TUHH
Hamburg
University of
Technology

Nagalikhitha Reddipalli

Relapse Prediction of Prostate Cancer from Histopathology Images using Vision Transformers

Institute of Medical Technology and Intelligent Systems in collaboration with the Group of Computational Pathology at the Institute of Medical Systems Biology (University Medical Center Hamburg-Eppendorf)
Building E | 21073 Hamburg
www.tuhh.de/mtec

Institute of Medical Technology and Intelligent Systems

A master thesis written at the Institute of Medical Technology and Intelligent Systems in collaboration with the Group of Computational Pathology at the Institute of Medical Systems Biology (University Medical Center Hamburg-Eppendorf) and submitted in partial fulfillment of the requirements for the degree Master of Science.

Author: Nagalikhitha Reddipalli

Title: Relapse Prediction of Prostate Cancer from Histopathology Images using Vision Transformers

Date: October 2, 2023

Supervisors: Marina Zimmermann (Dr.)
Alexander Schlaefer (Dr.-Ing.)

Referees: Prof. Dr.-Ing. Alexander Schlaefer
Prof. Dr. Marina Zimmermann

Declaration

I hereby certify that this thesis has been composed by me and is based on my own work, unless stated otherwise. No other persons work has been used without due acknowledgment in this thesis.

Date: .....

(Signature)

Contents

List of Figures

List of Tables

Abstract

Computational pathology, an emerging field in pathology, uses advancements in tissue scanning and computer vision to automate disease analysis, particularly in cancer tissues. This thesis centres on prostate cancer, the most prevalent non-skin cancer in men, with diagnosis relying primarily on tissue samples graded by pathologists using the Gleason score method. The reliability of diagnosis is, however, hampered by high interobserver variability in Gleason scores. Thus, this thesis makes a transition towards automatic cancer relapse prediction to identify whether a patient has a relapse within five years after the radical prostatectomy, especially using vision transformers (ViT). The primary topics investigated are how well vision transformers predict relapse using internal pathology data and how performance may be enhanced by adding domain-specific data pre-training from the publicly available Prostate cANcer graDe Assessment Challenge dataset (PANDA). The research also explores the effects of utilising several image resolutions in histopathology images with hierarchical vision transformers for analysis of histopathology images, with the goal of capturing information at various image scales for improved prediction accuracy. The findings spotlight a superior performance of vision transformers coupled with multiple instance learning (MIL) over convolutional neural networks with MIL, especially on the in-house data regarding relapse classification. The results show that ViT combined with MIL outperforms CNN with MIL when applied to the the internal data for relapse classification. However, the pre-training on PANDA dataset did not significantly improve the relapse prediction performance despite considerable hyperparameter testing and the use of multiple instance learning with different pooling strategies. The resulting performance of hierarchical vision transformers did not increase substantially over ViT with the MIL model, implying that hierarchical vision transformers may not be inherently appropriate for this task, given the amount of data. This study highlights the potential and pitfalls of using vision transformers in computational pathology, specifically in prostate cancer recurrence prediction, offering information on a path towards more precise prognostic models.

Keywords: Prostate cancer, multiple instance learning, vision transformers, hierarchical vision transformers, pre-training

1 Introduction

1.1 Motivation

Computational pathology is a discipline of pathology that aims to automate the disease analysis of histopathology images from pathology departments (Abels et al., 2019). Automated analysis of cancer tissue to extract helpful information for decision support has evolved as a result of the emergence of tissue scanners and simultaneous advances in computer vision techniques. Improving objectivity and speeding up cancer staging by using computer-based models to analyse cancer tissue automatically is possible. Other benefits include making better treatment decisions and identifying novel image components (Abels et al., 2019; Li et al., 2021b).

Prostate cancer is the most common non-skin cancer among men worldwide, with up to one in eight men diagnosed (Cancer.org, 2021) and is also the second most common cancer overall (Rawla, 2019). Several methods exist for prostate cancer diagnosis, among which tissue biopsy examination by pathologists remains the gold standard. The tissue is routinely stained with hematoxylin and eosin (H&E) to highlight structural patterns graded with Gleason scores by pathologists (Ikromjanov et al., 2022). However, the Gleason score varies largely from pathologist to pathologist, depending on the experience (Allsbrook Jr et al., 2001; Nagpal et al., 2019). Therefore, instead of Gleason score prediction, this thesis proposes to directly predict cancer relapse after radical prostatectomy, which could help reduce the bias and remove the necessity for annotations of Gleason grading. Current advancements in medical image analysis, including histopathology, to detect and grade tumours and to predict survival rates (Dietrich et al., 2021) using deep learning, have immensely impacted the scope of medical diagnosis at improved precision (Ikromjanov et al., 2022). Therefore, this thesis explores methods for automated relapse prediction that might help to improve the diagnosis of patients.

Whole slide images (WSI), which are digitized versions of tissue on the biopsy slide, enable a path to use deep learning in the field of pathology. However, the sizes of slides, e.g. of up to 100,000 $\times$ 100,000 pixels, pose a disadvantage in applying standard convolutional neural network (CNN), vision transformer based supervised learning techniques (Shao et al., 2021). To tackle these massive image sizes and obtain classification results, a technique like multiple instance learning is employed (Shao et al., 2021; He et al., 2022).

From a technical perspective, relapse prediction of prostate cancer using vision transformers is a relevant task since it has not been explored much on histopathology images. Several researchers have reduced relapse prediction to a straightforward binary task to predict whether a patient has relapse within a given time by using CNNs. This thesis primarily focuses on the same task using vision transformers and their extended version, hierarchical vision transformers. Transformers, which have paved their way from natural language processing to computer vision, are currently producing comparable or

better results to state-of-the-art techniques on images(He et al., 2022). In addition, the novel transformer architectures extract global information, in contrast to CNN, which is more locally focused (Chen et al., 2022a). Since their advent, several adaptations to vision transformers, like Swin Transformer (Cai et al., 2022), and Data efficient image transformer (DeIT) (Deininger et al., 2022), have been implemented on histopathology images. Hierarchical structural representation of pathology images on different scale images (Chen et al., 2022b; Li et al., 2022) are used for relapse prediction in our current work.

1.2 Research questions

This thesis aims to predict the relapse of prostate cancer patients after radical prostatectomy using histopathology images. The following research questions are addressed.

R1: Are vision transformers suitable for relapse prediction of in-house pathology data?

The research question pertains to evaluating the effectiveness of vision transformers in the context of relapse prediction using our in-house pathology dataset. Also, vision transformers are compared to a state-of-the-art CNN model.

R2: Can the performance of ViTs in relapse prediction be enhanced through the incorporation of additional domain-related data for pre-training?

As ViTs are data-hungry and require a vast amount of data, additionally, pre-training with publicly available prostate cancer data from the Prostate cANcer graDe Assessment (PANDA) Challenge dataset(Bulten et al., 2022) are explored in order to improve our models. The question mainly investigates the comparison between ImageNet pre-training and histopathology data pre-training.

R3: How does leveraging multiple image sizes in histopathology images impact prediction using hierarchical ViTs in medical diagnostics?

This research question focuses on utilizing a hierarchical ViT, which takes in different image levels to predict relapse and helps in obtaining information at different image scales.

1.3 Thesis outline

The outline of the thesis is presented below, describing the contents of each chapter.

In the **Background** chapter, medical and technical information is provided. In the medical background, prostate cancer diagnosis and brief details on histopathology images are provided. The basics of deep learning are explained in technical background.

In **State of the art**, current developments in computational pathology, along with its challenges and the theory of architectures used in this thesis, are explained. In recent

work, the papers involving cancer detection and survival prediction are summarized.

In the **Materials and Methods** chapter, the datasets utilised in this thesis are exemplified, and the approaches with their respective architectures are illustrated.

In the **Results** chapter, the results from all the experiments conducted are provided, and the results are discussed in the following **Discussion** chapter based on the research questions.

Finally, future steps are elucidated in the **Conclusion and Outlook** chapter by proving the conclusion of the thesis.

2 Background

An overview of prostate cancer epidemiology, relapse prediction of prostate cancer, and treatment methods are presented in the following chapter. The objective is to outline the relevant clinical background data for the scientific issue represented in this thesis and to illustrate why computer-assisted cancer relapse evaluation is essential.

Later in this section, a comprehensive explanation of machine learning concepts and an introduction to neural networks, convolutional neural networks, and vision transformers are provided.

2.1 Medical background

In this section, a brief description of histopathology, prostate cancer, diagnosis options and the grading of cancer are elucidated.

2.1.1 Prostate cancer

Prostate cancer is the most frequent cancer in men, with 70,100 predicted diagnoses in 2022, immediately followed by skin cancer (RKI, 2021). Also, prostate cancer-related deaths stand in fourth place. Although the risk of getting prostate cancer under 50 years is rare, a yearly check-up for males starting at 45 is included in Germany's statutory early prostate cancer detection program (RKI, 2021). The screening includes a digital rectal examination (DRE), which is free of cost, whereas a test for prostate specific antigen (PSA) has to be paid by the patient. DRE is a procedure where a healthcare professional examines the rectum and surrounding area, looking for signs of prostate cancer. Regarding the PSA test, a blood test is taken to check the levels of PSA in the bloodstream (Luiting and Roobol, 2019). PSA is a protein produced in the prostate gland. Increased levels of PSA indicate the risk of prostate cancer (Sohn, 2015). Prostate cancer has notably slow growth, which is mostly restricted to the prostate gland, and it can be very aggressive in later stages and can spread to nearby organs (Wessels et al., 2021). PSA levels alone cannot determine the presence of prostate cancer since it can be elevated due to other noncancer-related issues. Therefore, the PSA test alone cannot confirm prostate cancer.

The conceivable stages of a patient's care from diagnosis through treatment are shown in 2.1. The Anatomical location of the prostate is shown in the figure, which is underneath the blue urinary bladder and around the urethra, in orange colour (Sotelo et al.). A transrectal ultrasound scan biopsy is performed if a person is suspected of having prostate cancer during DRE or PSA test initial examination (Luiting and Roobol, 2019). In order to improve the localisation of the tumour, a magnetic resonance image (MRI)-guided biopsy can be taken (Luiting and Roobol, 2019). Once the biopsies are collected by these procedures, pathologists examine the biopsies visually. The so-called Gleason

score is assigned to each biopsy, indicating the severity of the prostate cancer. Apart from this, the quantification of the tumour is also considered for further treatment options (Grignon, 2018). Based on the severity of prostate cancer, doctors tend to choose treatment options ranging from radical prostatectomy (RPE) and radiotherapy to hormonal therapy (Heidenreich, 2007). Some patients with initial stages of cancer are just monitored regularly without applying any of the above treatment options. RPE is performed on patients with severe prostate cancer, involving the complete removal of the prostate from the body. Gleason score grading has high inter-observer variability leading to misdiagnosis of cancer. Depending on the studies, there is a large variation of about 1.7 % to 47 %, which are overdiagnosed, leading to prostate removal, which in turn leads to side effects (Loeb et al., 2014). Once the prostate is removed, the pathologists again examine the extracted tissue to under the condition properly and provide better treatment options (Grignon, 2018). The patient is constantly monitored after the RPE, as there is a chance of relapse of prostate cancer. This relapse is called biochemical recurrence (BCR), where PSA levels in the blood are elevated again.

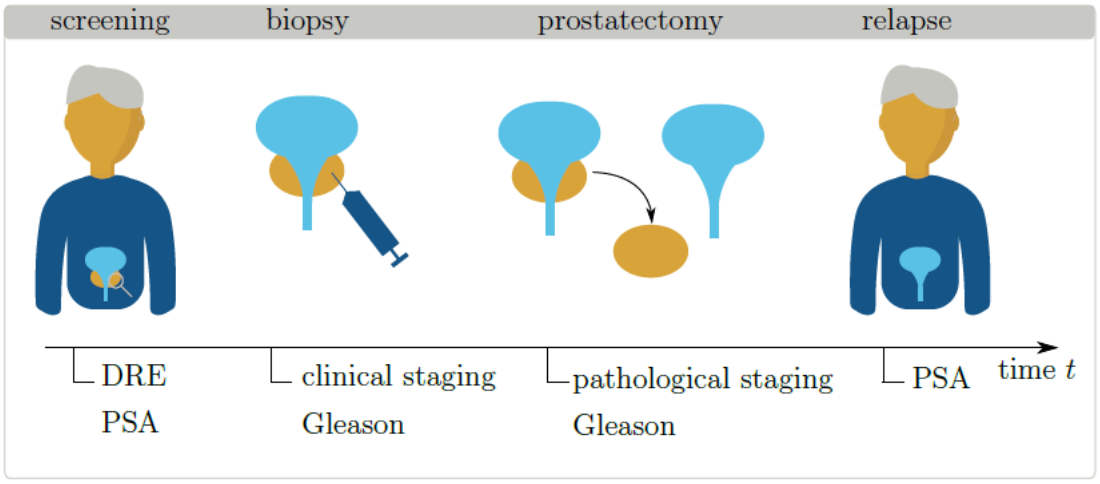


Fig. 2.1: Potential prostate cancer patient treatment pathway: A biopsy is carried out following a screening for prostate cancer that raises suspicion, such as a digital rectal exam (DRE) or a prostate specific antigen (PSA) value measurement. It is determined that a prostatectomy is required based on the clinical stage. The tissue is once more analyzed after prostate removal, this time using pathological staging to determine the best course of therapy. An examined PSA level can identify the potential for prostate cancer to spread once again (Dietrich, 2022).

