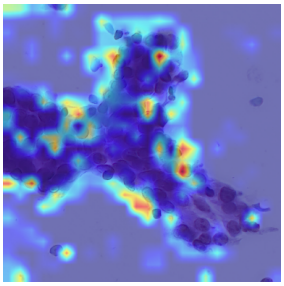
(e) Layer 4



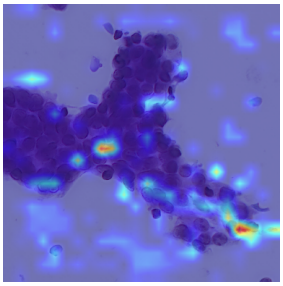
(f) Layer 5



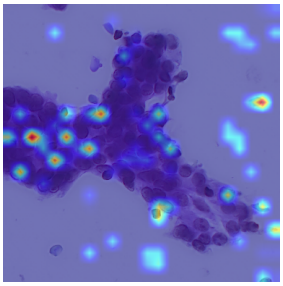
(g) Layer 6



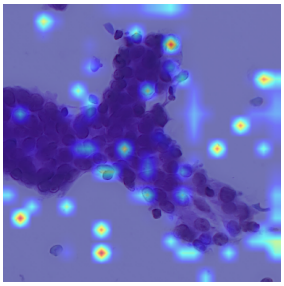
(h) Layer 7



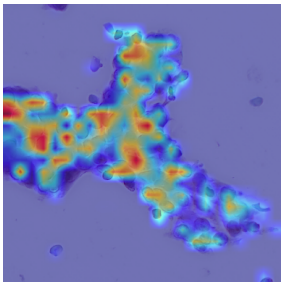
(i) Layer 8



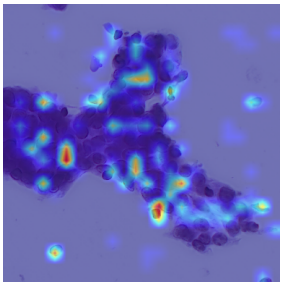
(j) Layer 9



(k) Layer 10



(l) Layer 11



(m) Layer 12

Figure 4.6: Attention Map Visualization on a random image from test set on ViT-Base with a 384x384 input dimension and a 32x32 patch size, as outlined in the fourth scenario

the result, which is different from what Shamshiri (2022) [31] found for CNNs. We speculate that since transformers require a lot of labeled data to be trained compare to CNNs, the number of samples in the pretrained dataset plays a more significant role compared to the dataset being compatible with the target dataset. Our method outperformed other benchmarks, highlighting the strength of transformers compared to CNNs.

Article	Segmentation Method	Classification Method	Evaluation Approach	Best Results: Test Accuracy
Kowal et al. (2014) [20]	HBMO Algorithm	Naive Bayes, Decision Trees, SVM, KNN (based on morphometric features of cell nuclei)	Leave-one-out patient	77.64%
Kowal et al. (2017) [19]	Image Thresholding + Fast Marching	Naive Bayes (based on morphometric features of cell nuclei)	Leave-one-out patient	87.64%
Kowal et al. (2018) [22]	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. (2020) [29]	No Segmentation	Deep CNNs: AlexNet, GoogleNet, SqueezeNet, DenseNet-121, Inception-V3	Training-Validation	91.86%
Kowal et al. (2021) [21]	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%
Shamshiri et al. (2023) [32]	Image Thresholding (MINIMUM Algorithm)	DenseNet-169	Training-Validation-Test on Patch-Level	94.55%
Proposed Method	No Segmentation	ViT-Swin Ensemble	K-fold CV	95.01%

Table 4.9: 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. All previous studies, including ours, have focused on image-level classification. However, Shamshiri (2022) [32] considers patch-level classification.

