Jari Lindroos

Transformers for breast cancer classification

Master’s Thesis in Information Technology

June 6, 2022

University of Jyväskylä

Faculty of Information Technology

Author: Jari Lindroos

Contact information: jari.m.m.lindroos@jyu.fi

Supervisors: Sami Äyrämö, and Ilkka Pölönen

Title: Transformers for breast cancer classification

Työn nimi: Transformerit rintasyövän luokitteluun

Project: Master's Thesis

Study line: Specialisation in Mathematical Modelling in Science and Decision Analytics

Page count: 53+1

Abstract: Breast cancer is the most common cancer worldwide in females apart from non-melanoma skin cancer. Detecting breast cancer as early as possible could significantly reduce its death rates. Histopathological analysis of the breast tissues is needed for determining the malignancy of the tumor on a cellular level. Manual analysis of histopathological images is time consuming and sensitive to human errors. Deep learning has introduced methods for recognizing breast cancer to assist pathologists in their diagnostic workflow. The convolutional neural networks have for long been the bandwagon deep learning model for breast cancer classification, but they are mostly limited at focusing on local variations in image patterns. The Vision Transformer, which originated from the dominant Transformer architecture in natural language processing has shown to outperform convolutional neural networks on several image classification benchmarks, due to its ability to focus on long range dependencies in images. In this thesis we aim to evaluate the performance of Vision Transformer based models by comparing them to the commonly used convolutional neural network ResNet-50 on the PCam-dataset. For training we utilize both the conventional transfer learning based approach and also an pre-training approach based on domain adaptation. We demonstrate the effectiveness of the implemented Vision Transformer models in the medical domain, by obtaining better results than the ResNet-50 on the PCam-dataset, with the best model B/16 achieving the best AUC score of 0.97315. The use of domain-based pre-training shows a performance gain for every model except the Ti/16-family models.

Keywords: vision transformer, transformer, deep learning, image classification, breast cancer, convolutional neural network

Suomenkielinen tiivistelmä: Rintasyöpä on maailmanlaajuisesti naisten yleisin syöpä, sen varhainen havaitseminen voi merkittävästi vähentää siihen liittyvää kuolleisuutta. Histopatologista analyysiä tarvitaan kasvainten laadun määrittämiseksi solutasolla. Histopatologisten kuvien manuaalinen analyysi vie kuitenkin aikaa ja on altis virheille. Syväoppimiseen pohjautuvassa tutkimuksessa on esitetty menetelmiä rintasyövän tunnistamiseen, jotka voivat auttaa patologeja diagnosoimisessa. Konvoluutioneuroverkot ovat pitkään olleet käytetyin menetelmä rintasyövän luokitteluun syväoppimisessa, mutta ne ovat enimmäkseen rajoittuneet keskittymään kuvien paikallisiin ominaisuuksiin. Vision Transformer on osoittautunut suoriutumaan konvoluutioneuroverkkoja paremmin useissa kuvanluokittelutehtävissä, koska se pystyy keskittymään kuvien pitkän matkan riippuvuuksiin. Tämän tutkielman tavoitteena on arvioida Vision Transformer -pohjaisten mallien suorituskykyä vertaamalla niitä yleisesti käytettyyn konvoluutioneuroverkkoon ResNet-50, kokeilut suoritetaan PCam-aineistolla. Mallien koulutuksessa hyödynnämme sekä perinteistä siirto-oppimiseen perustuvaa lähestymistapaa että myös toimialuekohtaiseen esikoulutukseen perustuvaa lähestymistapaa. Osoitamme, että implementoiduilla Vision Transformer -malleilla saadaan parempia tuloksia kuin ResNet-50 -mallilla. Parhaalla mallilla B/16 saavutettiin paras AUC-tulos arvolla 0.97315. Toimialuekohtaisen esikoulutuksen käyttö parantaa suorituskykyä kaikissa malleissa paitsi Ti/16 malleissa.

Avainsanat: vision transformer, transformer, syväoppiminen, kuvan luokittelu, rintasyöpä, konvoluutioneuroverkko

Glossary

CNN	Convolutional Neural Network
ViT	Vision Transformer
MHSA	Multi Head Self-Attention
PCam	PatchCamelyon
CAD	Computer-Aided Diagnosis
ROC	Receiver Operating Characteristic curve
AUC	Area Under the ROC Curve
NLP	Natural Language Processing
CV	Computer Vision
WSI	Whole Slide Imaging
MLP	Multi-Layer Perceptron
LayerNorm	Layer normalization
SGD	Stochastic Gradient Descent

List of Figures

List of Tables

Contents

1 Introduction

Breast cancer is the most common cancer worldwide in the female population after non-melanoma skin cancer (Waks and Winer 2019). Detecting breast cancer as early and as accurately as possible can significantly lower its mortality rate. (Wang 2017). Mammograms are one of the most commonly used techniques for detecting breast cancer (Smith, Cokkinides, and Brawley 2008), however histopathological analysis is also needed for determining the malignancy of the tumor on a cellular level (Gurcan et al. 2009).

Manually analyzing the histopathological images takes a lot of time and may be sensitive to human errors (Gurcan et al. 2009). Thus, a Computer-Aided Diagnosis system that can be used as an aid for making decisions in numerous vision tasks, such as classification, segmentation and detection can be highly helpful. Moreover the recent increase in computational power led to the development of applications and algorithms for medical image analysis for processing and analyzing histopathological images of breast cancer (Gurcan et al. 2009).

Deep learning has introduced several methods for recognizing breast cancer with high applicability in breast cancer diagnostics, and has basically inserted itself as a practical tool in Computer-Aided Diagnosis systems to further aid pathologists in their diagnostics workflow (Kwong and Mazaheri 2021). The advancements in deep learning had long been revolving around Convolutional Neural Networks (CNN). Since 2016 CNNs were by far the most used deep learning models for breast cancer classification (Mridha et al. 2021; Kwong and Mazaheri 2021). The CNN is widely used for its capabilities to extract useful features from images and it has made a significant contribution to various tasks in medical imaging, including classification, detection and segmentation (Litjens et al. 2017). Eventhough CNNs are great at feature extraction tasks, the downsides are that they are mostly limited at focusing on local variations in image patterns so they lose the global information of the features (Parvaiz et al. 2022).

The big breakthrough in Natural Language Processing introduced the Transformer architecture (Vaswani et al. 2017), which inspired many researchers to leverage its architecture for a variety of tasks in computer vision due to its attention mechanism, which focuses on extracting more global information (Parvaiz et al. 2022). The Vision Transformer (ViT) (Dosovitskiy et al. 2021) was the first transformer-based architecture, which was applied for image data. The Vision Transformer model swiftly demonstrated its effectiveness by pushing the state-of-the-art results in various computer vision tasks, including image classification, detection and segmentation. Additionally, recent research shows that the prediction errors of Vision Transformers are more human-like than the prediction errors of CNNs (Tuli et al. 2021).

In this thesis we aim to use Vision Transformer models based on the recommended models by, Steiner et al. 2021 to assess their performance on a histopathological image classification task and compare the results on a conventionally used CNN. We also aim to demonstrate whether domain-based pre-training can provide an increase in classification performance.

1.1 Research questions

For the following questions we wish to seek answers:

- How do the Vision Transformer based models perform in a breast cancer classification task compared to the conventionally used CNNs?
- Can the Vision Transformer based models pre-trained on domain specific datasets perform better on task-specific image classification than Vision Transformer models, that are pre-trained with ImageNet?

1.2 Structure

The structure of the thesis is laid out as follows: Chapter 2 introduces us to the concepts of breast cancer, histopathological imaging and the previous work done on tumor classification in deep learning, to get a glimpse of what kind of methods are more commonly represented. Chapter 3 sets the base foundation on the theoretical background of the thesis, focused on deep learning, convolutional neural networks and transformers. Chapter 4 gives us a description of the datasets that we are going to use for our experimental part. In chapter 5 we introduce our chosen methods and test environment used for the experiments. Chapter 6 presents the results obtained from the chapter 5 experiments. Chapter 7 presents our findings and discussion about the results obtained. Chapter 8 lays the final foundation and conclusion from the work of this thesis.

2 Deep Learning in Tumor Classification

In the first section, we briefly get an overview of breast cancer, its diagnosis and the situation of breast cancer worldwide and in Finland. The second section is focused on histopathological imaging, the importance of its analysis, the growing computing power, and the development of computer-aided diagnosis tools, which can help pathologists in their diagnostic workflow. In the third section we comprehensively review the applications of CNNs and ViTs in tumor classification, with an emphasis on breast cancer, first the status of classification methods for tumor classification is introduced. Next, we summarize findings on the common CNN methods and finally, we focus on the findings of the ViT-based methods and how they compare to the conventionally used CNNs.

2.1 Breast cancer

Breast cancer is a term used to describe the uncontrollable development and growth of the cells in the breast tissue (Khuwaja and Abu-Rezq 2004). Several types of tumors can develop in the breast tissues, but most of the tumors are the result of non-cancerous (benign) changes within the breast (Sharma et al. 2010). The most common symptoms of breast cancer is a lump, located either in the breast or armpit, but as breast cancer rarely causes pain, a painless mass is much more worrisome for malignancy than one that is causing symptoms (Richie and Swanson 2003).

Currently, breast cancer is the most common cancer worldwide among the female population, apart from non-melanoma skin cancer (Waks and Winer 2019). 5136 new cases of breast cancer was diagnosed in 2019, and approximately over 5000 new cases of breast cancer is diagnosed every year in the female population of Finland (Pitkäniemi et al. 2020). Furthermore, worldwide in 2020, breast cancer had 2.3 million new cases in the female population, which makes it the highest number of new cases out of the 36 cancer types (Sung et al. 2021). Also, the amount of new deaths caused by female breast cancer was over 684.000, making it also the fourth highest compared to the other cancer types (Sung et al. 2021).

Breast cancer is usually diagnosed by biopsy of breast nodules, which was detected by mammogram or by palpation (Smith, Cokkinides, and Brawley 2008). The screening of the breast is mostly performed in women who show no signs or symptoms of breast cancer for early detection, but the elements for performing breast screening also vary on different factors such as the age of the patient and the previous medical history (Bevers et al. 2009).

Previous studies have suggested that early and accurate breast cancer detection with fitting treatment could long-term significantly reduce breast cancer death rates and improving the prognosis (Wang 2017).