2.1.2 Histopathology of prostate cancer

Histopathology is the scientific study of examining and interpreting architectural patterns in tissues under a microscope in a clinical context (Belsare and Mushrif, 2012). The tissue from biopsies or surgical samples has to be chemically processed, stained and mounted onto glass slides before being examined under a microscope. Because of the cells' transparency and colourless nature, they are stained with chemicals to distinguish between cell structures and highlight the structures of interest (Gurcan et al., 2009).

Hematoxylin and eosin (H&E) staining is one of the most common staining methods in histopathology, which gives nuclei a purplish blue colour and a brighter pink to the cytoplasm (Alwahaibi et al., 2015). The stained samples are now digitized for research purposes mostly. The digitized image data is stored in so-called whole slide images (WSI), which can amount to $100,000 \times 100,000$ pixels in dimensions and frequently surpass 1 GB in size (Campanella et al., 2019). These WSIs are stored in a pyramidal pattern, with the bottom of the pyramid being high-resolution images and the top being lower-resolution images and between the intermediate resolutions (Chen et al., 2022b).

2.1.3 Gleason patterns

A pathologist evaluates the prostate cancer tissue specimen and annotates it with Gleason patterns to determine tumour severity. Based on the architectural pattern of the glands, (Gleason and Mellinger, 1974) created the Gleason grading system, which categorizes glands into five distinctive patterns. Both tissues from a biopsy and an RPE can be graded with a Gleason pattern. There are several methods for using the Gleason patterns to stratify risk groups. The following sections summarize the several methods used to classify the cancer level.

Gleason grade

The architectural pattern of the glands is distinguished into five risk categories from grade 1 to grade 5. Grades 1 and 2 are no cancer or benign cancer. Grade 3 has slightly abnormal glands, grade 4 is malignant cancer, and the most severe cancer type is grade 5. As shown in figure 2.2, 1 and 2-grade glands are regular. With the increase in severity, the irregularity of the gland is increased (Egevad et al., 2012).

Fig. 2.2: Gleason grades of prostate glands. Grade 1 and 2 indicate no cancer, Grade 3 to 5 indicate the increasing severity of cancer with distorted glands (Furihata and Takeuchi, 2017).

Gleason score

A Gleason score is assigned to each part of the prostate cancer tissue, which consists of two Gleason grades, to determine the severity of the cancer. There is a slight difference in the Gleason scores assigned to tissue after biopsy and after RPE. For the tissue obtained after biopsy, the most common and the most severe patterns are taken into account. The most common and the second most common patterns are considered for the tissue after RPE (Gordetsky and Epstein, 2016). In case a tissue only has a single pattern, for example, Gleason pattern 3, a Gleason score 3+3 is assigned.

ISUP score

ISUP score is a new grading system which tries to address the issues in the previous system, which was proposed in 2014 at the International Society of Urological Pathology (ISUP) (Epstein et al., 2016). ISUP score ranges from 0 to 5, depending on Gleason scores, with lower values being benign and higher values being aggressive cancer. The relationship between the two systems is such that ISUP grade 1 corresponds to Gleason score 6, ISUP grade 2 corresponds to Gleason score 7 (3+4), ISUP grade 3 corresponds to Gleason score 7 (4+3), ISUP grade 4 corresponds to Gleason score 8, and ISUP grade 5 corresponds to Gleason scores 9 and 10.

2.1.4 Cancer relapse

In oncology studies, relapse-free survival is the time between patient disease treatment and cancer relapse (Cheon et al., 2016). The study of cancer recurrence plays a significant role in treatment options for patients and doctors by better estimating the condition to avoid misdiagnosis (Kumar et al., 2017). Usually, patients have a follow-up every six months after their treatment, where the PSA levels in the blood are checked. An observed increase in PSA level indicates prostate cancer relapse (Lobel, 2007). There is a possibility that certain patients never have a relapse in their entire lifetime, or patients might die due to some other reasons like metastasis, onset of other diseases, other than relapse.

2.1.5 Data acquisition

Digital images are necessary for a computer-aided support system for the detection of prostate cancer and binary survival prediction. The steps necessary to obtain digital images from prostate tissue are illustrated in Figure 2.3. There are several techniques by which prostate tissue is collected. When the patient is suspected of having prostate cancer, a biopsy is taken with a hollow needle from the cancerous region of the prostate, which is under the urethra. The biopsy is embedded into a paraffin block. Then the paraffin block is sliced into $1 - 10\mu m$ thin sections (Junqueira and Carneiro, 2005). Each section is then stained with H&E as the preparatory step for visualizing in the microscope. In histopathology, H&E is common staining, resulting in purple to pink colours. Acidic structures bind to hematoxylin, giving a dark blue or purple colour to the cell nuclei, whereas eosin binds to basic structures, such as cytoplasm in cells leading to a pink colour (Junqueira and Carneiro, 2005). After the staining, pathologists view the tissue under a light microscope and assign a Gleason score that reflects the severity of the cancer. For digitization, the stained tissues are scanned with a scanner. If the prostate

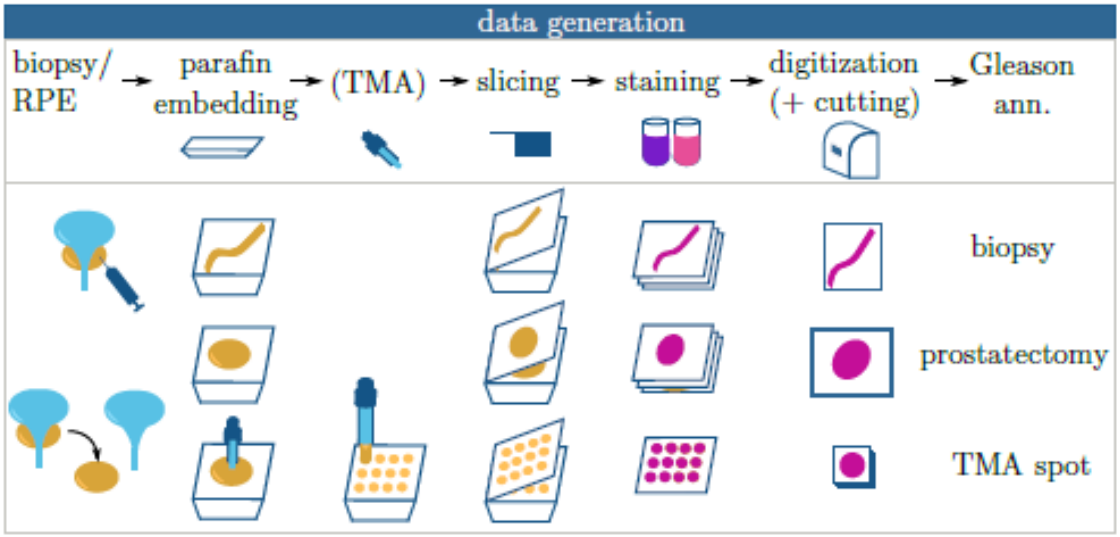


Fig. 2.3: Overview of data acquisition. **Biopsy and prostatectomy:** The collected prostate tissue is embedded into the paraffin block and sliced entirely. Each slice is digitalized after staining. Hence, the whole biopsy or the whole prostate is examined. **TMA spot:** After prostate removal using the RPE procedure, several tissue cores from different regions of the prostate are collected and placed on the single tissue microarray. Each TMA may have spots from multiple patients. Later, it is sliced, stained and digitalized. TMA spots can be cut separately for research purposes. RPE: Radical prostatectomy, TMA: Tissue microarray, ann.: annotation (Dietrich, 2022).

cancer is severe, the prostate is removed by RPE, and multiple cores are extracted from the removed prostate for research purposes, placing them on tissue microarray (TMA), and a similar procedure to the one mentioned above takes place (Parsons and Grabsch, 2009). Pathologists look at the whole prostate tissue to decide the treatment options. For research purposes, a single TMA spot is used (Simon et al., 2004).

2.2 Technical background

In the following section, the fundamental principles of machine learning, deep learning and the key concepts utilized in this thesis are explained.

2.2.1 Machine learning and Deep learning

Machine learning, a subfield of artificial intelligence, automatically enables computers to learn meaningful associations and patterns from examples and observations by using algorithms (Dutton and Conroy, 1997). Data required for machine learning algorithms to extract relevant features to make accurate predictions depends on the task, e.g. for simpler tasks like classification, less amounts of data are sufficient, whereas for deep learning methods with complex tasks, huge amounts of data are required (Jordan and Mitchell, 2015). In this section, the fundamental concepts of supervised learning, weakly supervised learning, and optimization process are briefly summarized by following

definitions in Mitchell and Mitchell (1997); Goodfellow et al. (2016).

Supervised and weakly supervised learning

In supervised learning, a set of input and output pairs are used to obtain the relationship between them, and the model, which is a system that is trained on the given data to recognize patterns, predicts the outputs for new data based on the learned patterns (Cunningham et al., 2008). In machine learning, data in structured form is treated as features and underlying information is obtained for further prediction of similar kinds of data. In the context of image analysis in machine learning, the input data are the features extracted from images, and the output data are the labels associated with each image. Deep learning is a subset of machine learning, which has algorithms inspired by neural networks in the brain and is used to recognize complex patterns in images, videos, text etc. In deep learning, images are given as input, and the features are automatically extracted, making supervised learning feasible for medical image analysis (Aljuaid and Anwar, 2022).

Weakly supervised learning is a machine learning approach in which the model is trained with data that is only partially labelled to obtain the desired results (Zhou, 2018). This method intends to find patterns or characteristics in data with little supervision, making it cost-effective. In medical imaging, where the acquisition of annotated data is very expensive, time-consuming and difficult, weakly supervised learning comes in handy. In recent times, multiple-instance learning, one of the weakly supervised algorithms, is playing a significant role in predicting diseases by having just one label for the WSI (Li et al., 2021a).

Artificial neuron

An artificial neuron or perceptron serves as the primary computing unit of an artificial neural network (ANN). This neuron is comparable to the biological neuron, which fires when information is received and transmits the information to the linked neurons (McCulloch and Pitts, 1943).

As shown in Figure 2.4, each neuron receives m inputs, each with a weight of $w_{i,m} \in \mathbb{R}^m$, from nearby nodes or external sources.

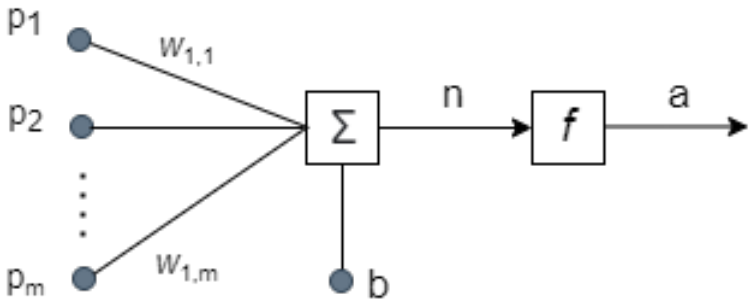


Fig. 2.4: Schematic diagram of an artificial neuron with activation function adapted from Demuth et al. (2014).

The weighted sum, obtained by the dot product of weights w and inputs p, is added to an additional bias component, b, making up $\mathbf{n}$. Equation

$$\mathbf{a} = f(\mathbf{n}) = f(\mathbf{w} \cdot \mathbf{p} + \mathbf{b}), \tag{2.1}$$

which is in vector form, results from applying an activation function on $\mathbf{n}$. For the activation function f, a variety of linear and non-linear functions. Non-linear functions are used to obtain non-linear relationships within the data by allowing models to capture and learn patterns which are not captured by linear activations. Some of the activation functions are Sigmoid, and GeLU, which are explained in 2.2.1.

Multi layer perceptron

Multi-layer perceptron (MLP) is the combination of artificial neurons in multiple layers (Gardner and Dorling, 1998). An example of an MLP is shown in Figure 2.5, which has an input layer which takes m inputs, one output layer with k outputs and the hidden layers that connects input and output layers. The input and hidden layers have a bias term, b, which is added to the sum of the weights multiplied with the inputs. The weight matrix $W \in \mathbb{R}^{l \times m}$ connects multiple inputs with multiple weights, in which l is the number of neurons in the next layer, and m is the neurons in the input layer. The formulation of the MLP from the Figure 2.2 is given by the equation,

$$\mathbf{y} = f^2(\mathbf{W}^2 f^1(\mathbf{W}^1 \cdot \mathbf{p} + \mathbf{b}^1) + \mathbf{b}^2), \tag{2.2}$$

with W_i as the weight matrix from each layer, p is the first layer input, b^i is bias terms, f is the activation function, and y is the output. The superscript in equation 2.2 and in Figure 2.5 refer to the corresponding MLP, which is also referred to as the neural network layer. MLP is a feedforward neural network widely used in deep learning models.

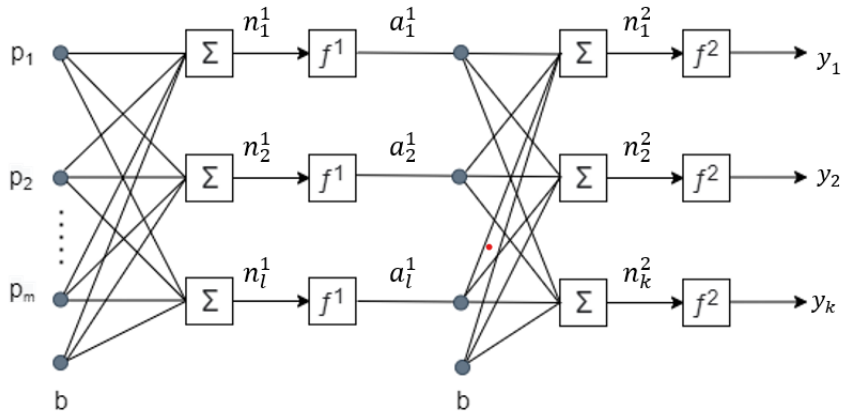


Fig. 2.5: Schematic diagram of multi-layer perceptron adapted from (Demuth et al., 2014).