Attention maps are one of the distinctive and pivotal features that set transformers apart from convolutional neural networks (CNNs). They are visual representations that show where a model focuses its attention within input data. They highlight which parts of the input are most important for generating an output, helping us understand how the model works. The attention maps revealed that the network primarily focused on cell nuclei features, aligning with medical knowledge. Transfer learning with a compatible dataset allowed the network to learn effectively from a small sample size.

Our method demonstrates a 3.06% improvement in classification accuracy over the current state-of-the-art image-level classification studies, which rely on traditional machine learning and deep learning techniques. Our approach even surpasses previous patch-level classification studies, showing 0.46% increase in test accuracy, ultimately achieving

95.01% on the test set and 100% on the validation set. Experimental results suggest that our method, despite utilizing a very limited number of training images, achieves performance comparable to that of experienced pathologists and holds promise for clinical application.

Pretraining each model on the histopathological dataset required approximately 46 minutes, while training on the cytological dataset took about 32 minutes per model. This resulted in a total time of 78 minutes for each model, which is less than half the duration reported in Shamshiri (2022) [31]'s work, where the best model required 159 minutes.

Despite vision transformers' larger size, ViTs can be faster than CNNs in certain scenarios. This speed advantage can be attributed to several factors:

- Parallel Processing: The self-attention mechanism in transformers processes all input data (image patches) simultaneously, as opposed to the sequential processing in CNNs. This allows for more efficient utilization of computational resources, particularly on hardware that supports parallel processing like GPUs.

- Reduced Computational Complexity for Larger Inputs: For large input image sizes, the computational complexity of CNNs increases significantly due to the convolution operations over the entire image. In contrast, ViTs divide the image into fixed-size patches and process these patches independently, which can be more efficient for larger images.

- Elimination of Certain Layers: ViTs do not require certain layers that are standard in CNNs, such as pooling layers. This can reduce the computational overhead during the forward pass.

- Improved Hardware Optimization: Modern hardware, especially GPUs, is becoming increasingly optimized for the types of matrix operations used in transformers. This hardware evolution can lead to faster processing times for ViTs compared to CNNs.

The duration required for training all the models across different scenarios is as follows:

- In the first scenario, 13 models were each trained for 32 minutes, cumulating to 416 minutes.
- The second scenario involved training 12 models for 32 minutes each, totaling 384 minutes.
- The third scenario saw 13 models undergo training, each for a combined duration of 78 minutes (32 minutes for cytology dataset training and 46 minutes for histopathological dataset pretraining), resulting in 1014 minutes.
- Similarly, the fourth scenario required a total of 936 minutes for training 12 models, each for 78 minutes.
- Moreover, each fold of the 4 fold cross validation took 32 minutes, two models undergo training which took 2*(4*32 + 46) resulting in 348 minutes

Collectively, this equates to 3098 minutes, which is approximately 52 hours. This figure solely represents the time expended in model training and does not account for the extensive duration dedicated to optimizing hyperparameters.

4.7 Summary

In this chapter, we presented the experimental outcomes resulting from various scenarios explored in this thesis. The main goal was to demonstrate how effective vision transformers are in classifying breast cancer images. We investigated four different training scenarios for the cytological dataset.

The first scenario included training transformers from scratch. In the second scenario, we performed binary classification using pre-trained networks from the ImageNet dataset. The third scenario revolved around classifying cytological images using pre-trained networks based on the BreakHis dataset. Lastly, in the fourth scenario, we fine-tuned the

model on cytological images after fine-tuning BreakHis images using pretrained networks from the ImageNet dataset.

The results distinctly indicated that transformers surpassed CNNs in performance. Additionally, ViT marginally outperformed the Swin Transformer. We were able to combine the advanteges of both models through an ensemble approach. Furthermore, initiating training with the ImageNet dataset had a substantial impact. Despite the compatibility between histopathology and cytological images, training directly from histopathology images proved less effective. The most effective strategy involved initial training on the ImageNet dataset, followed by further training on the BreakHis dataset, and concluding with fine-tuning on cytological images, which led to optimal result.