2.2 Histopathological imaging

Histopathology is the basis for cancer recognition and refers to the microscopic study of the disease of tissues and its diagnosis (Gurcan et al. 2009).

The importance of quantitative analysis of pathological images has been acknowledged by researchers in the domains of image analysis and pathology. Considering that the majority of current pathology diagnoses are based on the actual pathologists subjective opinions, quantitative image-based analysis of digital histopathology image slides is needed. It is critical not only for diagnosis, but also for understanding the underlying reasons behind a given diagnosis (Gurcan et al. 2009).

Increased computing power and improved image analysis algorithms have made it possible to develop powerful computer-aided diagnosis (CAD) methods for histopathological data. Because of the digital whole-slide scanners, the slides from the histopathological images can be digitized and saved as digital images (Gurcan et al. 2009). This opens digital histopathology up for the applications of computational image analysis and different deep learning techniques.

Pathologists can use the CAD tools to increase diagnosis accuracy and detection rates while also lowering the total rate for misdiagnosis (Gurcan et al. 2009). Automating some tasks in their diagnostic workflow could help increasing the total efficiency and objectivity (Bayramoglu, Kannala, and Heikkilä 2016).

2.3 Classification methods

Since cancer is the second biggest cause of death, numerous machine learning studies have been published to tackle this problem. Having a computer-based analysis tool for providing second hand opinions and accurate medical image classifications, can certainly help medical practitioners in treatments and clinical care.

Even though machine learning methods have achieved positive results analyzing histopathological images of breast cancer, their performance is greatly limited by being dependant on the extracted features of the data and task they are trained for (J. Xie et al. 2019). Furthermore, they lack the understanding for extracting discriminative information from the data (Bengio, Courville, and Vincent 2013).

Deep learning approaches are generally based on deep neural networks, especially the Convolutional Neural Networks (CNN). CNNs have been widely used in the field of medical imaging, since they have the ability to automatically extract more useful feature representations from the images and the image labels (D. Wang et al. 2016), without requiring the manual annotation of the features (Bengio, Courville, and Vincent 2013; Spanhol et al. 2016).

CNNs have been dominant in both computer vision and the field of medical image analysis before the emergence of Vision Tranformers. Mridha et al. 2021; Kwong and Mazaheri 2021 found on their surveys that the CNN was by far the most commonly used deep learning model for breast cancer classification since 2016. CNNs have been frequently employed in breast cancer classification because of its capability for extracting useful feature representations from the images. A lot of effort has gone into improving the performance of CNN-based classifiers. In recent years various new CNN architectures were developed: AlexNet (Krizhevsky, Sutskever, and Hinton 2012), VGG (Simonyan and Zisserman 2014), GoogleNet (Szegedy, Liu, et al. 2015), ResNet (Kaiming He et al. 2016), ResNeXt (S. Xie et al. 2017), and EfficientNet (Tan and Le 2019).

Previous implemented methods on medical imaging, which are based on CNNs tend to focus on local variations in image patterns. Vision Transformers (ViT) are more focused on modelling long range dependencies and thus are able to extract more global information from images (Parvaiz et al. 2022). ViTs have shown that pure transformers can also perform well on image classification tasks (Dosovitskiy et al. 2021; Khan et al. 2021) by being ranked on top of several image classification benchmarks. This has inspired many researches to further evaluate the performances between ViTs and CNNs. Vision transformer models not only outperform CNNs based on image classification benchmarks, but also show that their error is also more consistent with human errors (Tuli et al. 2021) (Kelei He et al. 2022).

In the next section, the relevant CNN-based models and ViT-based models to classify tumors are summarized.

2.3.1 CNN-based methods

D. Wang et al. 2016 identified metastatic breast cancer on the Camelyon16 dataset. Their method is based on comparing the performance of 4 different types of CNNs: GoogLeNet (Szegedy, Liu, et al. 2015), AlexNet (Krizhevsky, Sutskever, and Hinton 2012), VGG16 (Simonyan and Zisserman 2014) and FaceNet (Schroff, Kalenichenko, and Philbin 2015). GoogLeNet was selected by its performance to generate the tumor probability heatmaps, and finally a random forest classifier was used to classify the cancerous whole-slide images and the negative whole-slide images. The proposed method was the winning solution to the Camelyon16 Grand Challenge (Ehteshami Bejnordi et al. December 2017), and the results demonstrated the power of using deep learning by achieving classification performance of nearly human-level on the test dataset.

Spanhol et al. 2016 first propose a dataset called BreakHis for classifying histopathological images of breast cancer. They used a modified version of the AlexNet convolutional neural network, which showed better results than the conventional machine learning models.

Bayramoglu, Kannala, and Heikkilä 2016 proposed to classify breast cancer histopathological images on the BreaKHis dataset regardless of their magnification level, using convolutional neural networks. The two different architectures proposed: one CNN to predict the level of tumor malignancy, and the second CNN used for predicting both the tumor malignancy and the image magnification levels at the same time. Their obtained results were also competitive with the best results obtained from the traditional machine learning methods as above.

Gecer et al. 2018 use a CNN for detecting and classifying whole-slide-images and the region of interest in breast cancer. Their classifier obtained comparable results, which were based on the diagnosis of 45 different pathologists who were focused on diagnosing breast cancer.

Graham, Epstein, and Rajpoot 2020 propose a CNN called Dense Steerable Filter for histological image analysis. Their model achieves state-of-the-art performance in breast tumor classification on the PCam dataset (Veeling et al. 2018), while having fewer model parameters than any of the previous implemented models.

R. Zhang et al. 2022 perform a large transfer learning evaluation for one ResNet architecture for domain-adaptation across nine histopathological datasets. The results show that knowledge is transferred between histopathological datasets and the datasets which shared the same organ class had also shared knowledge more effectively.

2.3.2 Transformer based methods

Attention mechanisms, which utilize self-attention have for long been implemented as supplementary modules of convolutional neural networks for the analysis of medical images, before the implementation of Vision Transformers started to gain popularity in the computer vision community. Attention modules have successfully shown to improve the performance of deep learning models (Y. Liu et al. 2019; Han et al. 2021), which has inspired many researchers to directly integrate the Transformer architecture into the field of computer vision. Some of the methods directly focus on using pure transformers to replace CNNs altogether (Dosovitskiy et al. 2021), while others use a hybrid framework of both CNN and transformer, such as the Bottleneck Transformer (Srinivas et al. 2021).

This has also led to the recent rise of pure Transformer-based models in the field of medical image analysis. The first published transformer based method for medical image classification is TransMed (Dai, Gao, and Liu 2021), a hybrid transformer and CNN-based architecture which focuses on the classification of parotid tumors in multi-modal magnetic resonance medical images.

GasHis-transformer (Chen, Li, et al. 2021) , a Vision Transformer model is proposed for diagnosing gastric cancer in the stomach, where the microscopic images are classified into abnormal and normal cases. GasHis-Transformer model consists of two modules for feature extraction. The global information module (ViT-based) and the local information module (CNN based). The proposed model achieved greater classification performance than its compared CNN counterparts. The model also shows generalizability on other histopathological image datasets, such as breast cancer classification on the BreakHis dataset.

TransMIL (Shao et al. 2021) is a transformer-based architecture for exploring morphological and spatial information for classifying whole-slide images. The proposed model effectively dealt with both binary and multiclass classification. TransMIL achieves state-of-the-art classification performance on three different cancer datasets: breast cancer (CAMELYON16) (Veeling et al. 2018), lung cancer (TCGA-NSCLC) (Napel and Plevritis 2014), and kidney cancer (TCGA-R) (National cancer institute 2022)

Gheflati and Rivaz 2021 use both hybrid and pure Vision Transformers for classifying breast tissues in ultrasound images using different data augmentation strategies. The performance of their used models are compared with some of the state-of-the-art CNNs. The results of the performance were more in favor of the Vision Transformers than the CNNs for classification of ultrasound breast images.

Stegmüller et al. 2022 propose ScoreNet, a Vision Transformer based architecture for histopathological breast cancer classification on the BRACS dataset (Pati et al. 2022). The results show state-of-the-art performance but ScoreNet also demonstrated robustness and generalization in other breast cancer datasets like Camelyon16 (Ehteshami Bejnordi et al. December 2017) and BACH (Aresta et al. 2019).

3 Theory

This chapter will present the theoretical background for the thesis. The first section contains the key aspects of deep learning and the important contents and terms used in this thesis focused on image classification and transfer learning. The second section gives an overview of convolutional neural networks and the common layers in its architecture. In the third and last section we focus on the Transformers, the attention mechanism behind it, and the Vision Transformer with its training process.

3.1 Deep learning

Deep learning makes it possible for multi-layer computational models to learn representations of data with many feature levels (LeCun, Bengio, and Hinton 2015). The deep learning models learn complicated concepts by building them from simpler concepts (Goodfellow, Bengio, and Courville 2016). One example of a simple deep learning model is called feedforward neural networks, also known as Multi-Layer Perceptrons (MLP) (Goodfellow, Bengio, and Courville 2016).

Let us describe the MLP through a function f^* that is to be approximated. When from the input $\mathbf{x}$ a mapping $y = f^*(\mathbf{x})$ is done to its corresponding class of output y, the best approximation is defined as $\mathbf{y} = f(\mathbf{x};\sigma)$ where σ is parameters that yields the best results (Goodfellow, Bengio, and Courville 2016).

Deep learning methods have greatly improved the state-of-the-art results in various areas such as object detection (Wu, Sahoo, and Hoi 2020), natural language processing (Brown et al. 2020), speech recognition (Y. Zhang et al. 2020), drug discovery (Lavecchia 2019) and autonomous vehicles (Kuutti et al. 2020).

Deep learning methods can generally be split into supervised and unsupervised learning algorithms, which stem from machine learning (LeCun, Bengio, and Hinton 2015).