Activation functions

Activation functions are key components in neural networks, which help models learn complex information through non-linearity (Sharma et al., 2017). The sigmoid function is one of the common activation functions implemented in binary classification tasks, where the task is to classify the input into either of the classes (Han and Moraga, 1995). The sigmoid function, also called a logistic function, outputs a value between 0 and 1, which in turn is used to estimate probabilities for the given input, which is described in the formula 2.3. When z is extremely negative or extremely positive, this activation function saturates to 0 and 1, respectively, as shown in 2.6.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.3}$$

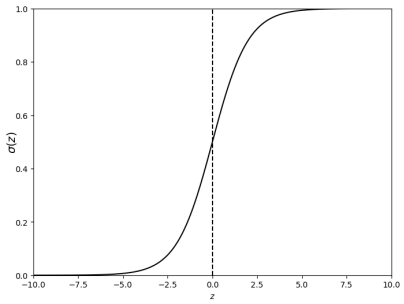


Fig. 2.6: Sigmoid activation function adapted from (Narayan, 1997).

2.2.2 Model training

During the training phase, machine learning models learn patterns and related representations from the data in order to predict the outputs accurately. During training, models iteratively adjust the weights based on the optimization algorithm to minimize the chosen loss function, which tries to bring predicted values closer to ground truths, which are actual or true labels associated with the input data (Wei et al., 2019).

Datasplits: The data used in ML models are split into training sets, validation sets, and test sets. The training set is used to train the model for predictions. The validation dataset is utilized to evaluate the model and tune the hyperparameters for increased performance (Xu and Goodacre, 2018). Hyperparameters are external settings for algorithms that are not learned from data but are set prior to training (Probst et al., 2019). In the end, the trained model is tested on test datasets, which are unseen during training and validation. One of the ways of dataset splitting is in the ratio 80:10:10, with training being the largest and validation and test sets with the same number of samples.

Hyperparameters: A significant amount of parameters in MLP in machine learning models are tuned during training to produce a good approximation of the mapping function (Probst et al., 2019). Hyperparameters are additional parameters that are not directly tuned during training. The common hyperparameters are batch size, learning rate, and weight decay.

The learning rate determines the step size for each iteration during the training process (Jacobs, 1988). While a low learning rate might result in gradual convergence or the algorithm becoming trapped in a suboptimal solution, a high learning rate could help the algorithm to converge fast but could also cause it to overshoot the ideal solution.

Loss functions: The functions that calculate the error between predicted values and ground truths are termed loss functions or cost functions (Janocha and Czarnecki, 2017). The goal is to minimize the loss function in order to bring actual and predicted values closer together by model training. For binary classification problems, binary cross entropy (BCE) is employed widely (Ho and Wookey, 2019). The dissimilarity between predictions and the true labels is observed in BCE loss. The loss is calculated for each sample and is averaged over the whole dataset. The BCE loss is given by the formula,

$$BCELoss = \frac{1}{N}\sum_{i=1}^{N} -(y_i * log(p_i) + (1 - y_i) * log(1 - p_i)), \tag{2.4}$$

where N is the total number of samples, y_i is ground truth, p_i is the prediction and suffix i indicates the sample. The first part of the equation 2.4 is activated when the sample corresponds to binary class 1 or positive, and for the class 0 sample, the second part is activated.

Optimization: The strategy of iteratively modifying a model's internal parameters to minimize a predefined loss function is called optimization in deep learning, and the algorithms that aim to find the best internal parameters, weights and biases to produce the best possible output are optimizers (Sra et al., 2012). Gradient descent and adaptive moment estimation optimizer (Adam) are some of the most common optimizers (Ruder, 2016; Kingma and Ba, 2014). The efficiency of the model is directly influenced by optimization algorithms used while training the model. The learning rate is changed for each network weight separately using the Adam optimizer. The optimizer takes past gradients and their second moments into consideration for dynamically computing the individual learning rate. Adam has a faster running time, lower memory requirements and less tuning of parameters than other optimizers making it widely used.

Generalizability: The ability of deep learning algorithms to function effectively on unexplored real-world data is a crucial prerequisite, and this is referred to as generalizability (Kawaguchi et al., 2017). The models are trained on the training dataset, and two separate datasets, validation and test dataset mentioned, are set aside to check the generalization of the models (Xu and Goodacre, 2018). The model is said to be generalizable when it performs well on the validation and test set, along with good performance metrics on the training set, where the validation set is used to optimize the model training and the test set is used to calculate the performance of the model. If the model performs poorly, it is indicated by either so-called overfitting or underfitting (Lever et al., 2016). In the case of overfitting, a common problem, the model learns to perform well on the training data, indicated by high training accuracy or low training loss. However, with these models, validation and test dataset performance are poorer. In underfitting, the chosen model is unable to capture the underlying features and performs poorly on both the training set and also on test datasets. Regularization methods, described in the section below, are used during the training to overcome this issue.

2.2.3 Regularization

Regularization is a technique used in machine learning, deep learning and statistical models to prevent overfitting and improve the model's generalization performance (Kukačka et al., 2017). Regularization methods aim to find a balance between fitting the training data well, keeping the model's complexity in check, and giving better performance on validation and test data. Dropout and early stopping are used in this thesis as the regularization methods.

Dropout is one of the most common strategies designed for regularization in neural networks. Dropout arbitrarily sets a portion of the neurons in a neural network layer to zero during each update, thus dropping out those units throughout training (Baldi and Sadowski, 2013). The random dropout causes distinct subsets of neurons to fire throughout each training iteration, forcing the network to acquire more reliable and independent features. Dropout decreases the network's dependence on specific neurons or neural network configurations, preventing overfitting. The entire network with all units is typically used during inference or testing.

According to the generalization performance of a validation set, a model's training can be terminated early before it has fully converged. It entails keeping track of the model's performance while being trained on a different validation dataset. When the validation performance begins to decline, signalling that the model is about to overfit, the training is stopped. By **early stopping**, the model's generalization performance is improved, preventing it from becoming too specialized on the training data (Yao et al., 2007). For early stopping, usually, the patience parameter and the threshold parameter are defined. The patience parameter limits the number of epochs or iterations to wait before stopping the training once the validation performance worsens. The threshold parameter defines the minimum change in the validation performance that is considered significant. Training is stopped if the validation performance does not improve by at least the threshold value within the patience period.

2.2.4 Convolutional neural networks

Convolutional neural networks are a type of feedforward neural networks that are popularly used for computer vision tasks (LeCun et al., 1998). Building upon the foundational structures of ANNs, CNNs can perform image classification and object identification tasks more efficiently than manual feature extraction with traditional machine learning. A CNN has several main layers like convolutional layers, pooling layers and fully connected layers.

Convolutional layers

Convolutional operation in the context of CNNs is expressed as

$$F(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n), \tag{2.5}$$

where I is the input to the convolutional operation, with F as the feature map, which is the output of one filter applied to the previous layer, K is the kernel or filter, which moves across the input data, executing element-wise multiplication and combining the

results to produce a single value in the feature map, and i, j are the rows and columns. Sparse interaction, parameter sharing and equivariant representations are the three significant advantages of convolution. In CNNs, there are fully connected layers, where every neuron in one layer is connected to every neuron in the next layer. With the sparse interactions, remove the necessity of fully connected neural networks, where neurons in one layer are connected to a small number of neurons in the next layer, making CNNs computationally efficient. Kernels are primarily used for pattern identification by convolutional operation by extracting localized features. In the first few layers, kernels can identify basic patterns like edges and textures, while the deeper layers can reveal sophisticated, abstract patterns. Often, filter or kernel sizes of 3×3, 5×5 are used. A stride S is the stepsize with which the kernel is shifted on input data. Padding is the process of adding extra pixels or data to an input image's or feature map's border and is used to maintain the specific size requirements. The reuse of kernels for various input areas is made possible by parameter sharing, which also improves the storage efficiency of convolutional processes compared to fully connected layers. Equivariance to translation denotes that when the input of a convolution operation is translated, the output is also translated in the same manner. In images, the change in the input image is reflected as the shift in the output image, which is helpful in identifying edges or multiple similar features at different locations in images.

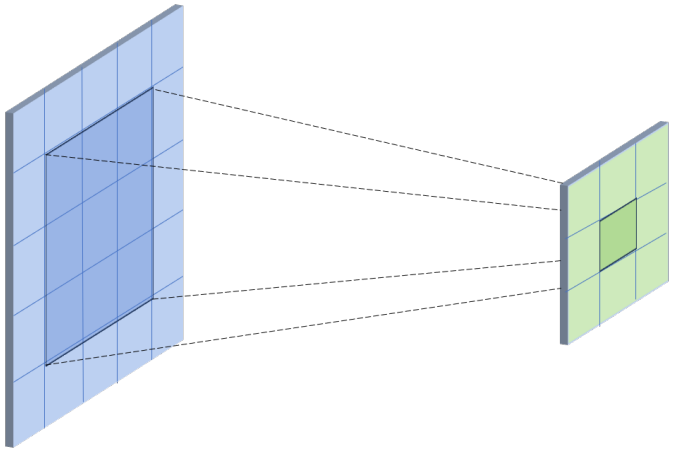


Fig. 2.7: 2D convolutional operation with kernel size 3×3, no padding and stride 1

Pooling layers

CNNs often contain pooling layers which seek to downsample the feature maps to minimize the spatial dimensions of data and computational costs in the network. Pooling layers are also advantageous for preventing overfitting and better generalization by providing a kind of spatial invariance, allowing the network to recognize features from various positions. The input for pooling operations is a tiny region called a receptive field with pooling or window size often selected as 2×2, 3×3, and the output is a single integer that serves as the region's representation. The typical approach for calculating the representative value of a receptive field involves using either the max function, known as max-pooling, or the average function, referred to as average pooling.

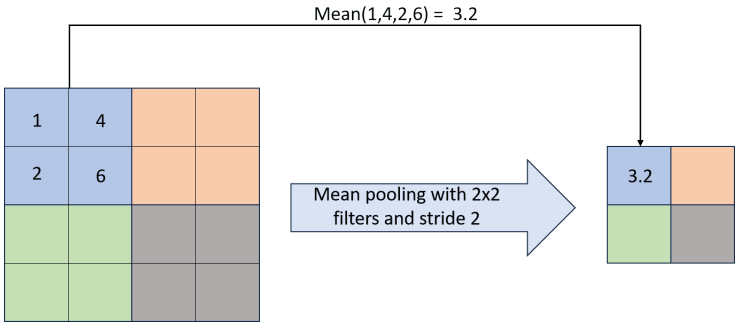


Fig. 2.8: Mean pooling operation with kernel size 2×2 and stride 2.

2.2.5 Transformers

Transformers have emerged as a groundbreaking architecture that has dramatically impacted the field of natural language processing (NLP). Since their introduction in the paper "Attention is all you need" (Vaswani et al., 2017), transformers have revolutionised sequential data processing. Transformers operate on the self-attention mechanism to capture long-range dependencies, unlike other sequential models like recurrent neural networks (RNNs). Since their outstanding performances in NLP, they have been adapted to computer vision tasks and called vision transformers (ViT) and outperformed CNNs, becoming state-of-the-art models in the imaging domain. In this section, the core concepts of attention and architectural components of transformers are explained in detail. Further explanation on Vision transformers is provided in detail in the next chapter 3.3.3.

Transformer architecture

Standard transformers contain an encoder and decoder structure as shown in Figure 2.9. The input series of symbol representations $(x_1, ..., x_n)$ is mapped by the encoder into a sequence of continuous representations $(z = (z_1, ..., z_n))$. This z is later fed into the decoder, which outputs the sequence $(y_1, ..., y_n)$ of symbols, each element sequentially. As a result of the model's auto-regressive characteristic, the previous symbols, y_{n-1}, are employed as extra input to the decoder to create the subsequent output. The transformer architecture incorporates stacked self-attention, which is explained in 2.2.5 and fully connected feed-forward layers in the encoder and decoder components.

Encoder and Decoder The encoder and decoder of the initial transformer architecture have N identical layers, as shown in Figure 2.9. Each encoder layer consists of two sub-layers. One is multi-head self-attention, and the other is a position-wise fully connected feed-forward network, with each layer having a residual connection before normalization. LayerNorm(x + Sublayer(x)) is the output of each sub-layer, where x is the input and Sublayer(x) is the output of each sub-layer function mentioned above. The decoder has similar layers to the encoder but with one additional layer, the so-called mask multi-head attention, which takes in the output from the encoder. This masking helps the model guarantee that the predictions made for position i solely rely on the known outputs at positions preceding i. This is used in inference ensuring training consistency to obtain relationships in the data.

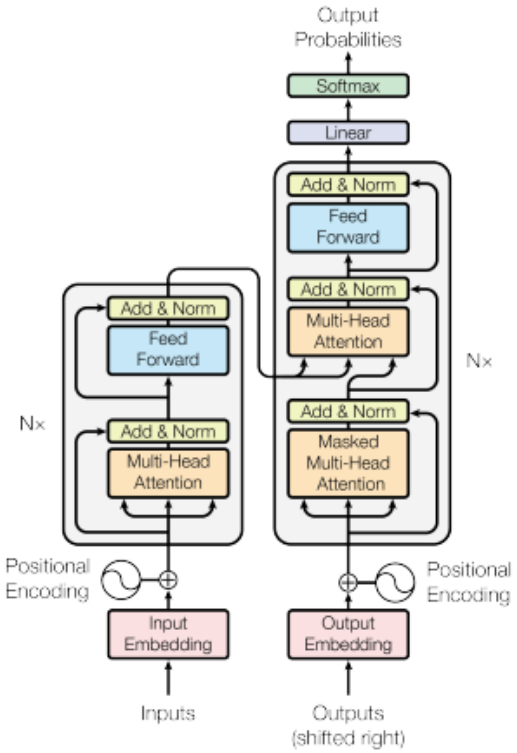


Fig. 2.9: Transformer architecture taken from (Vaswani et al., 2017).