Chapter 5

Conclusions and Future Work

In this thesis, we have introduced an innovative transfer learning (TL) strategy that harnesses the capabilities of vision transformers, particularly the ViT and Swin Transformer, to precisely classify cytological biopsy specimens of breast cancer. By utilizing an auxiliary dataset of histopathological images, our approach has demonstrated remarkable performance in classification, achieving substantial accuracy with a constrained number of annotated samples. We rigorously evaluated the proposed method across four distinct training scenarios for each model, revealing an enhancement of about 0.5% in training accuracy compared to the last best model by Shamshiri (2022) [31], although their work focused on patch-level classification.

The success of our TL framework, powered by these state-of-the-art Transformer architectures, highlights the significant potential of such models in the field of medical image analysis, consistently outperforming standard CNNs in our experiments.

5.1 Future Work:

- Given that the dataset is relatively small, including only 550 images, exploring the use of Generative Adversarial Networks (GANs) to augment the dataset could be

beneficial. This approach has the potential to generate additional images for training and merits further investigation.

- Graph Network Integration: Our objective is to combine graph networks with transformer models for whole-slide image classification. This integration is expected to leverage both the contextual understanding of graph networks and the detailed feature extraction capabilities of transformers.

- Diverse Disease Applications: We intend to apply our transfer learning strategy to a wider array of diseases, aiming to enhance the precision and effectiveness of medical diagnoses and treatments.

- Computational Efficiency: We will explore more efficient network designs to evaluate whether similar classification accuracy can be achieved with fewer computational resources.

- Cross-Modality Generalization: Future research should consider applying our transfer learning methods across various imaging modalities, potentially improving diagnostic techniques for a range of medical conditions.

To facilitate further research, we have made our pre-trained models accessible on GitHub, inviting collaboration and continued innovation in the domain of medical image classification.