Supervised learning can be explained as presenting the model with a dataset $Data = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N}$. Where $\mathbf{x}$ represents our input features of the data, whereas $\mathbf{y}$ represents an set of fixed class labels. The objective of training in supervised learning usually translate into finding model parameters σ, such that it is the most effective at predicting the data based on a certain loss function $Loss(\mathbf{y}, \mathbf{y}_o)$. Where $\mathbf{y}_o$ points at the models attained output of the obtained from the inserted data point $\mathbf{x}$ in the function $f(\mathbf{x}; \sigma)$ which represents the model. (Litjens et al. 2017)

Compared to supervised learning, in unsupervised learning the model learns the data with unlabeled examples, by distinguishing useful patterns in the images. The training for unsupervised learning can be conducted with different loss functions, such as the reconstruction loss $Loss_r(\mathbf{x}, \mathbf{x_o})$, in which the model learns to reconstruct its original input $\mathbf{x}$ through noisy representation to obtain the reconstructed input $\mathbf{x_o}$ (Litjens et al. 2017).

3.1.1 Image Classification

Image Classification is a subdomain of computer vision with the goal of associating one or more labels to a given image. (Wang and Su 2019)

The ImageNet database (Fei-Fei, Deng, and Li 2009) is a visual dataset consisting of 1.3 million images in 1,000 different classes, created to aid in the computer vision research as a classification benchmark. The contribution of Krizhevsky, Sutskever, and Hinton 2012 to the ImageNet challenge in December 2012 was a big breakthrough in image classification (Litjens et al. 2017) with the introduction of AlexNet, a convolutional neural network which placed first in the 2012 ImageNet challenge by a large margin. Since then, each year newer and deeper deep learning models which have been shown to achieve near humanlike performance have been proposed. (Szegedy, Liu, et al. 2015) (Kaiming He et al. 2016).

3.1.2 Transfer Learning

Transfer learning focuses on transferring latent knowledge between two domains: the source domain, which is generally a large dataset on which the networks pre-training is performed on for solving the issue of lack of data on the target domain where the networks fine-tuning happens (Aneja et al. 2021).

The domain can be defined as $\mathscr{D} = \{\mathscr{X}, P(\mathrm{X})\}$ where $\mathscr{X}$ is the feature space and $P(\mathrm{X})$ is the marginal probability distribution, where $\mathrm{X} = \{x_1, \ldots, x_n\} \in \mathscr{X}$. Given a particular domain $\mathscr{D}$, a task $\mathscr{T}$ can be defined as $\mathscr{T} = \{\mathscr{Y}, f(\cdot)\}$, which consists of the label space $\mathscr{Y}$ and the task prediction function $f(\cdot)$ (Weiss, Khoshgoftaar, and Wang 2016). There exists training data $\{x_i, y_i\}$, where $x_i \in \mathscr{X}$ and $y_i \in \mathscr{Y}$, from where the task $\mathscr{T}$ can be learned.

Following the presentation by Weiss, Khoshgoftaar, and Wang 2016 transfer learning can thus be defined as: Given the source domain $\mathscr{D}_S = \{\mathscr{X}_S, P(\mathrm{X}_S)\}$ and the source learning task $\mathscr{T}_S = \{\mathscr{Y}_S, f_S(\cdot)\}$, and the target domain $\mathscr{D}_T = \{\mathscr{X}_T, P(\mathrm{X}_T)\}$ and the target learning task $\mathscr{T}_T = \{\mathscr{Y}_T, f_T(\cdot)\}$, the purpose of transfer learning is to learn the prediction function of target $f_T(\cdot)$ by using the information from $\mathscr{D}_S$ and $\mathscr{T}_S$ in a way that $\mathscr{D}_S \neq \mathscr{D}_T$ and $\mathscr{T}_S \neq \mathscr{T}_T$ (Weiss, Khoshgoftaar, and Wang 2016).

Transfer learning makes it possible to start a deep learning training process with a better starting point where a more favorable local optimum for the training criterion is found. Being in a more favorable region of the parameter space means that the optimization of the model is easier, which means faster learning in the training process and can lead to better model performance (Aneja et al. 2021).

Despite the fact that transfer learning offers many benefits in transferring the learned representations of the features between the source domain and target domain, it can also be disadvantageous for the performance if the source domain and the target domain lack the same kind of features. This phenomenon can also be known as negative transfer (Weiss, Khoshgoftaar, and Wang 2016).

3.2 Convolutional neural networks

Convolutional neural networks (CNN) are a deep-learning framework that is the cornerstone of deep learning. CNNs take an input image for which it assigns certain weights and biases to be able to recognize the different features located in the images (Goodfellow, Bengio, and Courville 2016).

Pioneered by (LeCun et al. 1989), CNN were originally used for classifying digits and recognizing hand-written numbers, but as stated in chapter 3.1.1, a big breakthrough in the history of CNN was made in the 2012 ImageNet challenge (Krizhevsky, Sutskever, and Hinton 2012). Since then, CNNs have shown state-of-the-art results in areas such as image recognition and classification (Klang 2018).

CNNs are designed to handle a wide range of two-dimensional shapes, and thus are widely used in different tasks of computer vision such as, image classification, semantic segmentation and medical image analysis. (Sarker 2021).

A convolutional neural network can be seen as a sequential order of layers (Figure 1), which transform an input image into an output vector with the size of predefined classes, which means that a vector with a size of 1000 indicates the probability for an input image to belonging to one or several of the 1000 predetermined classes (Goodfellow, Bengio, and Courville 2016). A CNN architecture, consist of three main types of layers named: (1) convolutional layer, (2) pooling layer, and the (3) fully-connected layer (FC) (D. Liu et al. 2018).

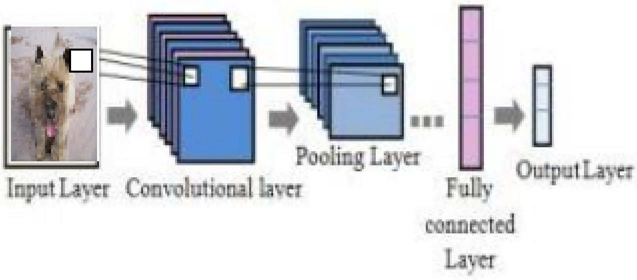

Figure 1: Example of a CNN architecture which contains the three main layer types, with an example image from the ImageNet (Fei-Fei, Deng, and Li 2009). Adapted from (Hamid and Walia 2021)

Convolutional layers are the main components of the CNN, they are in charge of feature extraction from the images and the feature maps, which are formed by the previous layers to form new feature maps (Rawat and Wang 2017).

In image classification, the input of a convolutional layer consists of one or more two-dimensional matrices, and multiple two-dimensional matrices are formed as the output. The process for computing a single output matrix can be defined as:

$$\mathbf{A}_j = f_l(\sum_{i=1}^{N} \mathbf{I}_i * \mathbf{K}_{i,j} + B_j) \tag{3.1}$$

First for every input matrix $\mathbf{I}_i$ the convolution operation $*$ is applied with the corresponding kernel matrix $\mathbf{K}_{i,j}$, which represents the feature extractors that handle the extraction of localized features from the input matrices. Next the convoluted matrices are summed together, the bias value B_j is summed to every element of the resulted matrix. Lastly a non-linear activation function f_l is added for every element of the previous matrices to produce one final output matrix (feature map) $\mathbf{A}_j$. The goal of the learning process is to find the kernel matrices $\mathbf{K}$ for extracting useful features, which can be used for classification purposes (Li et al. 2014).

Pooling layers are typically located between consecutive convolution layers to merge similar features into one by performing downsampling on the spatial dimensions, which reduces the amount of parameters and the amount of computation needed in the CNN (Rawat and Wang 2017).

All previous layers including convolution layers and pooling layers are concentrated on extracting and mapping the helpful features into lower-dimensional level representations; the function of the FC-layer is to map these representations into the desired target space and thus completes the classification. The output of the last fully-connected layer is then inserted into a classifier, which handles computing the probabilities for the input image belonging to the corresponding classes. (D. Liu et al. 2018)

3.3 Transformers

Transformers introduced by Vaswani et al. 2017 are a type of neural network architecture originally proposed for the machine translation task but achieved state-of-the-art performance in a large number of Natural Language Processing (NLP) tasks, after that it has also recently been widely applied in various computer vision tasks, including object detection (Carion et al. 2020), segmentation (Chen, Lu, et al. 2021), image enhancement (Yang et al. 2020) and video processing (Zeng, Fu, and Chao 2020).

The Transformer is the first architecture that revolutionized the use of self-attention mechanisms for calculating the representations of its input and output without utilizing any convolution layers (Vaswani et al. 2017). The architecture of transformers are multi-layered, consisting of an encoder layer and a decoder layer, each formed by stacking multiple Transformer blocks on top of each other. Each Transformer block is characterized by a multi-head self-attention mechanism (Tay et al. 2020; Dosovitskiy et al. 2021).

Self-Attention

Self-Attention (SA), is a type of attention mechanism that can be compared to convolutions in CNNs. It helps with learning long range dependencies across image regions (Parvaiz et al. 2022).

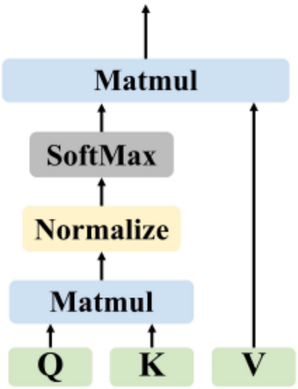

Figure 2: An overview of the Scaled Dot-Product Attention (Dai, Gao, and Liu 2021)

The Self-Attention layer transforms the input into three different embedding matrices : the query matrix $\mathbf{Q}$, which represents the input. The key matrix $\mathbf{K}$ represent what the query is compared with. Finally the value matrix $\mathbf{V}$, which tells how much each of the keys is relevant to the query. The output of the SA-layer is the weighted sum between all of the value vectors. The weights that are allocated to each value are determined by the scaled dot-product (Figure 2) between the query and its matching key (Vaswani et al. 2017), thus the function for the SA-head between the query, key and value can be calculated as following:

$$SA_i(\mathbf{QW}_i^{\mathbf{Q}}, \mathbf{KW}_i^{\mathbf{K}}, \mathbf{VW}_i^{\mathbf{V}}) = \underbrace{softmax\left(\frac{\mathbf{QW}_i^{\mathbf{Q}}\left(\mathbf{KW}_i^{\mathbf{K}}\right)^T}{\sqrt{d_k}}\right)}_{\mathbf{P}}\mathbf{VW}_i^{\mathbf{V}}, \tag{3.2}$$

where i is the index of the SA-head, $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n \times d_m}$ are the input embedding matrices, n is the sequence length, d_m is the embedding dimension, $\mathbf{W}_i^{\mathbf{Q}}, \mathbf{W}_i^{\mathbf{K}} \in \mathbb{R}^{d_m \times d_k}, \mathbf{W}_i^{\mathbf{V}} \in \mathbb{R}^{d_m \times d_v}$

are the learned projection matrices and d_k, d_v are the hidden dimensions of the projection subspaces. The projection matrices are responsible for calculating the attention weights from the feature values which are projected from these matrices (S. Wang et al. 2020; Vaswani et al. 2017). SA, which was defined in Equation 3.2 refers to a context mapping matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$. The Transformer utilizes $\mathbf{P}$ for capturing the context of an input for a given patch or token, based on the combination of every previous patch or token in the sequence (S. Wang et al. 2020).