Attention mechanism The attention mechanism is described as a function that maps a query to a collection of key-value pairs and then generates an output in which all parameters, key, value, and query are vectors. A weighted total of the values is used to compute the result, and each key's weight is determined by how similar the query is to each key. The weights show how important each value is to the result.

The scaled dot product is utilized in the architecture, which is shown in 2.10. The input matrices Q, K, and V, which represent Query, Key, and Value, are fed into the first layer. The Q matrix denotes the queries, which is the information the model aims to focus on. The K matrix symbolizes Keys, which are used to compare to queries and assess the relevance of different sections of the input sequence. The V matrix expresses values with the actual content or information associated with each input sequence. As described in the formula 2.6, the dot product of Q and K is computed and divided by $\sqrt{d_k}$, where d_k is the dimension of K. The softmax is applied to compute weights, which are later multiplied with V to get attention scores. Multiplicative or dot-product attention is one of the two most commonly used attention mechanisms, with the other being additive attention. $\frac{1}{\sqrt{d_k}}$ is the additional scaling factor used in the transformer network, which differs from the regular dot-product attention mechanism. Division by $\sqrt{d_k}$ is employed to reduce the impact of large dot-product values, which might cause the softmax function to move into regions with weak gradients.

Instead of performing attention once, multi-head attention performs it h times in parallel,

where h is the number of heads. The attention is performed in parallel on the input sequences Q, K and V to yield d_q, d_k and d_v dimensions with different learned projections, respectively. The outputs from each head are then concatenated to the final values. By using the multi-head attention strategy, the model is able to simultaneously capture data from various representational subspaces. The transformer has more power to encode several relationships for each word in the input due to this approach.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.6}$$

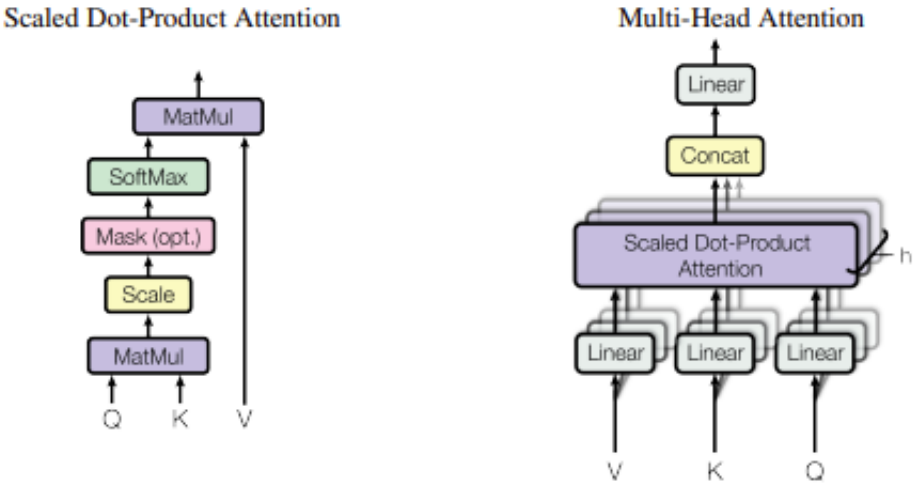


Fig. 2.10: Scaled dot-product attention (left) and multi-head attention (right), which consists of several attention layers taken from (Vaswani et al., 2017). Q, K, V indicates query, key, and value respectively.

2.2.6 Transfer learning and fine-tuning

A neural network's weights are not deliberately specified, instead, they are learnt through training following random initialization. As models are highly data-driven, the performance of the model depends on the training data. Each deep learning model has millions of parameters to be trained, which requires a sufficient amount of data. (Raghu et al., 2019) has called a dataset of $5,000$ samples small, whereas several papers worked on as small as below 100 samples, depending on the complexity of the problem. Transfer learning is commonly used in medical imaging due to the limitations of data availability. It is a standard machine learning technique that reuses previously gained knowledge in one domain to improve performance in another. For this purpose, the models pre-trained on ImageNet, one of the largest natural image datasets, are used as the initial weights instead of initializing with random weights and later, models are finetuned with task-specific data. Fine-tuning is the process of modifying the parameters of a pre-trained model on a new, generally smaller dataset in order to transfer learnt characteristics and enhance performance on a specific task related to or expanding beyond the initial training job. The retraining of the models, is the process of training

a pre-existing model on new data, either to update its knowledge or to adjust it for a change in the task, can be done on the last few layers in CNNs or only one last layer, like in the vision transformer or the entire models. Even though the context in the natural images and medical images are entirely different, some basic structures like edges and texture can still be recognized, utilizing models on natural images can benefit tasks rather than random weights.

3 State of the art

In this chapter, a concise description of computational pathology with emphasis on prostate cancer classification and relapse prediction, along with the current state-of-the-art models employed in the pathology domain, is given. In continuation, the recent work on prostate cancer pathology images and other histopathology images is summarized.

3.1 Computational pathology

The digitization of the pathology images of tissues and storing them for further analysis is coined digital pathology (Abels et al., 2019). Pathologists' diagnosis using digital images does not vary a lot compared to diagnosis with physical images viewed under the microscope, as stated in (Azam et al., 2021). Computational analysis of digital images, defined as the use of machine learning and deep learning approaches to automatically analyse and interpret medical data, reduces repetition and time consumption when compared to manual analysis (Li et al., 2021b). Computational pathology is a rapidly expanding discipline as a result of the advancements in deep learning and histopathology slide digitalization, which have enhanced the accessibility of whole slide images. This has been further enhanced due to an increase in computational power in recent years (Hanna et al., 2020).

Several studies using digital images based on machine learning and deep learning have recently demonstrated performance similar to pathologists. Computational pathology has laid paths ranging from classification of cancerous and non-cancerous regions (Xu et al., 2017; Campanella et al., 2019), quantification of nuclei, lymph nodes (Hu et al., 2021), and also survival prediction of the patients (Wulczyn et al., 2020; Fan et al., 2021). Deep learning has exhibited good performance of histopathology images compared to traditional machine learning methods, similar to the natural image domain (Abels et al., 2019). The complexity of histopathology images makes it difficult to obtain manual features. Although there are sophisticated developments in this field, there is a gap between research and the application of the research in hospital settings (Rakha et al., 2021).

3.1.1 Challenges of computational pathology

Though deep learning algorithms have achieved significant performance levels in the classification and segmentation of images, some differences still have to be considered when handling histopathology data compared to natural image data. There are vital challenges in the appearance and sizes of the images because of their high resolution and varied scale, which may require specialised preprocessing steps to accurately identify pathological features among the complex and varied tissue structures present.

One of the apparent distinctions between histopathological images and natural images,

aside from the distinctive colour distribution, which primarily displays purple and pink colour variations with the H&E-staining, is the intended rotation invariance. Independent of the individual task, the cell orientation is unimportant. This is often handled by randomly rotating and flipping images during model training to make the networks robust to these variations.

Another crucial aspect is the size of the WSI. Natural images often have sizes of 256×256 pixels, whereas WSIs range to $100,000 \times 100,000$ pixels. WSIs are usually cropped into smaller patches which can be fed to deep learning models for further feature extraction. The multiple instance learning method has been a go-to method for WSIs for a long time due to their enormous image sizes, which is explained in detail in 3.3.2.

Training of DL models demands vast amounts of data. Unlike natural images from everyday life, which are easily accessible, histopathology images are hard to acquire as there is limited access to publicly available data for research purposes (Deng et al., 2009). Obtaining data from the laboratories is time-consuming and challenging (Abels et al., 2019).

Among the primary reasons for the low accessibility of the pathology images is that they are still routinely examined under a microscope and mostly not digitized. In accordance with research from (Williams et al., 2018), less than 50% of the tissues did not have digital slides. The digital slides are mainly used for academic purposes and not for diagnosis. According to a study from (Nam et al., 2021), less than two-thirds of pathologists in Korea use digitized images. If digitized images are acquired, there are legal concerns, and data privacy issues have to be addressed. As it is patient data, patients need to consent, and the data needs to be anonymized or pseudo-anonymized. For data-driven models, images from several hospitals are required for better generalization. Data acquisition will have high variability as it can be scanned in various scanners, or staining protocols can vary, as this can introduce inconsistencies in data for models and requires robust preprocessing techniques for the models to perform well when trained in this data.

For supervised learning algorithms, annotated data is required, which is time-consuming and costly. Having pixel-level annotations for WSIs is more tedious than other images due to their sizes, and these require high domain expertise (Tizhoosh and Pantanowitz, 2018; Montagnon et al., 2020; Kohli et al., 2017). For prostate cancer, the Gleason score or ISUP score is provided by pathologists. For relapse prediction, follow-up data of the patients is needed from the hospitals. There can be further difficulties in collecting follow-up data as they might not have been correctly stored in the electronic health records. For survival prediction, all patients need to be in the common time span. The time from the biopsy or prostatectomy to the relapse should, therefore, be noted correctly for training the model.

3.2 Deep learning for computational pathology

As discussed earlier, deep learning has shown outstanding results on imaging tasks ranging from everyday images to medical images. It is hard to obtain a balanced dataset for the training of deep neural networks, as in the hospital setting, people with abnormal health

or cancer are significantly less common compared to healthy humans. Apart from dataset imbalance, some challenges in medical images, especially with histopathology WSIs, their enormous sizes and difficulty in pixel-level annotations, will be addressed in this thesis. Multiple instance learning has been one of the major breakthroughs for histopathology images, which is further explained in detail in 3.3.2. Also, Vision Transformers (ViTs), the state-of-the-art models in computer vision, have recently penetrated the medical imaging field, particularly the pathology domain. In this section, MIL and ViTs are explained in detail as they are the main models that are used in this thesis, along with CNN models, as they are chosen as baseline comparison models.

3.2.1 Cancer detection

The process of locating cancerous cells or tumourous within the human body is known as cancer detection (Walhagen et al., 2022). Early diagnosis is essential for enhancing treatment results and raising the likelihood that therapies will be effective. Manual annotations of histopathology slides by pathologists are labour-intensive, time-consuming, subjective, and sensitive to interobserver variability (Singhal et al., 2022). Deep learning models have the ability to examine a vast number of histopathological images, identify patterns, and categorise them as cancerous or non-cancerous, even with weak labels. In this area of research, CNNs are frequently utilised because of their ability to learn and extract pertinent information from the images (Das et al., 2018). Currently, ViTs show outstanding performance in classification when they are pre-trained on large datasets and then fine-tuned on the required task. This current thesis focuses on pre-training models on cancer detection task with one of the largest datasets available on prostate cancer.

3.2.2 Relapse-free survival prediciton

The term survival prediction refers to determining the time-to-event interval between surgery or other treatments and passing away from a specific disease (Cheon et al., 2016), whereas relapse prediction is determining the occurrence of disease again once the patient has been treated to it. Relapse prediction plays a significant role in determining treatment of the diseases. The prediction of survival was ranked high among medical professionals during the survey conducted on AI in digital pathology by (Heinz et al., 2022), indicating its importance in the clinical domain. In this thesis, binary relapse prediction is addressed.

Presently, the life expectancy of prostate cancer patients is indirectly captured by the Gleason score (Epstein et al., 2016). Gleason score is very subjective, as the grading between pathologists varies highly. Contrary to Gleason annotating methods, relapse prediction has an objective endpoint, as it is distinct and observable occurrence after certain period, like 5 years, as annotation.

There are certain challenges in relapse prediction, as the relapse does not only depend on the structures visible in the images. Factors like diet, genetics and family history also play an essential role (Cheng et al., 2022). Further, the probability of relapse over time is not measurable, as the disease discretely occurs at distinct time points. In this thesis, the time point of 60 months is considered for the prediction of relapse (Kumar et al.,

2017; Yamamoto et al., 2019). If the patients have a relapse of cancer during this period, they are considered to be positive patients. The patients who remain healthy without relapse are regarded as negative/healthy. A patient "dropping out" of a study in the context of medical research is someone who stops participating in the study or does not finish it for any number of reasons, including lack of contact, withdrawal of permission, or the onset of a separate medical condition. These dropped out patients are removed from this study as we are making binary relapse predictions.

3.3 Network architectures

In this section, the architectures used in the thesis are explained in detail.

3.3.1 EfficientNet

EfficientNet is a collection of convolutional network models that, compared to other models competing in the ImageNet database, had achieved the most accurate state-of-the-art performance (Tan and Le, 2019). The eight models in the efficient model group range chronologically in precision and number of parameters from B0 to B7. It has gained significance for its impressive performance and, at the same time, maintaining compact model size and reasonable computational requirements.

The key components of the model are basic convolutional blocks, EfficientNets compound scaling, and efficient attention mechanism. The model has a stack of CNN blocks with various kernel sizes that help the model extract the features from the images. Compound scaling is the technique that optimizes the network depth, width, and resolution simultaneously and has been used in EfficientNets for better performance. Also, the model has a squeeze and excitation attention mechanism, which initially captures spatial global information to a single channel and then activating important channels via a gating mechanism, which controls the flow of information passing through the network layers. These blocks are added to CNN blocks to make them more focused on informative features and squelch unimportant ones.