Appendix A

My Appendix

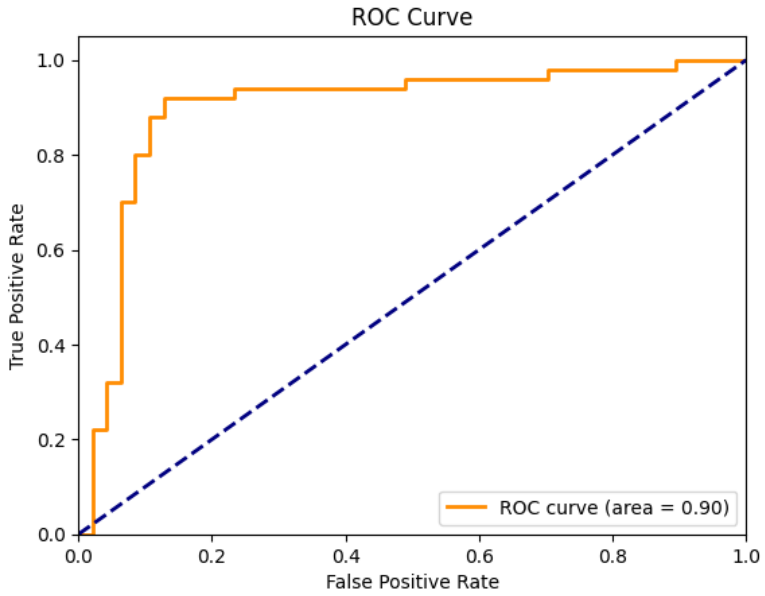


Figure A.1: Scenario1: ROC for the Custom ViT Model

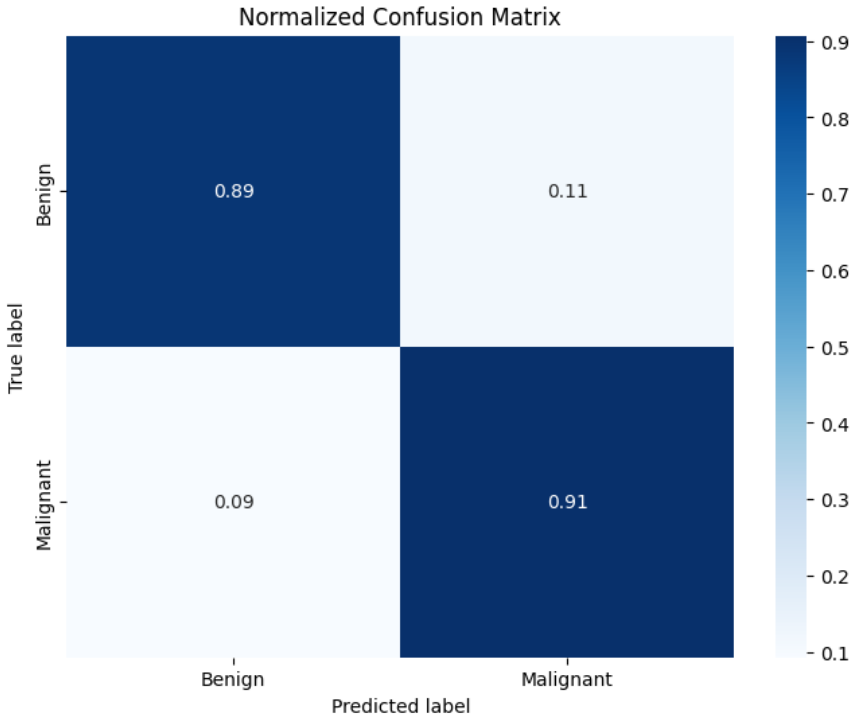


Figure A.2: Scenario1: Confusion Matrix for ViT-Swin Base 384x384 16 Patch-size

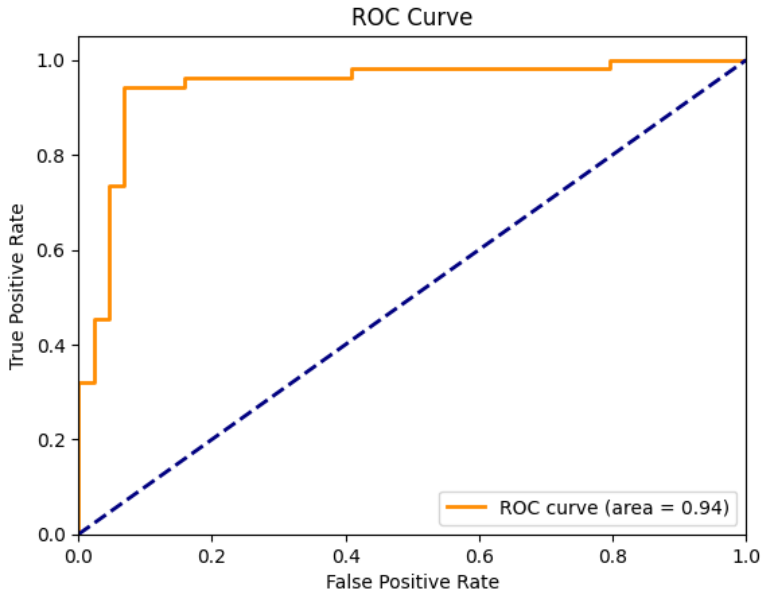


Figure A.3: Scenario2: ROC for the ViT Swin Base Model 384 x 384 16 patch size

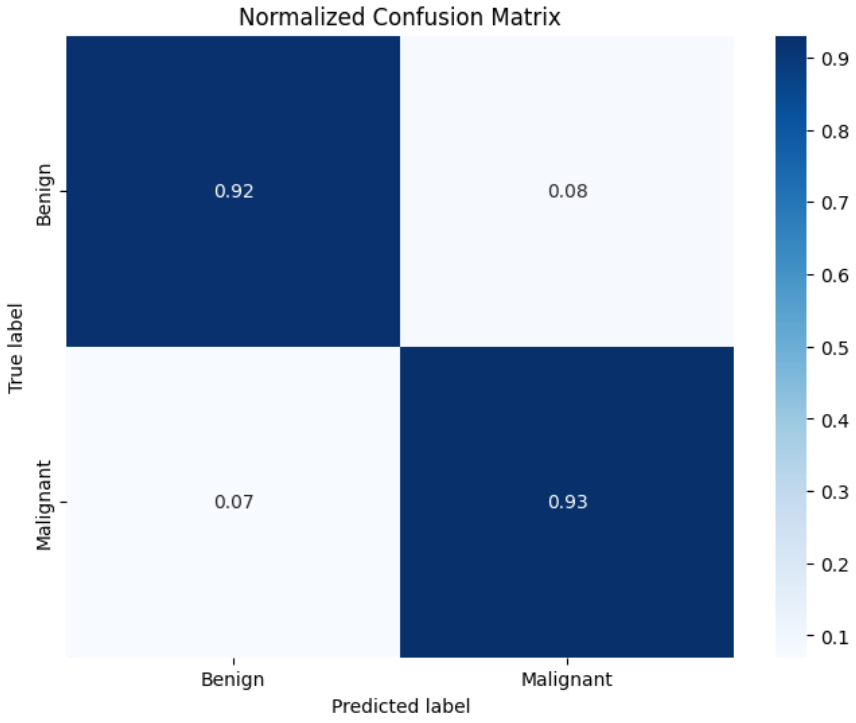


Figure A.4: Scenario2: Confusion Matrix for ViT-Swin Base 384x384 32 Patch-size

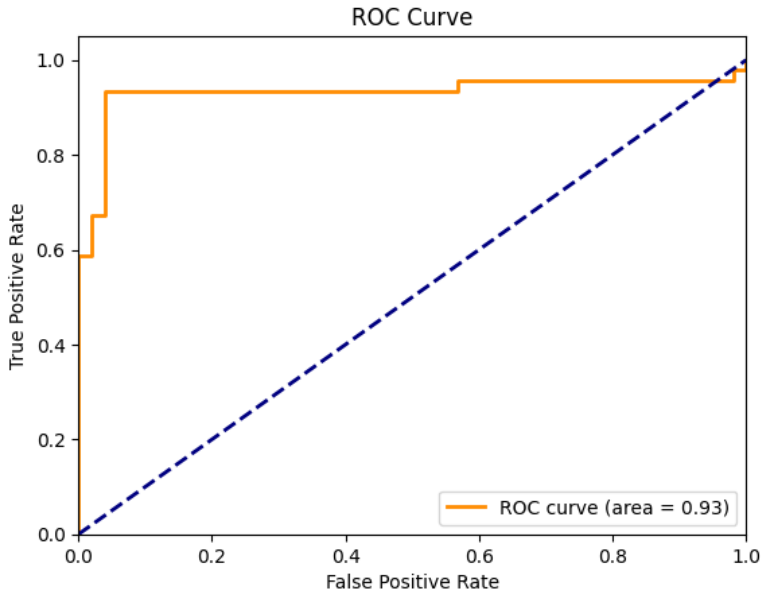


Figure A.5: Scenario3: ROC for the ViT Small Model 384 x 384 16 patch size

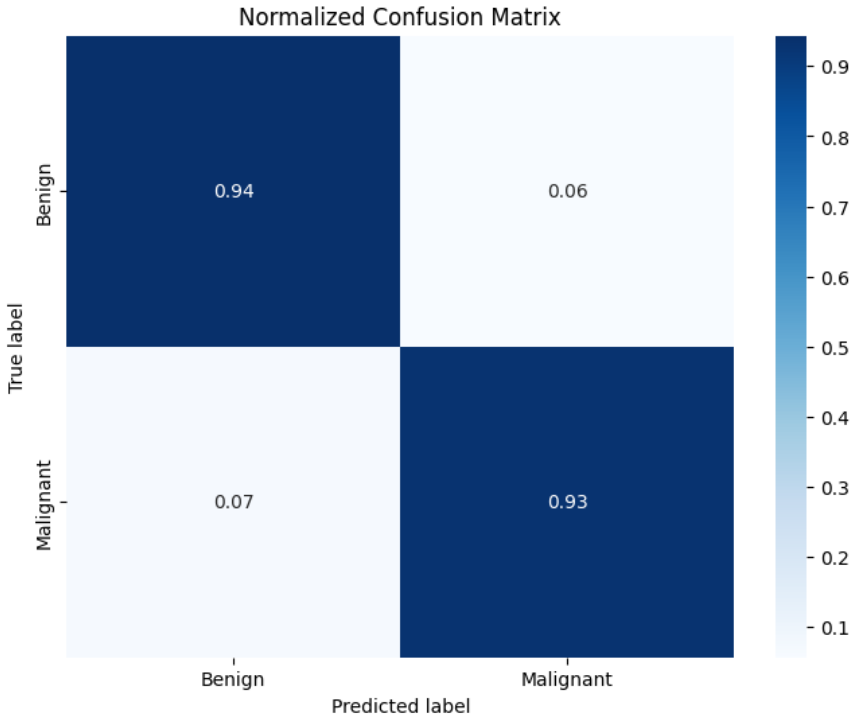


Figure A.6: Scenario3: Confusion Matrix for ViT-Swin Base 384x384 32 Patch-size

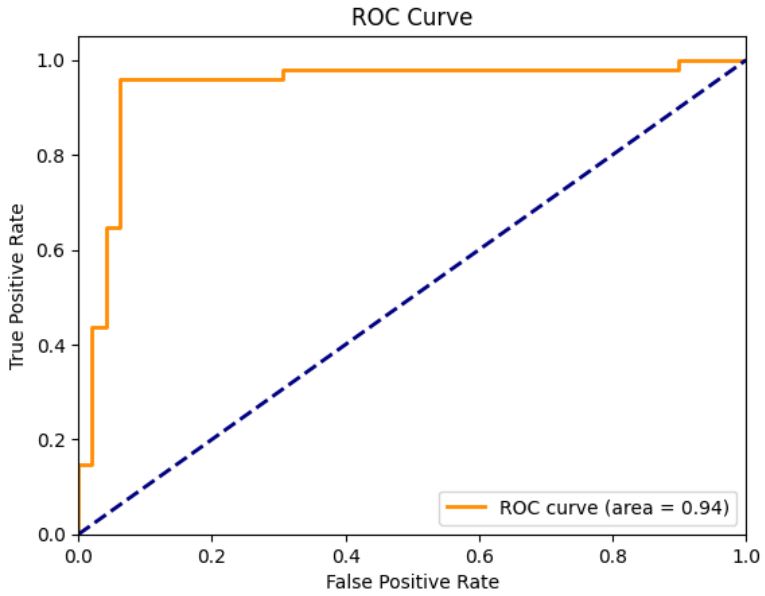


Figure A.7: Scenario4: ROC for the Swin Base Model 384 x 384 16 patch size