The Softmax is a function, which turns a input vector $\tilde{\mathbf{x}}$, into a vector of probabilities which sums to 1 (Goodfellow, Bengio, and Courville 2016). Softmax can be defined as following:

$$softmax(\tilde{\mathbf{x}})_i = \frac{e^{x_i}}{\sum_{j=1}^{K_n} e^{x_j}} \qquad x_i \in \mathbb{R} \tag{3.3}$$

and where K_n equals to the number of classes in the classifier.

Multi-Head Self-Attention (MHSA) , the essential component of the Transformer (Figure 3) consists of many SA-heads which are concatenated together to model the dependencies between the input sequence elements. If we denote each SA process as

$$head_i = SA_i(\mathbf{QW}_i^{\mathbf{Q}}, \mathbf{KW}_i^{\mathbf{K}}, \mathbf{VW}_i^{\mathbf{V}}), \tag{3.4}$$

then MHSA is defined as :

$$\mathrm{MHSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{ Concat } (\mathrm{head}_1, \text{ head }_2, \ldots, \mathrm{head}_h) \mathbf{W}^O, \tag{3.5}$$

where h is the number of heads and $\mathbf{W}^O \in \mathbb{R}^{hd_v \times d_m}$ is the learned projection matrix (S. Wang et al. 2020).

The benefits of MHSA is that it enables learning sequential and locational information in different representational subspaces for the model, since each self-attention head has its own internal representation of the inputs, thus sharing the information makes it possible for a more complete understanding of the relationships between the image patches in a sequence. (Dai, Gao, and Liu 2021; Dosovitskiy et al. 2021; Vaswani et al. 2017).

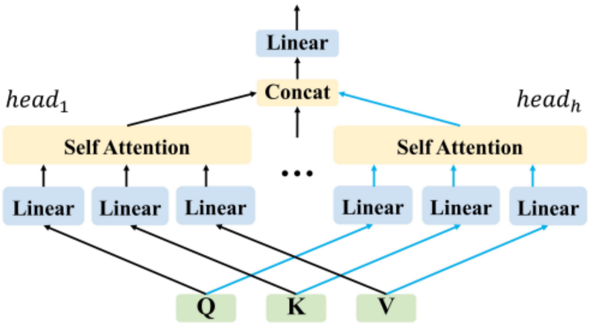

Figure 3: An overview of the Multi-Head Self-Attention (Dai, Gao, and Liu 2021).

3.3.1 Vision Transformers

The Vision Transformer (ViT) proposed by Dosovitskiy et al. 2021 was the first notable attempt for using a purely transformer-based architecture for computer vision related tasks, replacing the standard convolution operations and achieving impressive performance compared to the current state-of-the-art convolutional neural networks. However, training ViTs require a large amount of data which comes with a computational cost. Thus, Dosovitskiy et al. 2021 also proposed a hybrid ViT architecture, which conjugates the transformer with a CNN. In the hybrid ViT architecture, the CNN functions mainly as a feature extractor, while the transformer mainly focuses on global attention. The results of the hybrid ViT also showed comparable performance with the pure ViT models, but with relatively less amount of computational needs (Dosovitskiy et al. 2021).

An overview of the ViT model is shown in Figure 4, and following the presentation by Dosovitskiy et al. 2021, the training process can be explained as following:

For the two-dimensional input images, we first reshape the image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened two-dimensional patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 C)}$, where (H, W) stand for resolution of the original image, C is the number of channels, (P,P) is the resolution of each image patch, and $N = HW/P^2$ is the total number of patches resulted (Dosovitskiy et al. 2021).

The Transformer uses a constant latent vector with the size D throughout all of its layers, which means that the patches are flattened and projected to D dimensions with a trainable linear projection layer (Equation 3.6), the output of this projection is called patch embed-

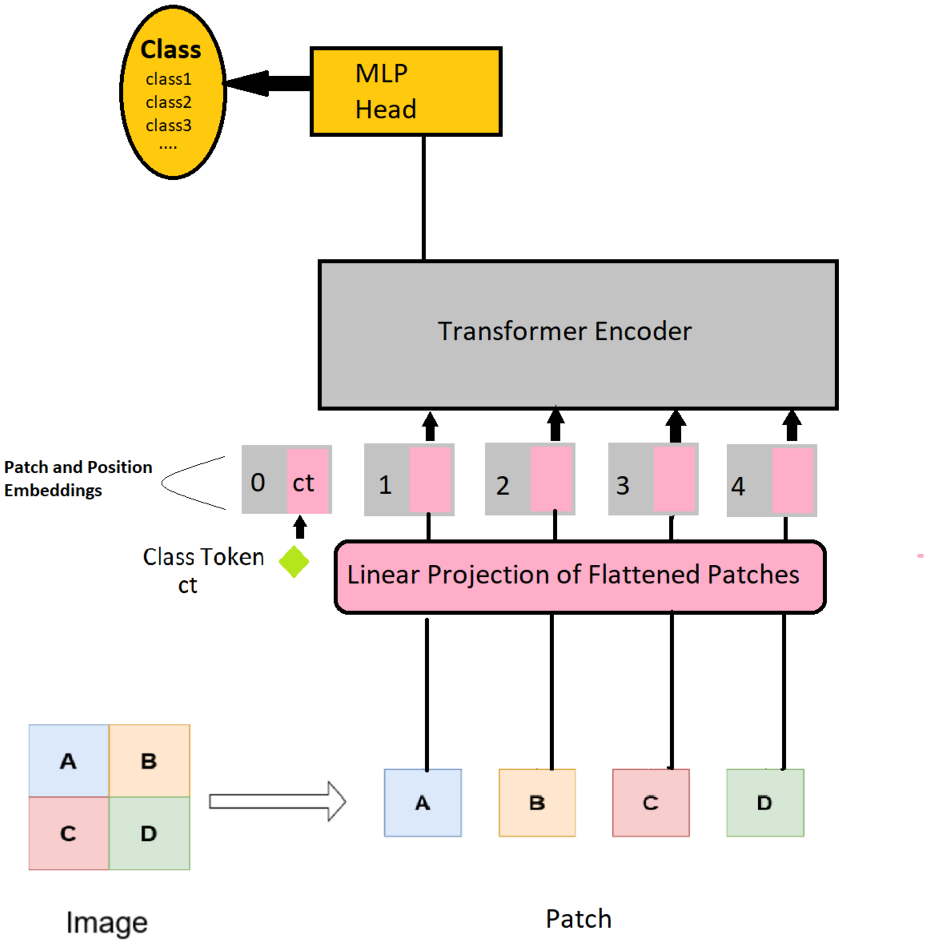

Figure 4: An overview of the Vision Transformer model and its training process. Adapted from (Dosovitskiy et al. 2021).

dings (Dosovitskiy et al. 2021).

A class token similar to the BERT-architecture (Devlin et al. 2018), is added to the beginning of the sequence of the embedded patches as $\mathbf{z}_0^0 = \mathbf{x}_{\text{class}}$. The class token serves as the representation of an entire image $\left(\mathbf{z}_L^0\right)$, where L equals to the last layer, which is used for classification purposes. During the pre-training and fine-tuning of the ViT, a classification head is added to $\left(\mathbf{z}_L^0\right)$ (Dosovitskiy et al. 2021).

One-dimensional position embeddings $\mathbf{E}_{pos}$ are then linearly added to the patch embeddings for providing the positional information of the images. This considers the inputs as a sequence of patches, such that the resulting sequence of embedding vectors act as an input to the ViT encoder (Dosovitskiy et al. 2021).

$$\mathbf{z}_0 = \left[\mathbf{x}_{\text{class}}\,; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}\right] + \mathbf{E}_{pos}, \quad \mathbf{E} \in \mathbb{R}^{\left(P^2 C\right)\times D}, \mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D} \tag{3.6}$$

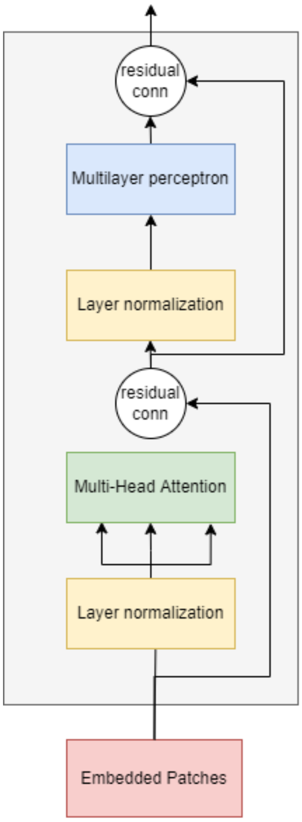

Figure 5: An overview of the Vision Transformer encoder layer. Adapted from (Dosovitskiy et al. 2021).