3.3.2 Multiple instance learning

Multiple instance learning is a type of weakly supervised learning technique, which has been extensively used for data where acquiring fully annotated data is expensive, unlike supervised learning, where each instance of training is annotated (Ilse et al., 2018). All the instances, which are training samples are arranged in sets called bags, and a single label is provided for the entire bag. The labels of the individual instances are unknown. Thus, weak labelling, labelling of the entire bag instead of each instance in the bag, is utilized for predictions.

Bag creation

In the context of pathology WSIs, each WSI is supposed to be a bag, and the image patches or regions in the WSI are regarded as instances of the bag. The patches are created using a systematic grid or sliding window technique, making the images easier to handle for analysis.

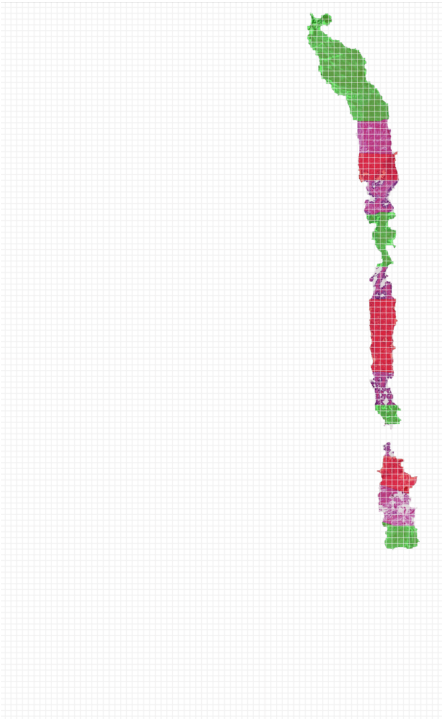


Fig. 3.1: Multiple instance learning image patches. The whole slide image is cut into 256×256 patches, with cancerous (red) and non-cancerous (green) masks overlapped on the tissue.

Instance labeling

One of the ways to define bag-level labels is based on the occurrence or lack of particular structures. As shown in the image 3.1, if one of the patches in the whole bag contains cancerous tissue, then the entire WSI is stated to have a positive label or to be abnormal. When all the patches in the WSI are healthy or non-cancerous regions, the WSI is considered negative or healthy. For tasks like relapse prediction, the bag-level labelling for WSI image is done based on labels available for the patient.

Feature extraction

The image patches contained within each WSI bag are used to extract a variety of characteristics. These characteristics may be colour-based, texture-based, morphological, or more sophisticated deep learning-based features. The objective is to identify distinguishing traits that can identify cancerous and non-cancerous instances. At the end of the model, a classifier is used to distinguish between the two classes. Various types of pooling methods, like max pooling or average pooling, are employed to obtain a single value for the bag of image patches.

3.3.3 Vision transformer

In order to make use of the transformer architecture's capability to preserve long-range relationships inside an image, Dosovitskiy et al. (2020) adapted it to computer vision problems. The key concept is to split the image into patches and consider each patch

as a token, similar to an NLP application. The patches are processed independently, considered as separate entities, and the attention mechanism is used to attend to one another, allowing the patches to model a long range of dependencies by letting each patch get information from other distant patches.

Initially, when the model trained on the mid-sized image classification datasets, it could not deliver better results than CNNs, as transformers struggle to generalize well if trained on limited data. Later, transformers are trained on huge data sets ranging from 14M to 300M images, producing the best results compared to CNNs (Dosovitskiy et al., 2020). Thus, most of the tasks which has less data will utilize pre-trained vision transformer models and then fine-tune on the task at hand for better results.

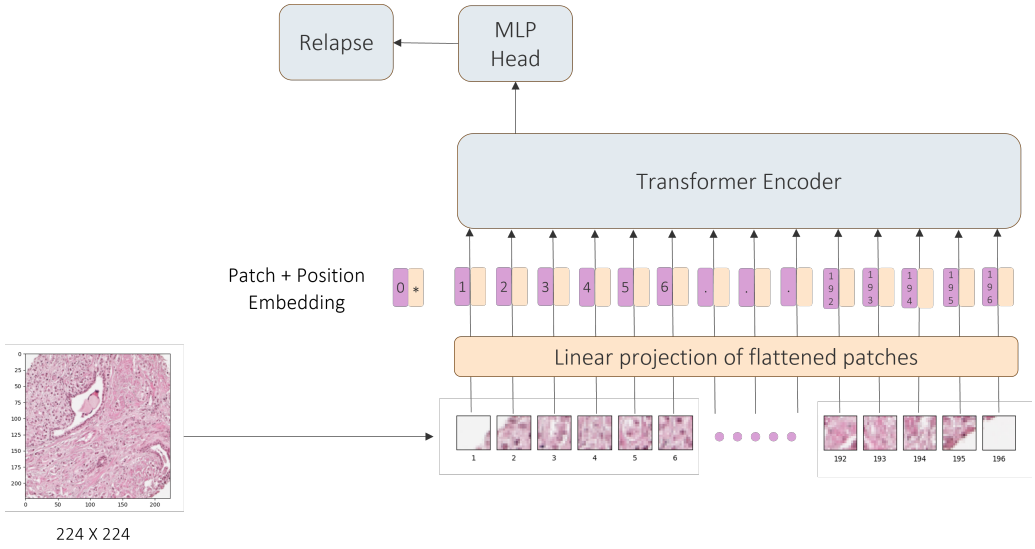


Fig. 3.2: Vision transformer overview. Image of size 224×224 pixels is converted into patches of 16×16 pixels and are linearly projected, then patch and position embeddings are added. A classification token is also added to the sequence and later fed to the encoder. MLP head at the end is used for classification.

The overall process of ViT is illustrated in Figure 2.9. As the first step, the image is split into a 2D sequence of non-overlapping or overlapping patches. These patches are linearly projected into higher dimensional space. Later, the positional encodings, and the learnable classification token at the beginning, are added to the patch projections. These projections and the positional embedding matrix, explained below, are fed into the encoder layers of the transformer, whose output is later given to the MLP head to get the final output.

Patch & positional embedding matrix Unlike a one-dimensional (1D) sequence of tokens in a standard transformer, the images that have to be fed to the vision transformer are two-dimensional with the shape $H \times W \times C$, where H is the height, W is the width of the image with C being the number of channels, often RGB (Red, Green, Blue) colour channels. The images are split into 2D sequences of patches of shape (P, P), where P is

the resolution of patches in pixels, which is usually 16×16 or 32×32. $N = \frac{H \cdot W}{P^2}$ is the number of patches obtained per image. A normalization layer is applied to each patch to get the required dimensionality. Afterwards, a linear transformation is applied where each patch is mapped into higher dimensional space to get interpretable and meaningful features from the image.

A classification token, which is added at the beginning of the embedding matrix, is a critical component in the architecture, as it helps the model to output the classes after training. This is necessary for the model to represent its learnings. A positional encoding matrix is added to retain the spatial information between the various regions of images, with the 0^{th} position for the class token and the rest for patches in the images.

Encoder ViT architecture has only an encoder component from the original transformer, as no output image is generated using a decoder network. Each encoder layer has a linear normalization layer (LN), multi-head attention (MHA) and MLP layers, as shown in Figure 3.3. A residual connection is applied to each layer, which helps in preventing exploding and vanishing gradients. Another advantage of the residual connection is that the network converges quickly, even with many layers. Input x is added to the output of the MHA layer leading to the first residual connection as explained in the equation 3.1 (Dosovitskiy et al., 2020). Similarly, the output of the first connection is added to the output of MLP to form the second residual connection, as expressed as

$$\begin{aligned} \text{First residual connection} &= y_1 = \text{LN}(x + \text{MHA}(\text{LN}(x))), \\ \text{Second residual connection} &= y_2 = \text{LN}(y_1 + \text{MLP}(y_1)). \end{aligned} \tag{3.1}$$

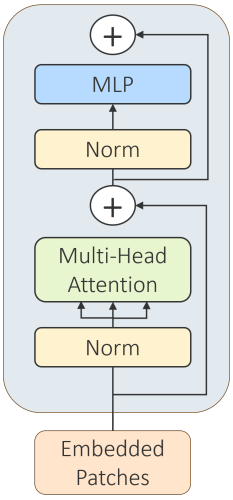


Fig. 3.3: Encoder of vision transformer adapted from (Vaswani et al., 2017).

Normalization layers are usually used to reduce the training time of the models, which several methods can achieve. Here, layer normalization is implemented, in which the mean and variance of the inputs are calculated, and the normalization function is applied to layers on training instances. After normalization, the output is fed into the attention layer, which is explained in detail in the 2.2.5.

The MLP is two-layered feed-forward network, which helps the ViT to capture complex and non-linear features. The layers in the MLP have a GELU activation function with dropout layers in between. GELU, which stands for Gaussian Error Linear Unit, has proven to outperform other activation functions in imaging and speech tasks (Hendrycks and Gimpel, 2016). GELU aims to combine features of activation functions and dropout regularizers. The inputs are multiplied by values from 0 to 1, but the output of this is determined by the input value. GELU is mathematically formulated as a cumulative distribution function (CDF) of the standard Gaussian distribution as expressed in 3.2.

$$\mathrm{GELU}(x) = xP(X \leq x) = x\phi(x) = x\frac{1}{2}[1 + \mathrm{erf}(x/\sqrt{2})] \tag{3.2}$$

$$\mathrm{GELU}(x) \approx 0.5x(1 + \tanh(\sqrt{(}2/\pi)(x + 0.044715x^3))) \tag{3.3}$$

In the end, the MLP head returns only one value at the beginning of the sequence, which determines the final class of the input image. This has one normalization and linear layer.

3.4 Recent work on histopathology images

According to a recent literature analysis, research on machine learning-based prostate cancer categorization has grown (Denysenko et al., 2022). Multiple imaging modalities like magnetic resonance imaging (MRI) and computed tomography (CT) can be used for prostate cancer studies apart from histopathology imaging. As there are few studies conducted with the combination of prostate cancer histopathology images and vision transformers, recent works with vision transformers on histopathology images with other diseases, along with prostate cancer, have been summarized in this section. The key similarities and variations in the methods used on H&E-stained histopathology images will be highlighted in the following paragraphs. A short overview of the papers is shown in Table 3.2.

3.4.1 Related work on the classification of cancer

In this section, related work on the classification of different types of cancer, the various pathologies and the datasets along with metrics the models are evaluated on is summarized.

Pathologies

There has been a growing interest in leveraging WSIs for the analysis and diagnosis of various types of cancers in digital pathology. Campanella et al. (2019); Duran-Lopez et al. (2020); Bhattacharjee et al. (2022); Shao et al. (2021) have used histopathology images of prostate cancer to detect the presence or absence of cancer. For classifying tumours in the brain, Li et al. (2023) have used weakly supervised ViT. Similarly, ViT models, both alone and in combination with CNNs, have also been employed to address other pathologies, such as breast, oral and colorectal cancer, as mentioned in Thomas et al. (2022); Singha Deo et al. (2022); Zeid et al. (2021) respectively.

Publications that studied the performance of prostate cancer with respect to ViT are

very few (Ikromjanov et al., 2022; Chen et al., 2022b). The methodologies are frequently improved to achieve good performance for disease detection. Consequently, despite the fact that a model employs the same image format across all diseases, comparing its performance is challenging. The several reasons hindering fair comparison could be differences in staining of the image, various magnification levels and color scheme of the scanner while digitizing. Various datatypes like biopsies and TMA spots are used for training, making it difficult to compare models. Also, models are performed on various disease types, which differ in cellular structure and their growth of the cells aiding for cancer.

Output objectives of models

Deep learning models provide various outputs, enabling pathologists and clinicians to make informed decisions. At a slide level, models determine the presence or absence of cancer by learning to identify characteristic patterns and features associated with cancerous tissues (Campanella et al., 2019). In addition, Arvaniti et al. (2018); Burlutskiy et al. (2019) performed localization, which involves identifying the precise location of cancerous regions within a pathology slide and segmentation of cancer, which outlines the boundaries of individual cancerous cells or tumour structures. This detailed delineation aids in quantifying the extent and size of the cancerous area to accurately assess the spread and distribution of cancer within the tissue. del Toro et al. (2017), identifies the cancer grading, which refers to the assessment of the aggressiveness level of cancer cells, which in turn helps in treatment planning decisions. Moreover, these models can also classify cancer into subtypes based on distinct molecular markers or histological characteristics.

Datasets

The datasets for research purposes have several sources of origin. Datasets can exhibit variations in terms of geographical regions, hospitals, and the diverse protocols employed by healthcare institutions for data collection. Some (Arvaniti et al., 2018) use TMA spots obtained after RPE for the diagnosis, while others (Li et al., 2021b; Campanella et al., 2019) use the entire WSI. Some slice the images after prostatectomy, and these slides are used for classification.

The size of prostate cancer datasets used in research can vary significantly depending on available resources and data collection efforts. The Prostate cANcer graDe Assessment (PANDA) challenge dataset is a publicly available dataset focused on prostate cancer detection and is used by Bulten et al. (2022) for classification. Hulsen (2019) aimed to classify Gleason grading from The cancer genome atlas (TCGA) prostate adenocarcinoma (PRAD) dataset with 16,790 images, is part of the larger TCGA project. Arvaniti et al. (2018) use TMA spots from the tissue micro arrays Zürich (TMAZ) dataset of about 864 images with segmentation annotations for each Gleason pattern. Some studies have access to large-scale publicly available datasets comprising 12,160 WSI, allowing for robust statistical analysis and generalizability of findings Campanella et al. (2019). Other studies may have access to smaller in-house datasets (Zhang et al., 2021), which confine the output. The lack of standardized datasets and variations in the sizes of datasets used in different publications have led to challenges in achieving comparability across studies.