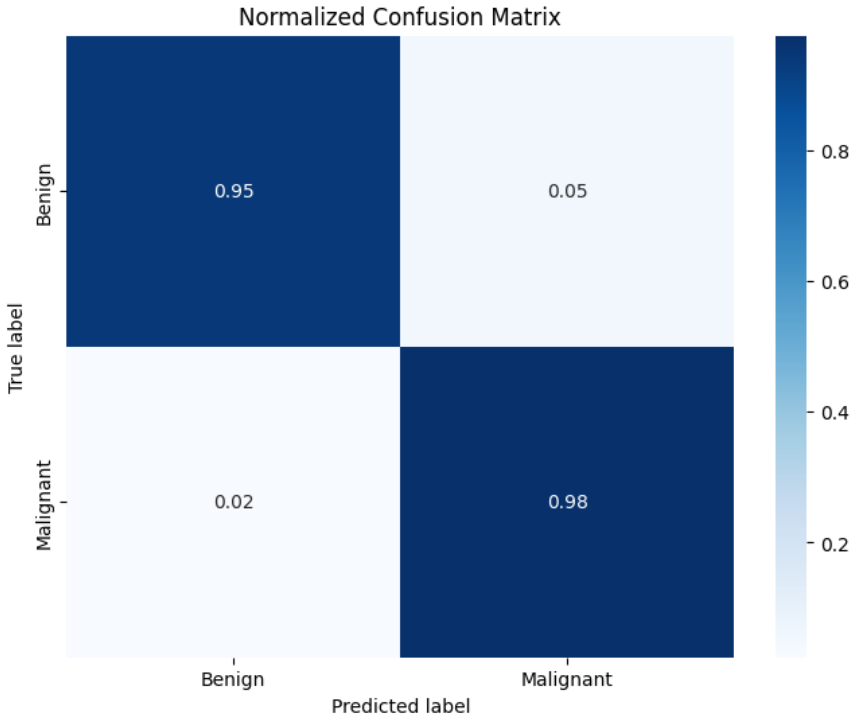


Figure A.8: Scenario4: Confusion Matrix for ViT-Swin Base 384x384 16 Patch-size

Bibliography

[1] Jay Alammar. The illustrated transformer. https://jalammar.github.io/illustrated-transformer/, 2018. Accessed: 2022-09-24.

[2] Amira Alotaibi, Tarik Alafif, Faris Alkhilaiwi, Yasser Alatawi, Hassan Althobaiti, Abdulmajeed Alrefaei, Yousef Hawsawi, and Tin Nguyen. Vit-deit: An ensemble model for breast cancer histopathological images classification. In *2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)*, pages 1–6. IEEE, 2023.

[3] Laith Alzubaidi, Mohammed A Fadhel, Omran Al-Shamma, Jinglan Zhang, and Ye Duan. Deep learning models for classification of red blood cells in microscopy images to aid in sickle cell anemia diagnosis. *Electronics*, 9(3):427, 2020.