The ViT encoder consists (Figure 5) of alternating layers ℓ of MHSA (Equation 3.8) and Multi-Layer Perceptron (MLP) blocks (Equation 3.7). Layer normalization (LayerNorm) (Ba, Kiros, and Hinton 2016) is also applied before every Transformer block, and residual connections after every transformer block, Residual connection, also known as a type of skip connection provide an alternative path for the data to reach deeper parts of the model by skipping some layers (Kaiming He et al. 2016).

$$\mathbf{z}_{\ell} = \mathrm{MLP}\left(\mathrm{LayerNorm}\left(\mathbf{z}'_{\ell}\right)\right) + \mathbf{z}'_{\ell}, \quad \ell = 1 \ldots L \tag{3.7}$$

$$\mathbf{z}'_{\ell} = \mathrm{MHSA}\left(\mathrm{LayerNorm}\left(\mathbf{z}_{\ell-1}\right)\right) + \mathbf{z}_{\ell-1}, \quad \ell = 1 \ldots L \tag{3.8}$$

LayerNorm is a technique for normalizing the distributions of the models previous layers, which increases the models training speed and generalization accuracy (Xu et al. 2019). Following the presentation by Xu et al. 2019 LayerNorm can be defined as re-centering and re-scaling input vector representation $\mathbf{x} = (x_1, x_2, \ldots, x_H)$ of an input of size H to normalization layers as:

$$\mathbf{h} = \mathbf{g} \cdot N(\mathbf{x}) + \mathbf{b}, \quad N(\mathbf{x}) = \frac{\mathbf{x} - \mu}{\sigma}, \quad \mu = \frac{1}{H} \sum_{i=1}^{H} \mathbf{x}_i, \quad \sigma = \sqrt{\frac{1}{H} \sum_{i=1}^{H} (\mathbf{x}_i - \mu)^2} \tag{3.9}$$

where $\mathbf{h}$ is the output from the LayerNorm-layer, $\mathbf{b}$ and $\mathbf{g}$ are defined as the bias and gain parameters, which both have the same dimension as H, μ and σ represent the mean and standard deviation of $\mathbf{h}$ (Xu et al. 2019).

Unlike the transformer blocks in vanilla Transformers, Vision Transformers have no decoder layers and the outputs of the transformer encoder are instead sent into an MLP head to classify the classes for the learned image representation from the last layers class token. (Dosovitskiy et al. 2021).

4 Data

For our experiments, we used the patched version of the Camelyon16 challenge dataset as our target dataset (Veeling et al. 2018) and for the domain-based pre-training we use the Grand Challenge on Breast Cancer Histology images (BACH) challenge dataset by (Aresta et al. 2019) as the source dataset.

4.1 Camelyon16

Camelyon16 (Ehteshami Bejnordi et al. December 2017) based on the Camelyon16 grand challenge, is a dataset suitable for the classification and detection of breast cancer in Whole Slide Imaging. The data of Camelyon16 is originally from the Radboud University Medical Center and the University of Utrecht Medical Center. Camelyon16 is made up of 170 phase I lymph node WSIs, in which 100 are normal tissues and 70 are metastatic tissues, and 100 Phase II WSIs in which 60 are normal tissues and 40 metastatic tissues. The test dataset includes 130 WSIs from both medical centers. Figure 6 shows an example of a lymph node.

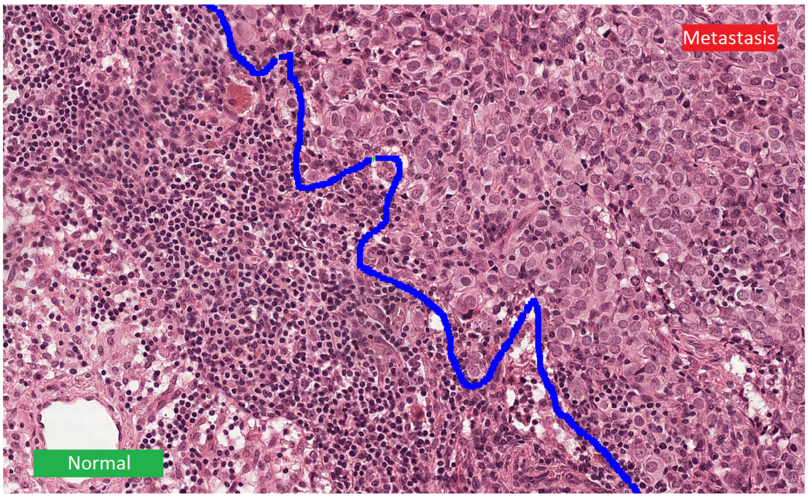

Figure 6: A pathological image of a lymph node in Camelyon16 (Ehteshami Bejnordi et al. December 2017). The left side referring to normal tissue while the right side has growth of metastatic cancer cells

4.1.1 PatchCamelyon

The PatchCamelyon benchmark (PCam) (Veeling et al. 2018) is a image classification dataset extracted from the Camelyon16 whole-slide-images lymph node sections. it consists of 327.680 images with a fixed size of 96x96 pixels. The two classes present in this dataset: normal and tumor tissues. Benchmark-wise PCam plays an important role in clinically relevant task of metastasis detection.

In this thesis, a modified PatchCamelyon dataset is used based on the histopathologic cancer detection competition organized by Kaggle due to the original PCam dataset containing duplicate images (Chandrasekhar et al. 2019).

With duplicate images removed the same data and splits as the PCam benchmark is maintained. The training set has 220,025 images, with 130,908 (60%) being benign images and 89,117 (40%) being images, where at the center of the 32×32 pixels image region exists at least one pixel of tumor tissue (Chandrasekhar et al. 2019). Samples from the dataset are shown in Figure 7.

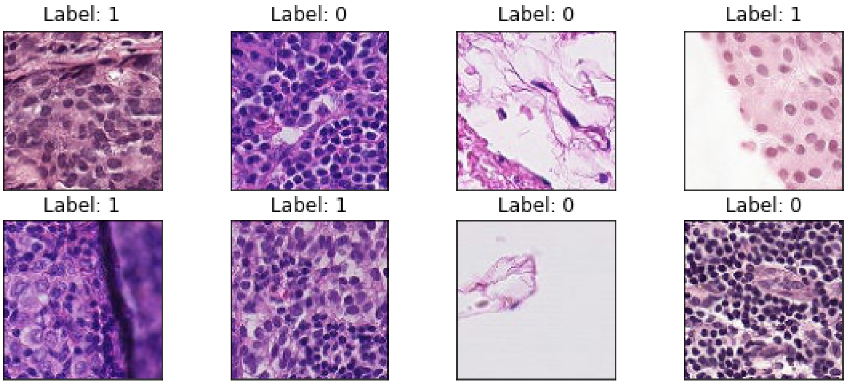

Figure 7: Sample images from the PCam-dataset. Label 0 referring to patches with no tumor tissue, and Label 1 referring to patches with at least one pixel of tumor tissue.

4.2 BACH

The Grand Challenge on BreAst Cancer Histology images (BACH) (Aresta et al. 2019) is aimed at the classification and localization of breast cancer in both microscopical images and whole-slide images from a large annotated dataset. For that, 400 training and 100 test samples, with equal class distribution, were given. The microscopical images are labeled as: normal, benign, in situ carcinoma or invasive carcinoma. Each class contains 100 images for the corresponding cancer. The annotation was done by two medical practicioners and if disagreement happened regarding the annotation, then those images were discarded from the data (Aresta et al. 2019). Samples from the dataset are shown in Figure 8.

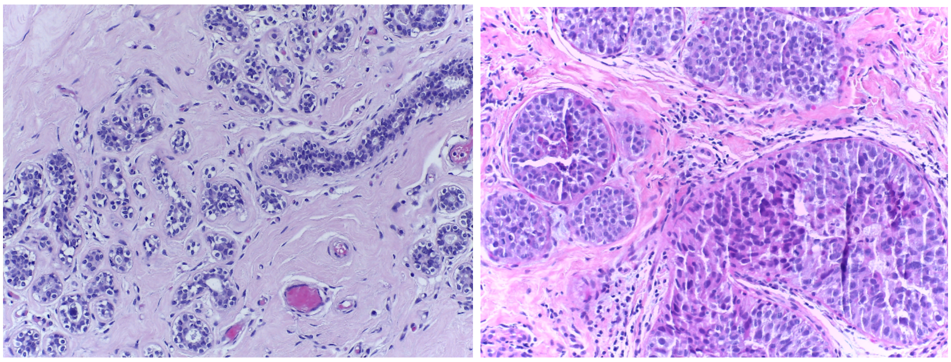

Figure 8: Example of a image labeled normal and in-situ (Aresta et al. 2019)

5 Methods

The following section of the chapter provides an overview of the test environment used for the experimental setup, the necessary preprocessing with an insight to data augmentation techniques and evaluation metrics used in our experiments. We present our model selections for both the CNN and Vision Transformer-based models before we go over the training and validation process for our experiments. Finally, we give an brief overview of how we evaluated our experiments to implement robust and efficient models.

5.1 Test Environment

For all experiments, training, validation and testing was conducted using PyTorch (Paszke et al. 2019) version 1.8.1 with CUDA 10.2 using a fixed seed number of 323 for reproducibility purposes, utilizing 3x NVIDIA Tesla P100 GPUs at most for the larger models. For the pre-trained models, we use the PyTorch Image Models (timm) library (Wightman 2019), which includes the pre-trained ImageNet weights for every model used in our experiments. For data augmentations the Albumentations library (Buslaev et al. 2020) is used.

5.2 Preprocessing and evaluation metrics

We split both datasets into training and validation sets respectively as shown in Table 1, while the test set is already predetermined in the PCam benchmark. The training set refers to the actual dataset, which the model is trained on to learn from the data. The validation set refers to an sample of data, which is used for frequent evaluation of the models, so during the validation the model sees the data but it never learns from it in this case. The test set is a sample of data used only once to evaluate the models only when they are finished training completely, the biggest difference from the validation set is that the test set is carefully curated regarding the selection of data samples.

All the experiments are conducted on the fixed PCam dataset to determine the best models for comparison purposes. The domain-based pre-training is done on the BACH dataset, which we modify for the purpose of maintaining a binary classification task. The original 4 labels normal, benign, in situ carcinoma and invasive carcinoma are now labeled as normal (tissue), and tumor (tissue) in which we combine the samples from the latter three original labels. Every image from both datasets is resized to 224x224 pixels to match the models input dimensions.

Since the PCAM-dataset is slightly imbalanced, the performance metric we use for evaluation purposes is a commonly used metrics also employed in the PCam-dataset classification (Veeling et al. 2018), which is the area under the receiver operating characteristic curve (AUC).

5.2.1 Evaluation Metrics

There are many metrics for evaluation purposes: the confusion matrix, cross validation, the receiver operating characteristic curve (ROC) and the area under the ROC curve (AUC) (Mridha et al. 2021).

ROC curve and AUC: The area under a receiver operating characteristic (ROC) curve, known as AUC, measures the performance of a binary classifier (Hanley and McNeil 1982). The value of AUC ranges mostly from 0.5 to 1, where the value of 0.5 represents that the model has no capacity to distinguish between positive and negative classes and the value 1 would correspond to a perfect classifier. Overall, AUC is a robust metric for evaluating the performance of classifiers since the calculations are based on the whole two-dimensional area under the entire ROC curve, and therefore involves all the possible classification thresholds (Melo 2013).