Pixel-wise data annotation for WSI is highly tedious as each image consists of thousands of pixels (Arvaniti et al., 2018). For most of the datasets, for WSI, slide-level labelling is available, indicating the existence of cancer or the grade of cancer (Campanella et al., 2019). As there is high inter-observer variability for pathologists grading, some studies aim to obtain annotations from more than one pathologist and select the majority grading as the final ground truth (Nagpal et al., 2019; Bulten et al., 2020). Apart from direct annotations by specialists, certain datasets have annotations obtained from clinical records. Instead of getting fully annotated by pathologists, a portion of the dataset is automatically segmented using additional histochemical stains (Burlutskiy et al., 2019), which can be used to provide colour information that identifies the absence of basal cells, which in turn indicates the presence of cancer.

When working with datasets, researchers often perform dataset splits to separate the data into different subsets for training, validation, and testing purposes. The most common approach is a random split, where the data is randomly divided into three parts, typically with a ratio of 70% for training, 15% for validation, and 15% for testing (Campanella et al., 2019; del Toro et al., 2017). Other publishers Bulten et al. (2020) utilize different image sources for training and testing. Some utilize different TMA arrays for validating and testing, and the rest for training the models.

Dataset preprocessing

Various approaches are employed in different papers to prepare the data for model training. Sikaroudi et al. (2023) used pre-trained weights from ImageNet, a large-scale dataset used for training visual recognition models, and implemented normalization with a mean and standard deviation of the ImageNet dataset. This normalization enables better convergence during model training. Burlutskiy et al. (2019) use histogram normalization techniques, such as Macenko normalization, to enhance the contrast and brightness of histopathology images, improving the visual quality and interpretability of the data. To increase the generalizability of the models, Chen et al. (2022b) employed data augmentation techniques. These techniques involve applying random transformations to the data, such as flipping images horizontally or vertically, performing random rotations, and applying random zooms.

Architectures

Springenberg et al. (2022); He et al. (2022) provide a thorough evaluation of deep learning models, from CNNs to vision transformers, for the analysis of histopathology images. CNNs have been in the limelight for diagnosing H&E images for a long time. There is no one specific CNN architecture that performs the best. (Nagpal et al., 2019; Ström et al., 2020), report InceptionV3 have better performance than ResNet (Campanella et al., 2019; Oner et al., 2022). Most authors use pre-trained networks on ImageNet, while some train the model from scratch (Sikaroudi et al., 2023). Computational pathology is witnessing extensive implementation of ViTs due to its recent advancements. Ikromjanov et al. (2022) implemented a ViT for detecting cancer on the PANDA challenge dataset. Ikromjanov et al. (2022) subdivided WSI into patches of smaller size, i.e., 256×256 pixels, and they are flattened and fed to the transformer encoder like the basic transformer model. In the end, the model classified the different Gleason grades and whether the

patch contained a stroma or was benign. Zeid et al. (2021) compares two approaches, a ViT and a convolutional compact transformer (CCT), for analyzing pathology images related to colorectal cancer. ViT employs eight layers of transformer blocks to process image patches, while CCT combines convolutional layers and four transformer encoder layers with sequential pooling. CNNs excel at capturing local spatial features, while vision transformers leverage self-attention mechanisms to capture global dependencies within the images, which can lead to improved performance when combining these approaches. In Thomas et al. (2022), the authors used a similar architecture like Ikromjanov et al. (2022) for breast cancer classification. The ViT model employed is pre-trained on the ImageNet dataset. Another recent approach, Chen et al. (2022b), has implemented a so-called hierarchical ViT, which leveraged the different scales of magnification of histopathology images. The method involves pre-training of ViTs in stages using self-distillation with no labels (DINO) on TCGA data. The embeddings from the low-level ViT are given as input to higher-level ViTs. The ViTs trained at different levels take input image sizes of 256×256 pixels, 4096×4096 pixels, and the entire WSI, thus making the model a three-stage ViT. Another prominent approach for WSI analysis is MIL. Most studies on MIL with CNN often overlook the issue of correlation, assuming independent and identical distribution. However, Shao et al. (2021) attempted to address this limitation by incorporating correlation into the model. The authors combined MIL and a ViT.

Metrics

Several measures have been used to assess the effectiveness of classification model performance in recent work on the classification of prostate cancer. An often-used metric for classification is accuracy, which assesses the overall correctness of predictions (Nagpal et al., 2019; Ren et al., 2018; Bhattacharjee et al., 2022). Furthermore, measures like area under the receiver operating characteristic curve (AUROC) have been used to gauge the accuracy of model predictions at various classification levels (Bhattacharjee et al., 2022; Duran-Lopez et al., 2020). This is crucial when working with datasets that are unbalanced since the AUC offers an overall performance indicator. Additionally, the F1-score, which combines accuracy and recall, which is the rate of true positives overall positive instances into a single metric, has been used to assess the balance between accurately detecting positive cases and preventing false positives.

3.4.2 Related work on binary survival analysis

In this section, the related work on survival analysis of various pathologies is summarized since strategies can be mutually exchangeable for different diseases. Apart from histopathology images, other medical images like CT and MRI are also included in this overview as similar models can be applied. The endpoints referring to survival for the studies and the datasets used and metrics analyzed are different across the studies not allowing for a fair comparison between the models. There are several categories for this specific problem statement, including binary survival prediction, risk score prediction and survival curves which give predictions of relapse. As this study is mainly focused on binary survival, only papers related to classification are concisely presented here.

Kumar et al. (2017) used CNNs to predict the biochemical recurrence of PCa within a

five-year period on TMA spots. Huang et al. (2022) have proposed a prediction for a three-year window using WSIs of biopsies using CNNs. Yamamoto et al. (2019) evaluated the CNN model on WSIs for the 1-year and 5-year relapse prediction. The metrics used for this approach are mainly AUROC, accuracy, and specificity, the rate of true negatives overall negative instances. In Chen et al. (2022b), hierarchical ViT is used for survival prediction by obtaining information from different scale levels along with the genetic features and combining it with MIL for survival prediction. The five largest datasets from TCGA are used for this study. Another approach by Chen et al. (2021) has combined radiology and histology data to obtain survival prediction. In this approach, the combination of CNN and hierarchical ViT is implemented on the WSIs dataset from TCGA and radiology images from the cancer imaging archive (TCIA). As ViTs require large data and computational resources, most models Sikaroudi et al. (2023) used pre-trained models on ImageNet and only fine-tuned on the necessary task.

Summary

The cancer detection and relapse prediction approaches mentioned are heterogeneous, each study uses varied image datasets, different image types and sizes of datasets. There is no one baseline model that can be used to compare the performances of the models because of the variations. Also, there are only a few papers related to prostate cancer survival prediction using vision transformers, making it more difficult to draw comparisons. In this work, a systematic approach to draw the comparison between CNN, ImageNet pre-trained, domain-related pre-trained transformers and hierarchical transformers is implemented, as all the models are implemented on the UKE dataset.

Tab. 3.1: Research papers on different histopathology tasks using deep learning. CNN: Convolutional neural networks, MIL+ResNet: Multiple instance learning combined with ResNet model, ViT: Vision transformers, ViT+CNN: ViT combined with CNN, ViT+MIL: ViT combined with MIL, MIL + attention: MIL based on attention, HViT: Hierarchical ViT

Paper	**Task**	**Architecture**	**Pathology**
(Ren et al., 2018)	low / high	CNN	prostate cancer
(Arvaniti et al., 2018)	Gleason grade classification	CNN	prostate cancer
(Campanella et al., 2019)	benign / malignant	MIL+ResNet	prostate cancer
(Nagpal et al., 2019)	Gleason grade classification	CNN	prostate cancer
(Burlutskiy et al., 2019)	binary segmentation	CNN	prostate cancer
(Bulten et al., 2020)	Gleason grade classification	CNN	prostate cancer
(Duran-Lopez et al., 2020)	benign / malignant	CNN	prostate cancer
(Ström et al., 2020)	Gleason grade classification	CNN	prostate cancer
(Zeid et al., 2021)	multi-class classification	ViT, ViT+CNN	colorectal cancer
(Shao et al., 2021)	Binary and cancer subtyping	ViT+MIL	various cancer types
(Zhang et al., 2021)	multi-class classification	MIL+attention	various cancer types
(Ikromjanov et al., 2022)	Gleason grade classification	ViT	prostate cancer
(Bhattacharjee et al., 2022)	benign / malignant	CNN	prostate cancer
(Thomas et al., 2022)	benign / malignant	ViT	breast cancer
(Singha Deo et al., 2022)	benign / malignant	ViT, CNN	oral cancer
(Chen et al., 2022b)	benign / malignant	HViT	various cancer types
(Li et al., 2023)	tumor subtyping	ViT	brain
(Sikaroudi et al., 2023)	tumor classification	CNN	various cancer types

Tab. 3.2: Research papers on survival prediction tasks using deep learning.CNN: Convolutional neural networks, ViT: Vision transformers, HViT: Hierarchical ViT

Paper	**Task**	**Architecture**	**Image type**
(Kumar et al., 2017)	relapse (5 years)	CNN	histopathology
(Yamamoto et al., 2019)	relapse (1 and 5 years)	CNN	histopathology
(Huang et al., 2022)	relapse (3 years)	CNN	histopathology
(Chen et al., 2021)	relapse	ViT	histopathology and radiology
(Chen et al., 2022b)	relapse	HViT	histopathology

4 Materials and Methods

In the following section, an overview of the datasets and the methods implemented in the thesis is provided. Firstly, datasets for cancer detection and survival prediction are explained, and the related statistics are provided. Later, the preparatory steps on data for training for evaluation and the approaches used are described in detail. At the end of the chapter, training and evaluation processes, along with the evaluation metrics, are furnished.

4.1 Datasets

Relapse prediction of prostate cancer is implemented on TMA spots acquired after RPE using ViTs in this thesis. ViTs are usually pre-trained with huge amounts of image data for better performance. As the relapse prediction related data is small, pre-training the models with biopsies which are similar to TMA spots will help the models for improved performance as both have similar cell structures. Biopsies are used to pre-train the models for cancer detection task. The biopsies do not have follow-up data, thus making relapse prediction not possible. In this thesis, two different datasets are used, the publicly available PANDA dataset and an internal dataset from the University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany. The two datasets are described below.

PANDA dataset

PANDA challenge dataset is one of the largest publicly available WSI datasets for prostate cancer (Bulten et al., 2020). It consists of 10,616 WSIs of prostate cancer collected from two centres, Radboud University Medical Center in the Netherlands and Karolinska Institute in Sweden. All the images have been recorded at 20x magnification. Among all WSIs, $10,515$ have been manually annotated by pathologists. Each pixel in the WSI is annotated, with each pixel representing a class. For the Swedish dataset, three categories are defined i.e., background, stroma and cancerous tissue, whereas for the Dutch dataset, six classes are defined, background, stroma, healthy, and Gleason grades 3, 4, 5. The annotations of the Karolinska Institute dataset are marked by one urologist with a clinical procedure. For the Radboud dataset, three uropathologists from two different institutes have provided annotations. The images vary in size with $5,000$ to $40,000$ pixels per dimension with approximately $0.5\mu m$. Figure 4.1 shows an example image of the PANDA dataset.

UKE dataset

The survival dataset for relapse prediction is the internal dataset provided by the UKE Institute of Pathology. Unlike the PANDA data, UKE data are tissue microarray (TMA) spots obtained after RPE as shown in 4.2. TMA spots are stained with hematoxylin for four minutes and eosin for 1:20 minutes, with each spot having a diameter of $0.6mm$ and

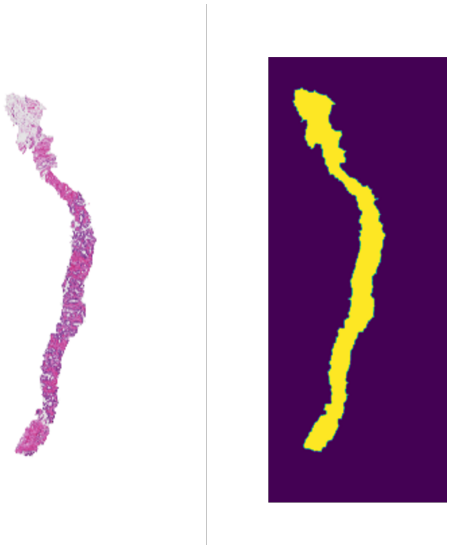


Fig. 4.1: Exemplary image from the PANDA dataset.

2.5μm thickness. The datasets are scanned with a Leica Aperio AT2 scanner, which is used to digitize images with 20x magnification, and these images are digitally magnified to 40x magnification. All the TMA spots used in the thesis are preprocessed to obtain 2048×2048 pixels image size. The recurrence-free survival time for each patient is provided for each TMA spot image in the survival dataset. Each spot corresponds to one patient in the dataset. The time is determined by BCR after RPE, which is obtained from the corresponding electronic health records. Patients whose relapse-free time is not available are removed from the data cohort. Also, the images with sparse or no tissue or overlapping tissues are eliminated from the data by manual control, leading to 14,479 images. Pathologists also provided, ISUP score for the patients related to TMA spots, which is displayed in Figure 4.3. Figure 4.3 represents the train, validation and test set which has different ISUP scores which are used for analysis.

4.1.1 Data splitting

For pre-training of the models, the PANDA dataset is divided into training, validation and test datasets in the ratio of 8:1:1. For fine-tuning on the relapse prediction task, the UKE dataset is also partitioned in a similar proportion of PANDA. The composition of the datasets is shown in Figure 4.4.

4.2 Approaches and architectures

The primary task of this thesis is the prediction of relapse of prostate cancer within sixty months (5 years) after prostatectomy/RPE using TMA spots as input data employing vision transformers. ViTs require massive amounts of data to perform better than CNNs, but with the UKE dataset, this is challenging as the dataset is not very large and has 6828 images. To overcome this limitation, pre-trained ViT networks are utilized. The research questions 1.2 are answered through the strategies and experiments discussed in the following section. All the experiments conducted are displayed in the Tables **??**, 4.1.