[4] Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al. Qupath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1):1–7, 2017.

[5] Christopher M Bishop and Hugh Bishop. Continuous latent variables. In *Deep Learning: Foundations and Concepts*, pages 495–531. Springer, 2023.

[6] Jie-Zhi Cheng, Dong Ni, Yi-Hong Chou, Jing Qin, Chui-Mei Tiu, Yeun-Chung

Chang, Chiun-Sheng Huang, Dinggang Shen, and Chung-Ming Chen. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. *Scientific Reports*, 6(1):24454, 2016.

[7] Yin Dai, Yifan Gao, and Fayu Liu. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8):1384, 2021.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv Preprint ArXiv:2010.11929*, 2020.

[9] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010.

[10] Paweł Filipczuk, Thomas Fevens, Adam Krzyżak, and Roman 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.

[11] Zhiyang Gao, Zhiyang Lu, Jun Wang, Shihui Ying, and Jun Shi. A convolutional neural network and graph convolutional network based framework for classification of breast histopathological images. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3163–3173, 2022.

[12] Zhiyang Gao, Jun Shi, and Jun Wang. Gq-gcn: Group quadratic graph convolutional network for classification of histopathological images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference,*

Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24, pages 121–131. Springer, 2021.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[15] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *ArXiv Preprint ArXiv:1602.07360*, 2016.

[16] D Friday King and Laura AC King. A brief historical note on staining by hematoxylin and eosin. *The American Journal of Dermatopathology*, 8(2):168, 1986.