5.2.2 Data Augmentation

Data Augmentation is a technique for increasing the total amount of data by generating transformed copies of the existing data or creating new synthetic data from the already existing data. Examples of these transformations are such as horizontal flipping, vertical flipping, rotation, reflection, blur, and color-space transformations like altering the contrast of an image. (Shorten and Khoshgoftaar 2019).

One sample of data augmentation applies each above listed transformation to an original image to generate 6 new augmented images, this increases the amount of images in the dataset with new unseen training examples which often improves the robustness and performance of the model (Zheng et al. 2016).

Test-Time Augmentation

Typically, data augmentation is performed during a models training process. However, it can also be applied to the test dataset to obtain stronger performance and improved accuracy (Shorten and Khoshgoftaar 2019). Test-Time Augmentation (TTA) combines the models predictions from several augmented versions of a given test input to obtain a more confident prediction by averaging them (Shanmugam et al. 2021).

Since transformers require a larger amount of data. We rely on extensive data augmentation to train our transformer-based models with more examples, transformers requiring strong data augmentation was also noted by Touvron et al. 2021. Almost every data-augmentation method excluding dropout layers proved to be useful.

In our experiment we use the following heavy type of data augmentations with the Albumentations library (Buslaev et al. 2020):

- **RandomRotate90**, in which images are randomly rotated between +/- 90 degrees with a probability of 0.5
- **Transpose**, where the rows and columns of an image is swapped with a probability of 0.5
- **Flip**, where images are reflected over the central vertical line with a probability of 0.5
- **OneOf**, where we randomly apply only one of the following transforms with a probability of 0.5
  - **CLAHE** which limits the contrast amplification to reduce amplified noise
  - **Sharpen**, where the input image is sharpened
  - **Emboss**, where the input image is embossed
  - **RandomBrightnessContrast**, randomly changing the brightness and contrast of the image
  - **ImageCompression**, where the quality of the image is decreased

Table 1: Specifications of the target and source datasets after train-validation-test splitting

Dataset	Train	Val	Test	Classes
PCam	187021	33004	57458	2
BACH	359	41	-	2

  - **Blur**, where the input image is blurred
  - **GaussNoise**, where we apply gaussian noise to the input image
- **HueSaturationValue**, randomly changing the images hue, saturation, and value with a probability of 0.5.
- **ShiftScaleRotate**, randomly applying affine transforms: shifting and scaling the image with a rotation limit of +/- 45 degrees, shift and scale limit of 0.15 with a probability of 0.5.

the code of the whole augmentations for the training set can be seen in Appendix A.

We also apply test-time augmentation (TTA) for the test set, in which we include random rotation and random flip as the augmentations.

5.3 Model Variants

In this section we discuss our model selection for both the CNN baseline and the Vision Transformer-based models.

5.3.1 Baseline model

ResNet-50 (Kaiming He et al. 2016) pre-trained on ImageNet was chosen as the baseline model due to its wide use throughout literature (Dosovitskiy et al. 2021; Touvron et al. 2021; Gheflati and Rivaz 2021), as well as its relatively small number of parameters compared to other commonly selected CNNs like VGG16 (Simonyan and Zisserman 2014) and InceptionV3 (Szegedy, Vanhoucke, et al. 2015). We modify the last output layer modified to handle binary classification for outputting two classes.

5.3.2 Vision Transformer based models

The training process of the experiments in this thesis is mostly based on the recommendations provided by Steiner et al. 2021. Furthermore, the criteria for selecting the models for our approach regarding the ViT-based models is done based on the recommended ViT architectures provided by (Steiner et al. 2021), which includes various pure ViT models with different sizes (Ti/16, S/32, S/16 and B/16) (Dosovitskiy et al. 2021), but also the hybrid ViT models (R+Ti/16, R26+S/32). Regarding the size of the models: Ti meaning tiny, S meaning small and B meaning base (Dosovitskiy et al. 2021). In the hybrid models R represents the ResNet-architecture used for feature extraction, so the hybrid model is a combination of the ViT and ResNet models. The numbers 16 or 32 at the end of the model refer to the models patch size (16x16 or 32x32).

Every chosen ViT model is pre-trained on ImageNet, with the last output layer modified to handle binary classification to output two classes.

5.4 Training, Fine-tuning & Validation

The first experiments are conducted by fine-tuning the chosen models on the PCam dataset utilizing the conventional transfer learning approach. The second experiments are conducted by first pre-training the chosen models on the BACH dataset by utilizing domain adaptation in transfer learning. After that, fine-tuning is performed on the PCam dataset for the domain-based pre-trained models.

For the fine-tuning and domain-based pre-training of the baseline model, we use Adam as the optimizer (Kingma and Ba 2014) with an initial learning rate of 0.001 with all other parameters left default. Optimizers are short for optimization algorithms, which are used to train deep learning models, and learning rate controls how fast the optimizer can reach the minima of a loss function (Bengio 2012).

For the ViT-based models, both the fine-tuning and domain-based pre-training is conducted based on the results provided in (Steiner et al. 2021), so stochastic gradient descent (SGD) is chosen as the optimizer with an initial learning rate of 0.001 but leaving the parameter momentum as default (0) instead of the proposed value of 0.9 in Steiner et al. 2021.

For both the fine-tuning and domain-based pre-training of the models, a batch size of 128 is selected for the ViT-based models and 64 for the baseline respectively. Batch size refers to the number of data samples being processed before the model is updated (Bengio 2012).

A learning rate scheduler with a cosine decay is selected for both models with the minimum learning rate set as $1\mathrm{e}{-7}$, which means that we start at an initial learning rate and slowly reduce it in accordance to the scheduler. The loss function we use for both models is a Binary Cross-Entropy Loss with Logits from the torch library (Paszke et al. 2019), which creates a criterion that measures the Binary Cross-Entropy (BCE) between the target and the output, the logits referring to combining the loss inside a Sigmoid layer. BCE compares the predicted probabilities to the original class output, which can either have the value 0 or 1. BCE penalizes the probabilities, which are further away from the actual value (PyTorch 2022).

While training happens the training loss and training AUC is calculated for every batch of samples. Validation is set to happen every 100 batches, which is when we validate our models performance on the samples of the validation set for which we calculate the validation loss and validation AUC. We always save the best model based on the value of validation AUC.

Every model is set to be trained for 50 epochs (i.e. the number of times the models runs through the whole training set) with early stopping happening to stop the training if no improvement after a given number of events (patience) is happening for validation AUC. For early stopping, the value of 25 is selected for patience.

5.5 Evaluation

Evaluating deep learning models is an important step in implementing robust and efficient models. After the initial training and validation phase, the trained model is evaluated on with the test images to evaluate its performance. In this thesis the evaluation is done on the PCam test set where we evaluate whether the images contain traces of breast cancer or not.

Each of the best model variants from the baseline ResNet-50 to the ViT-based models chosen by the validation process previously mentioned from the both experiments are now further evaluated on the unseen test examples for which the classification performance is measured by the test AUC.

6 Results

We present the classification results for the CNN and ViT models on the PCam-dataset in two ways: first, using conventional transfer learning, and second, utilizing domain-based pre-training.

6.1 Classification results for transfer learning (PCam)

Table 2 shows the classification results of the pre-trained ViT-based models and the ResNet50 on the PCam validation set. The classification results from the implemented ViT models show a way better validation AUC score for S/16 and B/16 variation of the attention-based models than the corresponding results for the ResNet-50 model. According to the table, the best result of the ResNet-50 model is nearly comparable to the result of the smallest ViT Ti/16 model while having almost 5 times less model parameters. Meanwhile the best validation AUC is 0.9933 for the ViT B/16 model.

Table 3 shows the classification results of the pre-trained ViT-based models and the ResNet50 on the PCam test set. The results of the ViT models show a better test AUC score for every variation of the attention-based models expect the S/32 and R+Ti/16 variations compared to the corresponding results of the CNN model. According to the table, the best result of the ResNet-50 CNN model is worse than the result of the ViT Ti/16 model, which has almost 5 times less model parameters. Meanwhile, the best performing model is still ViT B/16 model with a test AUC of 0.97305.

6.2 Classification results for transfer learning with domain-based pre-training (PCam + BACH)

Table 2 shows the results of the pre-trained ViT-based models and the ResNet-50 on the PCam validation set after being pre-trained on the BACH-dataset. The results indicate a performance boost for almost every model, except the R+Ti/16, which actually had a lower validation AUC score compared to the conventional transfer learning method. ResNet-50

Table 2: Performance comparison of the ViT and CNN models on the classification of PCam validation set with and without domain-based pre-training on the BACH-dataset

Method	**Parameters**	**Validation AUC** (PCAM)	**Validation AUC** (PCam+BACH)
ResNet-50	25.55 M	0.9892	**0.9945**
Ti/16	5.71 M	0.9885	0.9896
S/32	22.87 M	0.9879	0.9891
S/16	22.05 M	0.9924	0.9928
B/16	86.56 M	**0.9933**	0.9933
R+Ti/16	6.33 M	0.9888	0.9861
R26+S/32	36.43 M	0.9867	0.9890

Table 3: Performance comparison of the ViT and CNN models on the classification of PCam test set with and without domain-based pre-training on the BACH-dataset

Method	**Parameters**	**Test AUC** (PCam)	**Test AUC** (PCam+BACH)
ResNet-50	25.55 M	0.96565	0.96995
Ti/16	5.71 M	0.96645	0.96580
S/32	22.87 M	0.96435	0.96605
S/16	22.05 M	0.9717	0.9723
B/16	86.56 M	**0.97305**	**0.97315**
R+Ti/16	6.33 M	0.96505	0.96280
R26+S/32	36.43 M	0.96905	0.9717

seemed to gain a huge increase in validation AUC compared to any of the ViT-based models. The best validation AUC score is 0.9945 for ResNet-50.

Table 3 shows the results of the pre-trained ViT-based models and the ResNet50 on the PCam test set after being pre-trained on the BACH-dataset. The results indicate a smallish boost of performance depending on the size of the model. Smaller models belonging to the Ti/16 -family actually suffer a performance loss compared to the conventional transfer learning method, but we do see a boost in performance for the other models, especially for the baseline ResNet50. The best AUC score is 0.97315 for the ViT B/16 model indicating a small increase of performance in classification.