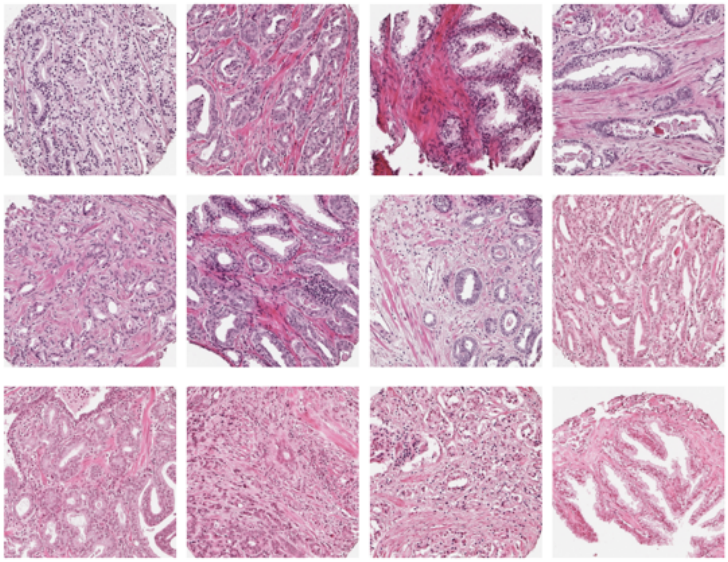


Fig. 4.2: Exemplary images from UKE dataset.

Tab. 4.1: List of experiments. CNN+MIL represents convolutional neural network combined with multiple instance learning, ViT+MIL represents vision transformers combined with multiple instance learning. Models are pre-trained on ImageNet and PANDA datasets. Max and mean aggregate layers are used for MIL.

Model	**Pre-training**	**Pooling**	**Name**
CNN+MIL	ImageNet	Mean	CNN+MIL: ImageNet-Mean
		Max	CNN+MIL: ImageNet-Max
	PANDA	Mean	CNN+MIL: PANDA-Mean
		Max	CNN+MIL: PANDA-Max
ViT+MIL	ImageNet	Mean	ViT+MIL: ImageNet-Mean
		Max	ViT+MIL: ImageNet-Max
	PANDA	Mean	ViT+MIL: PANDA-Mean
		Max	ViT+MIL: PANDA-Max
Hierarchical ViT	PANDA	-	HViT

4.2.1 Vision transformers with MIL

In this approach, vision transformers, which are state-of-the-art models for computer vision tasks, are integrated with MIL, a technique which is widely implemented for WSIs, because of their enormous sizes and their lack of annotations at an individual pixel level.

This approach has two stages, pre-training for detecting cancer from WSIs and fine-tuning these models on the UKE dataset for relapse prediction, which are explained in detail below.

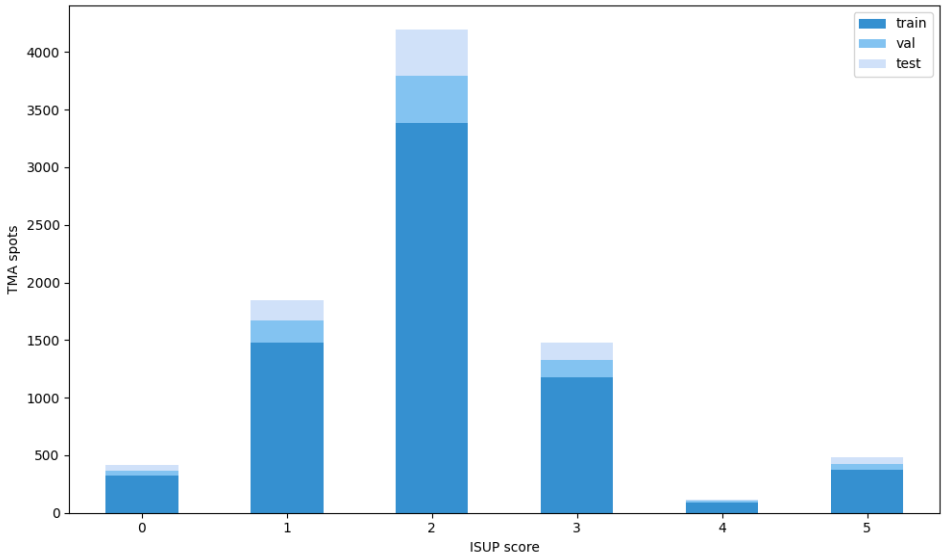
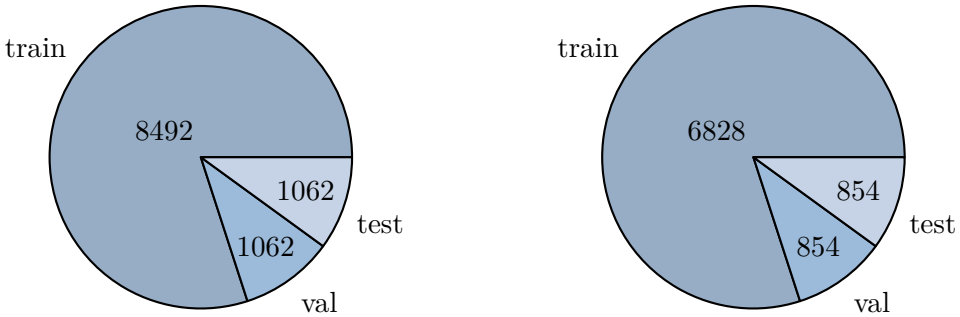


Fig. 4.3: ISUP scores of UKE dataset. Train, val, test indicates training, validation and test datasets.



(a) PANDA datasplit (b) UKE datasplit

Fig. 4.4: Dataset splits.

Pre-training

Pre-training the model entails learning generic characteristics and patterns on the extensive PANDA dataset before fine-tuning it. The ViT-b16, a variant of ViT architecture available at torch-vision models, is pre-trained on ImageNet. Instead of training the ViT from scratch by random initialization of weights, pre-training with WSIs in this task, is done by selecting weights from ImageNet trained model. As standard models trained on ImageNet consist of image input size 256×256, the histopathology images used for pre-training should also maintain the same size. From the enormous WSIs, patches of the required size are extracted, which is explained in detail here.

Patch extraction For the training of ViT, the PANDA dataset with 8,492 number of images is used which lead to 55,42,663 patches. Each image in PANDA has a pixel-level label indicating cancerous and non-cancerous regions. As shown in Figure 4.6, each WSI can thus be separated into background and foreground. In the foreground, cancerous

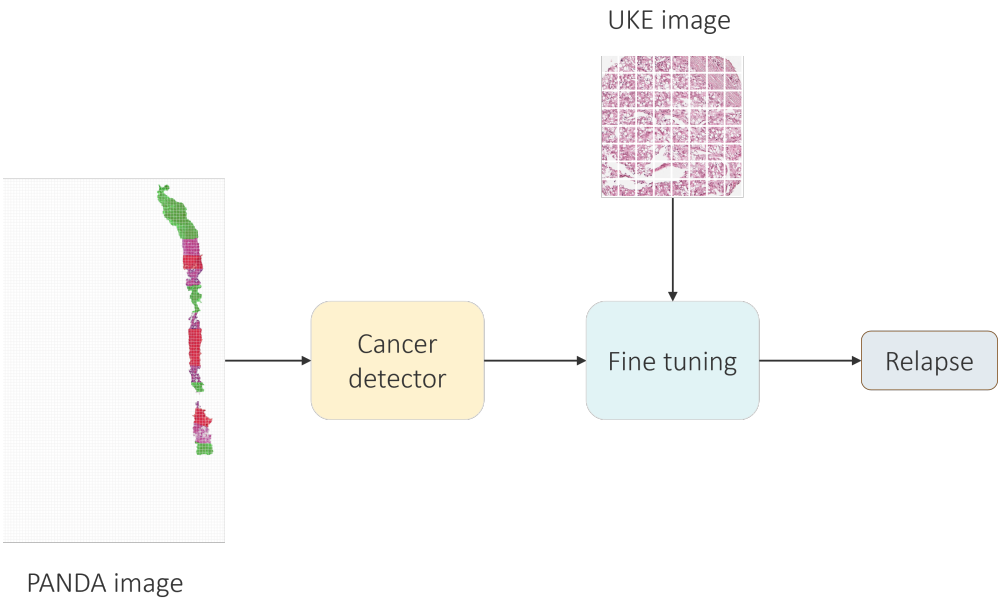


Fig. 4.5: Diagram illustrating the implemented pipeline. The models are first pre-trained with patches of PANDA dataset for cancer detection and later trained weights are used for fine-tuning on relapse prediction of UKE images.

Tab. 4.2: List of pre-trainings. CNN: Convolutional neural network, ViT: vision transformer

Model	**Pre-training**	**Image size**	**Patch size**
CNN	PANDA	256×256	-
$\text{ViT}_{256} - 16$	PANDA	256×256	16×16
$\text{ViT}_{1024} - 256$	PANDA	1024×1024	256×256

and non-cancerous regions are present. The patches of size 256×256 pixels are extracted from every WSI based on the mask threshold. Patches with the foreground with equal or more than the 0.1 threshold value are selected. The patches are labelled positive and negative samples depending on the cancer label threshold value of 0.9. If the tissue region in the patch contains more than 0.9 percent cancer, then it is labelled as the cancerous or positive sample, else negative. The patches are displayed in the Table 4.4.

Implementation Each patch of size 256×256 pixels is first resized to 224×224 pixels in order to comply with the dimensions used in ImageNet models, then images are normalized with mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$ with similar values are ImageNet. Then each image is divided into 196 patches with size 16×16. These are then flattened, and class tokens and positional embeddings are added. Later these are sent to the encoder, which has 12 layers. The embeddings of size 768 are maintained constant throughout the entire model to access skip connections. The input size of MLP is 3072. In the end, the model outputs only one prediction, which detects the presence of cancer in the patch.

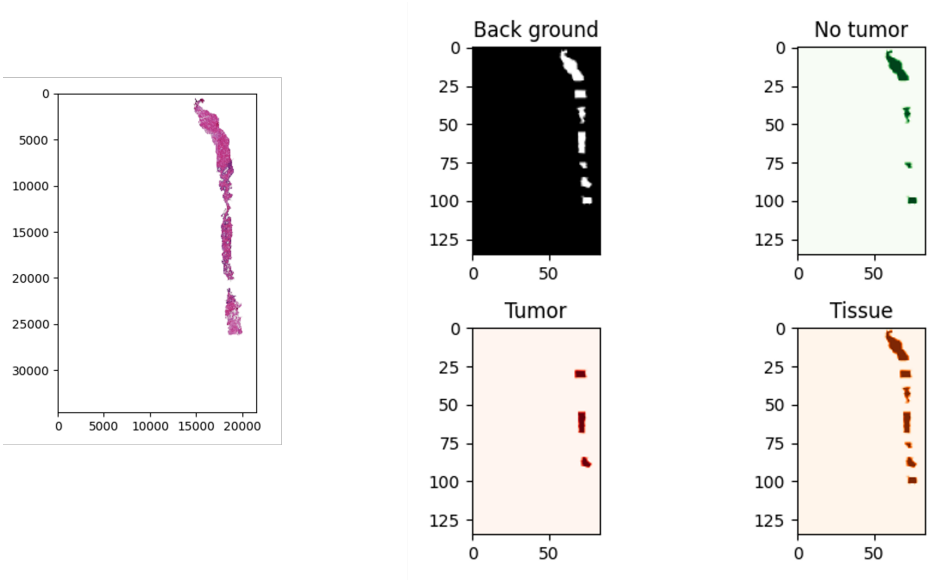


Fig. 4.6: PANDA image with background, no tumor, tumor, whole tissue masks.

Fine-tuning

The model is fine-tuned to a specific task using a smaller UKE dataset with task-specific labels after being pre-trained on a larger PANDA dataset. UKE dataset with each TMA spot of size 2048×2048 pixels is used as the input image for relapse prediction. Each image is divided into a bag of instances with each instance of size 512×512 pixels. Here 512×512 pixels is considered instead of 256×256 pixels, depending on the magnification of the PANDA images. PANDA image with 256×256 pixels and UKE image with 512×512 pixels have $0.5\mu m$ per pixel. Each image in the bag is sent to the pre-trained model. Following the literature on ViTs (Dosovitskiy et al., 2020), only the last layer in the architecture is trained and all other layers are frozen. These are then combined using mean and max aggregation layers to obtain a single label for each bag.

Tab. 4.3: Table showing number of trainable parameters. CNN: Convolutional neural network, ViT: vision transformer

Model	Trainable parameters
CNN	20M
$\mathbf{ViT}_{256} - 16$	330M
$\mathbf{ViT}_{1024} - 256$	236M

4.2.2 Hierarchical vision transformers

In this approach, hierarchical vision transformers are implemented, which involve various levels of hierarchy to capture information at different image resolutions. This hierarchy gives the model the ability to identify the precise features of tissue patterns and structures, as it can view image at different levels, which are essential for correctly detecting cancer

and predicting relapses. Similar to the above approach, this has two steps: pre-training and finetuning of the model.