[17] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ArXiv Preprint ArXiv:1609.02907*, 2016.

[18] Roshan Konda, Hang Wu, and May D Wang. Graph convolutional neural networks to classify whole slide images. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 754–758. IEEE, 2020.

[19] Marek Kowal, Przemysław Jacewicz, and Józef Korbicz. Combining image thresholding and fast marching for nuclei extraction in microscopic images. In *Image Processing and Communications Challenges 8: 8th International Conference, IP&C 2016 Bydgoszcz, Poland, September 2016 Proceedings*, pages 195–202. Springer, 2017.

[20] Marek Kowal, Andrzej Marciniak, Roman Monczak, and Andrzej Obuchowicz. Discovering important regions of cytological slides using classification tree. In *Image Processing & Communications Challenges 6*, pages 67–74. Springer, 2014.

[21] Marek Kowal, Marcin Skobel, Artur Gramacki, and Józef 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.

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

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[24] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570. Pmlr, 2015.

[25] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[27] Tariq Mahmood, Jianqiang Li, Yan Pei, Faheem Akhtar, Azhar Imran, and Khalil Ur Rehman. A brief survey on breast cancer diagnostic with deep learning schemes using multi-image modalities. *IEEE Access*, 8:165779–165809, 2020.

[28] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. Is it time to replace cnns with transformers for medical images? *ArXiv Preprint ArXiv:2108.09038*, 2021.

[29] Bartosz Miselis, Thomas Fevens, Adam Krzyżak, Marek Kowal, and Roman Monczak. Deep neural networks for breast cancer diagnosis: fine needle biopsy scenario. In *Current Trends in Biomedical Engineering and Bioimages Analysis: Proceedings of the 21st Polish Conference on Biocybernetics and Biomedical Engineering*, pages 131–142. Springer, 2020.

[30] Asmaa Mountassir, Houda Benbrahim, and Ilham Berrada. An empirical study to address the problem of unbalanced data sets in sentiment classification. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3298–3303. IEEE, 2012.

[31] Mohammad Amin Shamshiri. *Classification of Breast Cancer Cytological Images using Transfer Learning and Deep Convolutional Neural Networks*. PhD thesis, Concordia University, 2022.

[32] Mohammad Amin Shamshiri, Adam Krzyżak, Marek Kowal, and Józef Korbicz. Compatible-domain transfer learning for breast cancer classification with limited annotated data. *Computers in Biology and Medicine*, 154:106575, 2023.

[33] Wei Shen, Mu Zhou, Feng Yang, Caiyun Yang, and Jie Tian. Multi-scale convolutional neural networks for lung nodule classification. In *Information Processing in Medical Imaging: 24th International Conference, IPMI 2015, Sabhal Mor Ostaig,*

Isle of Skye, UK, June 28-July 3, 2015, Proceedings 24, pages 588–599. Springer, 2015.

[34] Yiqiu Shen, Nan Wu, Jason Phang, Jungkyu Park, Kangning Liu, Sudarshini Tyagi, Laura Heacock, S Gene Kim, Linda Moy, Kyunghyun Cho, et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical Image Analysis*, 68:101908, 2021.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*, 2014.

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

[37] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[38] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[40] Yang Wen, Leiting Chen, Yu Deng, and Chuan Zhou. Rethinking pre-training on

medical imaging. *Journal of Visual Communication and Image Representation*, 78:103145, 2021.

[41] Chii-Shuenn Yang, Ming-Chen Chang, Yee-Jee Jan, and John Wang. Fine needle aspiration of breast myofibroblastoma. *Acta Cytologica*, 54(3):356–358, 2010.

[42] Yu-Dong Zhang, Suresh Chandra Satapathy, David S Guttery, Juan Manuel Górriz, and Shui-Hua Wang. Improved breast cancer classification through combining graph convolutional network and convolutional neural network. *Information Processing & Management*, 58(2):102439, 2021.