7 Discussion

The previous chapter provided the results of this thesis. The results were shortly described, and in this chapter we present discussion about the results and its possible limitations.

Large scale datasets have been said to be required for obtaining desirable results with the Transformer architecture (Dosovitskiy et al. 2021). However, the availability of images and annotations can be quite limited in the field of medical image analysis.

In regards to our second research question, we show that Vision Transformer models, which we pre-trained on domain-specific datasets achieve better performance, than the ViT models, which are just fine-tuned for a task-specific classification for most variations of the ViT models. The interesting observation we made from the results in the section 6.2, is that the size of the BACH dataset, which we used for domain-based pre-training was only 359 training samples compared to the PCam dataset, where fine-tuning was performed with 187021 training samples. This demonstrates the possible effectiveness of utilizing domain associated pre-training for more accurate results even without needing large amounts of data, which has also been noted by Gheflati and Rivaz 2021 for ViTs and by Romero et al. 2019; R. Zhang et al. 2022 for CNNs. From this we could also even say that if we had potentially used a larger dataset for the domain-based pre-training, the results could have been even better, since as mentione, Transformers need large amount of pre-training, but in this case we demonstrated that even with a small amount of pre-training data, we can already achieve a increase in both validation and test AUC.

Some observations on why the Ti/16-family models performed worse after being domain-base pre-trained, which could be considered as their lack of parameters, which could translate into retaining less diverse information between two different histopathological datasets. Meanwhile, ResNet-50 had achieved the best validation AUC for domain-base pre-training, but on the test AUC it only performed better than the Ti/16-family models and the S/32. This could be because since CNNs are more focused in local variations in images and since the test set could contain different samples than the training set, thus the CNN might encounter difficulties understanding the unseen data. ViTs on the other hand, are able to extract more global information from images, which shows a steady performance for the ViT-based models. This answers our first research question, as four of the six of our implemented ViT models achieve better results than the ResNet-50 CNN on the PCam breast cancer classification test dataset. Transformer-based models outperforming the conventionally used CNNs has also been noted by numerous other studies (Chen, Li, et al. 2021; Gheflati and Rivaz 2021; Dai, Gao, and Liu 2021) in their respective datasets. Our best model B/16 also achieves comparable performance in terms of test AUC to the the current state-of-the-art result on the PCam dataset by Graham, Epstein, and Rajpoot 2020.

Another important observation in the results is that different ViT architectures are actually very close, in terms of both test and validation AUC for every ViT-model except the S/16 and B/16. Same kind of results were also noted by Gheflati and Rivaz 2021 for their selection of models. The similarities between the results in our ViT-based models shows that models with smaller patch size works better in regards to model size (i.e S/16 vs S/32), which has been noted in previous studies also (Steiner et al. 2021; Dai, Gao, and Liu 2021). Therefore, we could argue by our results that, if we should choose one best architecture for this particular classification task, then we should choose S/16, a small ViT architecture which achieved the second best performance in terms of AUC in our results, but in terms of its computational efficiency e.g. model parameter count, it is by far more convenient than the B/16.

The possible reasons why our implemented ViT-based models performed well on these types of breast cancer datasets might be that, unlike in natural images such as the ImageNet dataset, the relations and dependencies between the spatial information between the image patches is more connected in these types of histopathological datasets.

These findings show that ViT-based models can have great performance in the area of histopathological image classification due to their attention mechanism, but there remains a need for more implementations with ViT for breast cancer classification. ViTs achieve satisfactory and even state-of-the-art results in a lot of computer vision tasks, but due to being a recent discovery, overall more work is also needed with ViTs in the area of histopathological image classification and the exploration of the effects of domain-based pre-training (Chen, Li, et al. 2021).

8 Conclusion

Vision Transformers are slowly becoming the new rising trend in computer vision due to its outstanding performance and growing potential in utilizing self-attention mechanisms. CNNs have for long been dominant in the field of medical image analysis but, they are mostly limited at focusing on local variations in image patterns, in the field of medicine even an imprecise classification result might jeopardize lives. The introduction of the attention mechanism from the Transformer focuses on the global information of the image through the attention module . ViTs quickly achieved numerous state-of-the-art results in many computer vision tasks having surpassed the CNNs. The real potential of ViTs for medical image analysis is still yet to be fully discovered.

The main goals of this thesis was to evaluate different ViT architectures based on the recommended models and training strategy by Steiner et al. 2021 for breast cancer classification on the PCam-dataset. We compared our results to a conventionally used CNN architecture ResNet-50 in order to see how the ViT-based models perform. We also aimed to demonstrate whether the use of domain-based pre-training on the BACH-dataset could provide an increase in the classification performance. Our implemented ViT models obtained better results than the CNN-baseline in breast cancer classification on the PCam-dataset, while also having less model parameters. The use of domain-based pre-training showed an improvement in the results for every model except the Ti/16-family models.

Bibliography

Aneja, Sandhya, Nagender Aneja, Pg Emeroylariffion Abas, and Abdul Ghani Naim. 2021. “Transfer learning for cancer diagnosis in histopathological images”. *arXiv preprint arXiv:2112.15523.*

Aresta, Guilherme, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. 2019. “Bach: Grand challenge on breast cancer histology images”. *Medical image analysis* 56:122–139.

Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. “Layer normalization”. *arXiv preprint arXiv:1607.06450.*

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

Bengio, Yoshua. 2012. “Practical recommendations for gradient-based training of deep architectures”. In *Neural networks: Tricks of the trade,* 437–478. Springer.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. 2013. “Representation learning: A review and new perspectives”. *IEEE transactions on pattern analysis and machine intelligence* 35 (8): 1798–1828.

Bevers, Therese B, Benjamin O Anderson, Ermelinda Bonaccio, Sandra Buys, Mary B Daly, Peter J Dempsey, William B Farrar, Irving Fleming, Judy E Garber, Randall E Harris, et al. 2009. “Breast cancer screening and diagnosis”. *Journal of the National Comprehensive Cancer Network* 7 (10): 1060–1096.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. “Language models are few-shot learners”. *Advances in neural information processing systems* 33:1877–1901.

Buslaev, Alexander, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. 2020. “Albumentations: fast and flexible image augmentations”. *Information* 11 (2): 125.

Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. “End-to-end object detection with transformers”. In *European conference on computer vision,* 213–229. Springer.

Chandrasekhar, K, R Pavan, P Bharathi, and KV Triveni. 2019. “HISTOPATHOLOGIC CANCER DETECTION”. *IRJCS:: International Research Journal of Computer Science* 6:102–124.

Chen, Haoyuan, Chen Li, Xiaoyan Li, Ge Wang, Weiming Hu, Yixin Li, Wanli Liu, Changhao Sun, Yudong Yao, Yueyang Teng, et al. 2021. “GasHis-Transformer: A Multi-scale Visual Transformer Approach for Gastric Histopathology Image Classification”. *arXiv preprint arXiv:2104.14528.*

Chen, Jieneng, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. 2021. “Transunet: Transformers make strong encoders for medical image segmentation”. *arXiv preprint arXiv:2102.04306.*

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

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “Bert: Pre-training of deep bidirectional transformers for language understanding”. *arXiv preprint arXiv:1810.04805*

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2021. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* arXiv: 2010.11929 [cs.CV].

Ehteshami Bejnordi, Babak, Mitko Veta, Johannes P, al, Francisco Beca, Shadi Albarqouni, Rengul Cetin-Atalay, et al. December 2017. “Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer”. *JAMA* 318 (): 2199–2210. https://doi.org/10.1001/jama.2017.14585.

Fei-Fei, Li, Jia Deng, and Kai Li. 2009. “ImageNet: Constructing a large-scale image database”. *Journal of vision* 9 (8): 1037–1037.

Gecer, Baris, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. 2018. “Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks”. *Pattern recognition* 84:345–356.

Gheflati, Behnaz, and Hassan Rivaz. 2021. “Vision Transformer for Classification of Breast Ultrasound Images”. *arXiv preprint arXiv:2110.14731.*

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning.* MIT press. https://www.deeplearningbook.org.

Graham, Simon, David Epstein, and Nasir Rajpoot. 2020. “Dense steerable filter cnns for exploiting rotational symmetry in histology images”. *IEEE Transactions on Medical Imaging* 39 (12): 4124–4136.

Gurcan, Metin N, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. 2009. “Histopathological image analysis: A review”. *IEEE reviews in biomedical engineering* 2:147–171.

Hamid, Sofia, and Mrigana Walia. 2021. “Convolution Neural Network Based Image Recognition”. *International Journal of Science and Research Volume 10 Issue 2.*

Han, Changhee, Leonardo Rundo, Kohei Murao, Tomoyuki Noguchi, Yuki Shimahara, Zoltán Ádám Milacski, Saori Koshino, Evis Sala, Hideki Nakayama, and Shin’ichi Satoh. 2021. “MADGAN: unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction”. *BMC bioinformatics* 22 (2): 1–20.

Hanley, James A, and Barbara J McNeil. 1982. “The meaning and use of the area under a receiver operating characteristic (ROC) curve.” *Radiology* 143 (1): 29–36.

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

He, Kelei, Chen Gan, Zhuoyuan Li, Islem Rekik, Zihao Yin, Wen Ji, Yang Gao, Qian Wang, Junfeng Zhang, and Dinggang Shen. 2022. “Transformers in medical image analysis: A review”. *arXiv preprint arXiv:2202.12165.*

Khan, Salman, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2021. “Transformers in vision: A survey”. *arXiv preprint arXiv:2101.01169.*

Khuwaja, Gulzar A, and AN Abu-Rezq. 2004. “Bimodal breast cancer classification system”. *Pattern analysis and applications* 7 (3): 235–242.

Kingma, Diederik P, and Jimmy Ba. 2014. “Adam: A method for stochastic optimization”. *arXiv preprint arXiv:1412.6980.*

Klang, Eyal. 2018. “Deep learning and medical imaging”. *Journal of thoracic disease* 10 (3): 1325.

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

Kuutti, Sampo, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. 2020. “A survey of deep learning applications to autonomous vehicle control”. *IEEE Transactions on Intelligent Transportation Systems* 22 (2): 712–733.

Kwong, Timothy, and Samaneh Mazaheri. 2021. “A survey on deep learning approaches for breast cancer diagnosis”. *arXiv preprint arXiv:2109.08853.*

Lavecchia, Antonio. 2019. “Deep learning in drug discovery: opportunities, challenges and future prospects”. *Drug discovery today* 24 (10): 2017–2032.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep learning”. *nature* 521 (7553): 436–444.

LeCun, Yann, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. 1989. “Handwritten digit recognition with a back-propagation network”. *Advances in neural information processing systems* 2.

Li, Qing, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen. 2014. “Medical image classification with convolutional neural network”. In *2014 13th international conference on control automation robotics & vision (ICARCV),* 844–848. IEEE.

Litjens, Geert, 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. 2017. “A survey on deep learning in medical image analysis”. *Medical image analysis* 42:60–88.

Liu, Dongyu, Weiwei Cui, Kai Jin, Yuxiao Guo, and Huamin Qu. 2018. “Deeptracker: Visualizing the training process of convolutional neural networks”. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10 (1): 1–25.

Liu, Yongkai, Guang Yang, Sohrab Afshari Mirak, Melina Hosseiny, Afshin Azadikhah, Xinran Zhong, Robert E Reiter, Yeejin Lee, Steven S Raman, and Kyunghyun Sung. 2019. “Automatic prostate zonal segmentation using fully convolutional network with feature pyramid attention”. *IEEE Access* 7:163626–163632.

Melo, Francisco. 2013. “Area under the ROC Curve”. *Encyclopedia of systems biology,* 38–39.

Mridha, Muhammad Firoz, Md Hamid, Muhammad Mostafa Monowar, Ashfia Jannat Keya, Abu Quwsar Ohi, Md Islam, Jong-Myon Kim, et al. 2021. “A Comprehensive Survey on Deep-Learning-Based Breast Cancer Diagnosis”. *Cancers* 13 (23): 6116.

Napel, Sandy, and Sylvia K Plevritis. 2014. “NSCLC radiogenomics: initial Stanford study of 26 cases”. *Cancer Imaging Arch.*

National cancer institute. 2022. “The cancer genome atlas program”. Visited on January 20, 2022. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

Parvaiz, Arshi, Muhammad Anwaar Khalid, Rukhsana Zafar, Huma Ameer, Muhammad Ali, and Muhammad Moazam Fraz. 2022. “Vision Transformers in Medical Computer Vision–A Contemplative Retrospection”. *arXiv preprint arXiv:2203.15269.*

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. "Pytorch: An imperative style, high-performance deep learning library". *Advances in neural information processing systems* 32.

Pati, Pushpak, Guillaume Jaume, Antonio Foncubierta-Rodrıguez, Florinda Feroce, Anna Maria Anniciello, Giosue Scognamiglio, Nadia Brancati, Maryse Fiche, Estelle Dubruc, Daniel Riccio, et al. 2022. "Hierarchical graph representations in digital pathology". *Medical image analysis* 75:102264.

Pitkäniemi, Janne, Nea Malila, A Virtanen, Henna Degerlund, Sanna Heikkinen, and Karri Seppä. 2020. "Syöpä 2018". *Tilastoraportti Suomen syöpätilanteesta. Suomen Syöpäyhdistyksen julkaisuja nro* 93.

PyTorch. 2022. "BCELoss- loss function". Visited on May 14, 2022. https://pytorch.org/docs/stable/generated/torch.nn.BCELoss.html.

Rawat, Waseem, and Zenghui Wang. 2017. "Deep convolutional neural networks for image classification: A comprehensive review". *Neural computation* 29 (9): 2352–2449.

Richie, Rodney C, and John O Swanson. 2003. "Breast cancer: a review of the literature". *JOURNAL OF INSURANCE MEDICINE-NEW YORK THEN DENVER–* 35 (2): 85–101.

Romero, Miguel, Yannet Interian, Timothy Solberg, and Gilmer Valdes. 2019. "Training deep learning models with small datasets".

Sarker, Iqbal H. 2021. "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions". *SN Computer Science* 2 (6): 1–20.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin. 2015. "Facenet: A unified embedding for face recognition and clustering". In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 815–823.

Shanmugam, Divya, Davis Blalock, Guha Balakrishnan, and John Guttag. 2021. "Better aggregation in test-time augmentation". In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 1214–1223.

Shao, Zhuchen, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. 2021. “TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classication”. *arXiv preprint arXiv:2106.00908.*

Sharma, Ganesh N, Rahul Dave, Jyotsana Sanadya, Piush Sharma, and KK3255438 Sharma. 2010. “Various types and management of breast cancer: an overview”. *Journal of advanced pharmaceutical technology & research* 1 (2): 109.

Shorten, Connor, and Taghi M Khoshgoftaar. 2019. “A survey on image data augmentation for deep learning”. *Journal of big data* 6 (1): 1–48.

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

Smith, Robert A, Vilma Cokkinides, and Otis Webb Brawley. 2008. “Cancer screening in the United States, 2008: a review of current American Cancer Society guidelines and cancer screening issues”. *CA: A Cancer Journal for Clinicians* 58 (3): 161–179.

Spanhol, Fabio Alexandre, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. 2016. “Breast cancer histopathological image classification using convolutional neural networks”. In *2016 international joint conference on neural networks (IJCNN),* 2560–2567. IEEE.

Srinivas, Aravind, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. 2021. “Bottleneck transformers for visual recognition”. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 16519–16529.

Stegmüller, Thomas, Antoine Spahr, Behzad Bozorgtabar, and Jean-Philippe Thiran. 2022. “ScoreNet: Learning Non-Uniform Attention and Augmentation for Transformer-Based Histopathological Image Classification”. *arXiv preprint arXiv:2202.07570.*

Steiner, Andreas, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. 2021. “How to train your vit? data, augmentation, and regularization in vision transformers”. *arXiv preprint arXiv:2106.10270.*

Sung, Hyuna, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. 2021. “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. *CA: a cancer journal for clinicians* 71 (3): 209–249.

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. “Going deeper with convolutions”. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 1–9.

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. “Rethinking the inception architecture for computer vision. 2015”. *arXiv preprint arXiv:1512.00567.*

Tan, Mingxing, and Quoc Le. 2019. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In *International conference on machine learning,* 6105–6114. PMLR.

Tay, Yi, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. “Efficient transformers: A survey”. *arXiv preprint arXiv:2009.06732.*

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

Tuli, Shikhar, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. 2021. “Are Convolutional Neural Networks or Transformers more like human vision?” *arXiv preprint arXiv:2105.07197.*

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention is all you need”. In *Advances in neural information processing systems,* 5998–6008.

Veeling, Bastiaan S., Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. *Rotation Equivariant CNNs for Digital Pathology.* arXiv: 1806.03962 [cs.CV].

Waks, Adrienne G, and Eric P Winer. 2019. “Breast cancer treatment: a review”. *Jama* 321 (3): 288–300.

Wang, Dayong, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. 2016. “Deep learning for identifying metastatic breast cancer”. *arXiv preprint arXiv:1606.05718.*

Wang, Lulu. 2017. “Early diagnosis of breast cancer”. *Sensors* 17 (7): 1572.

Wang, Shuai, and Zhendong Su. 2019. “Metamorphic testing for object detection systems”. *arXiv preprint arXiv:1912.12162.*

Wang, Sinong, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. “Linformer: Self-attention with linear complexity”. *arXiv preprint arXiv:2006.04768.*

Weiss, Karl, Taghi M Khoshgoftaar, and DingDing Wang. 2016. “A survey of transfer learning”. *Journal of Big data* 3 (1): 1–40.

Wightman, Ross. 2019. *PyTorch Image Models.* https://github.com/rwightman/pytorch-image-models. https://doi.org/10.5281/zenodo.4414861.

Wu, Xiongwei, Doyen Sahoo, and Steven CH Hoi. 2020. “Recent advances in deep learning for object detection”. *Neurocomputing* 396:39–64.

Xie, Juanying, Ran Liu, Joseph Luttrell IV, and Chaoyang Zhang. 2019. “Deep learning based analysis of histopathological images of breast cancer”. *Frontiers in genetics* 10:80.

Xie, Saining, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. “Aggregated residual transformations for deep neural networks”. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 1492–1500.

Xu, Jingjing, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. 2019. “Understanding and improving layer normalization”. *Advances in Neural Information Processing Systems* 32.

Yang, Fuzhi, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. 2020. “Learning texture transformer network for image super-resolution”. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* 5791–5800.

Zeng, Yanhong, Jianlong Fu, and Hongyang Chao. 2020. “Learning joint spatial-temporal transformations for video inpainting”. In *European Conference on Computer Vision,* 528–543. Springer.

Zhang, Ryan, Jiadai Zhu, Stephen Yang, Mahdi S. Hosseini, Angelo Genovese, Lina Chen, Corwyn Rowsell, Savvas Damaskinos, Sonal Varma, and Konstantinos N. Plataniotis. 2022. *HistoKT: Cross Knowledge Transfer in Computational Pathology*. arXiv: 2201.11246 [eess.IV].

Zhang, Yu, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. 2020. "Pushing the limits of semi-supervised learning for automatic speech recognition". *arXiv preprint arXiv:2010.10504.*

Zheng, Stephan, Yang Song, Thomas Leung, and Ian Goodfellow. 2016. "Improving the robustness of deep neural networks via stability training". In *Proceedings of the ieee conference on computer vision and pattern recognition,* 4480–4488.

Appendices

A Data augmentations used for PCam training set

```
data_augmentations = albumentations.Compose([
    albumentations.Resize(224, 224),
    albumentations.RandomRotate90(p=0.5),
    albumentations.Transpose(p=0.5),
    albumentations.Flip(p=0.5),
    albumentations.OneOf
    ([
        albumentations.CLAHE(clip_limit=2),
        albumentations.Sharpen(),
        albumentations.Emboss(),
        albumentations.RandomBrightnessContrast(),
        albumentations.ImageCompression(),
        albumentations.Blur(),
        albumentations.GaussNoise()
    ], p=0.5),
    albumentations.HueSaturationValue(p=0.5),
    albumentations.ShiftScaleRotate(shift_limit=0.15, scale_limit=0.15,
                                    rotate_limit=45, p=0.5),
    albumentations.Normalize(),
    ToTensorV2(p=1.0)

    ])
```