Pre-training

Based on the image sizes of UKE dataset, a two-stage hierarchical vision transformer is employed. The pre-training is performed on two resolutions as shown in Figure 4.7. One is $\mathrm{ViT}_{256} - 16$, and the other is $\mathrm{ViT}_{2048} - 256$. In $\mathrm{ViT}_{256} - 16$, the suffix represents the input image sizes of 256×256 pixels, and 16 represents the patch sizes used in the transformers. Similarly, 20482048 pixels represents the input image size for the second stage, and 256×256 denotes patch sizes. To maintain pixels per resolution for both pretraining and fine-tuning, the images are resized to 1024×1024 pixels. As ImageNet weights are only available for certain image sizes of 256×256, in this section, it was not possible to use ImageNet weights as starting weights, instead, weights are randomly initialized for $\mathrm{ViT}_{1024} - 256$.

Patch extraction For stage-1 and stage-2 training of ViT, the same PANDA dataset as approach 1 is used, except different patch sizes are extracted from the images and are then fed to ViTs. Stage one is the same pre-training with 256×256 pixels image patches as approach one, ViT with MIL. For stage two training, patches of size 1024×1024 pixels are extracted from the PANDA dataset, and the same threshold as in the first approach, ViT with MIL, is applied to get patches. A mask threshold of at least value 0.9 to label the patch as cancerous and the rest as non-cancerous patches.

Implementation Two separate pre-training runs for cancer detection are implemented. For the first stage, pre-training is done precisely as 4.7. Each 1024×1024 pixels patch is fed to the ViT for stage two. Each image is then sub-patched into patches of size 256×256 pixels leading to 16 sub-patches per image. All the images are normalized with mean and standard deviation as ImageNet. Then the images pass through the layers of the encoder, and finally, outputs are predicted.

Fine-tuning

In order to predict relapse from the UKE dataset, the pre-trained models are arranged in a hierarchical manner. The images in the UKE dataset, which are of size 2048×2048 pixels, are first resized into 1024×1024 pixels to maintain the same pixel-level resolution as the pre-trained models. Now each image is first patched into 256×256 pixels patches. These patches are fed to ViT at stage one. The output embeddings of the last layer, the layer before the final classification layer, are given as input to the second stage ViT. Now these embeddings are fed to the pre-trained $\mathrm{ViT}_{1024} - 256$, and only the last classification layer is finetuned to get the class prediction, which gives probability with which patch has relapse.

4.2.3 Baseline model

As mentioned above, every approach implemented in this thesis has two specific tasks, one pre-training task and one final task. Even for the baseline model for fairer comparison, the model is pre-trained on the PANDA dataset with the same patch selection as ViT

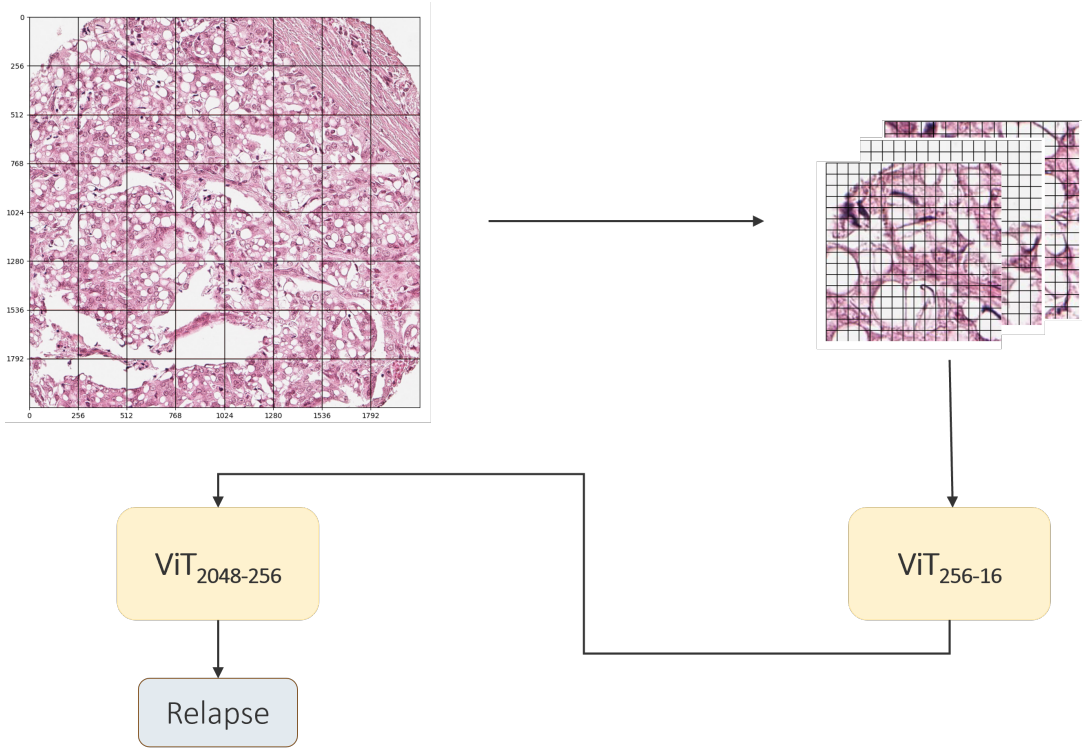


Fig. 4.7: Hierarchical vision transformer. Image of size 2048×2048 is made into patches of size 256×256 and are fed to the $\text{ViT}_{256} - 16$. The output embeddings of this are fed to another $\text{ViT}_{1048} - 256$, and relapse is predicted.

with MIL. EfficientNetB0, a CNN based architecture, has been used in combination with MIL as baseline model. Once the model is pre-trained on PANDA, fine-tuning is done on the UKE dataset for relapse prediction.

Tab. 4.4: Table showing the number of patches for PANDA pre-training. -ve patches indicate no cancer patches, +ve patches indicate patches with cancer.

Patch size	**No: of patches**	**-ve patches**	**+ve patches**
256	55,42,663	41,58,915	13,83,749
1024	5,00,100	3,96,884	1,03,216

4.3 Training

Generally, the training parameters are set at the same level throughout all experiments and models to avoid disproportionate tuning of specific techniques. Two distinct sets of parameters are chosen, one for pre-training with the PANDA dataset for cancer detection and the other set for fine-tuning UKE data for relapse prediction. For cancer detection, batch size 256 is set with a $3e^{-6}$ learning rate with an adam optimizer (Kingma and Ba, 2014). Early stopping is used based on the validation AUC curves. For the approach with MIL, a batch size of 8 is used and for hierarchical ViT, a single image is used per

batch, keeping all other parameters constant. All the experiments were conducted on 48gb NVIDIA RTX A6000 with Kubernetes cluster.

4.3.1 Hyperparameter tuning

The training parameters of the ViTs for pre-training on the PANDA dataset are selected based on the hyperparameter tuning of weight decay and learning rate. Experiments are conducted on a subset of the PANDA patch datasets which constitute 250,000 patches for training and 25,000 patches for each validation and test sets, maintaining the same distribution of labels as the whole dataset. The metric used for selection is the validation accuracy and shown in Figure 4.8, learning rate of $3e^{-6}$ and weight decay of 0.03 are selected.

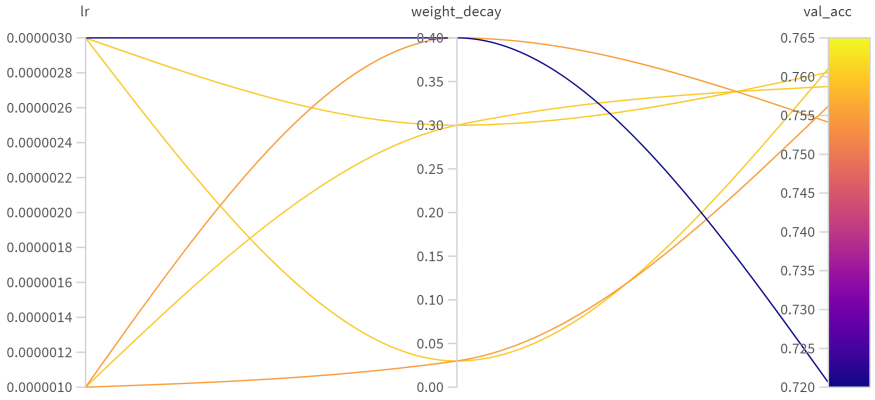


Fig. 4.8: Hyperparameter tuning on the subset of PANDA dataset for learning rate and weight decay. lr: learning rate, val_acc: validation accuracy.

4.4 Evaluation

The models are evaluated for different tasks: the cancer detection model predicts the presence or absence of cancer, and the survival prediction predicts relapse. Within the evaluation pipeline for the cancer detector, the dataset undergoes normalization as the first step. Then, images are resized to patch size 256×256 pixels and fed to the model. Binary cross entropy loss is used as a loss function to evaluate the models. The model outputs the predictions now fed to the sigmoid function to predict the probability between 0 and 1, which is later used to predict the class. These predictions and the ground truths are used to calculate a confusion matrix and other metrics explained below. For the relapse prediction, the evaluation pipeline has the same steps as above.

4.4.1 Metrics

Metrics are utilised to evaluate the model performance and help in choosing the best-performing model. There are a varied number of metrics which help in capturing various characteristics of the model. For binary classification, metrics like accuracy, precision, recall, F1-score, confusion matrix and area under the receiver operating curve (AUROC)

are generally used. For imbalanced datasets, AUROC and F1-score are majorly used to evaluate the performance. Compared to conventional metrics like accuracy, these offer a more fair assessment of the model's performance on both the minority and majority classes.

Confusion matrix

The values in the confusion matrix are the basis for many of the metrics used to evaluate how well a model performs when used for classification. Confusion matrix values are calculated based on the actual and predicted values. Table 4.5 represents the confusion matrix with two classes. True positive (TP) indicates the number of positive examples predicted as positives, and true negative indicates samples which are negative are classified as negative. False positives (FP) are obtained when ground truths are negative, but the model classifies them as positive. Similarly, false negatives (FN) are when the model predicts positive samples as negatives.

Tab. 4.5: Confusion matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

In the context of prostate cancer detection, positive samples are when the patch has cancer, and negative samples imply healthy patients. Similarly, for relapse prediction, if the patient has a relapse within a specific period, the sample is considered to be positive. If not, then it is considered negative. Clinically, the more the true positives are identified, the model is better. If the model has high false negatives, the model is unable to predict the positive cases, which might lead to patient's undertreatment and with high false positives, the patients might get overdiagnosis.

Accuracy

The ratio between accurate predictions and the overall predictions is called accuracy. The formula for accuracy, which indicates the overall correctness of a model is given as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{4.1}$$

Precision

The proportion of true positive predictions to the total number of positive samples is coined as precision. It shows the percentage of positive incidents that really occur when patients have cancer with the formula,

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{4.2}$$

Recall

The term recall is defined as the proportion of true positives to actual positive samples. This is one of the primary metrics for an imbalanced dataset. Recall is especially useful when the cost of false negatives is high, as is the case of the medical domain. It is formulated as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4.3}$$

F1-score

The harmonic mean between recall and precision is presented as F1-score, which combines two metrics into one score. The F1-score provides a mean of how well the model predicts positive across all the positive labels and how many of the positive predictions were incorrect. This metric is mainly considered when models are trained on an imbalanced dataset, and it also serves as a trade-off between recall and precision. It is formulated as,

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4.4}$$

AUROC

The receiver operating curve gives a graphical representation of binary classification problems. It is a popular metric used for classification, as it distinguishes the classes. This metric is a trade-off between true positive rate (TPR) and false positive rate (FPR). TPR and FPR formulae are given as,

$$\begin{aligned} \text{TPR (Sensitivity)} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{FPR (1 - Specificity)} &= \frac{\text{FP}}{\text{FP} + \text{TN}}. \end{aligned} \tag{4.5}$$

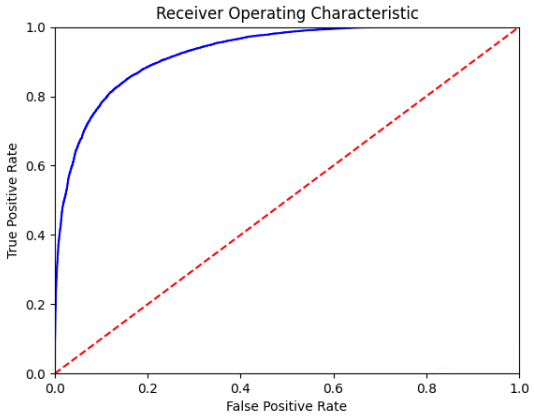


Fig. 4.9: Receiver operating curve. The blue line represents ROC, and the red dotted line represents the line of no discrimination or random choice.

The graph is drawn between FPR and TPR for several decision thresholds, demonstrating how well a model performs for all possible thresholds, not just one. with values on

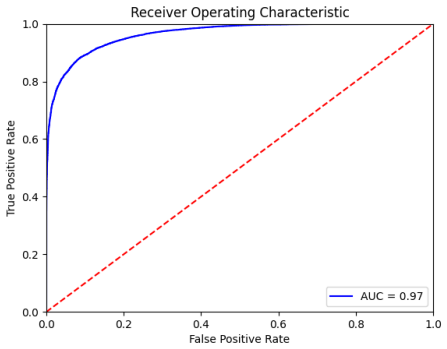
the x-axis and y-axis, respectively as shown in Figure (4.9). Area under the ROC curve (AUROC), assesses the model's overall capacity for discriminating between the positive class and the negative class across all threshold levels. If the value of the AUROC is 1, the model perfectly distinguishes positive and negative samples, whereas if the AUROC is 0.5, the model performs as a random classifier. This is a significant metric when one class outweighs the other class, as it focuses on distinguishing between classes rather than just reflecting majority class predictions. TPR, FPR mentioned above are operated on a fixed threshold basis, whereas thresholds are not considered in calculating AUROC. Rather than assessing performance at a single, set threshold, the AUROC metric gives a summary evaluation of model performance over all possible thresholds.

5 Results

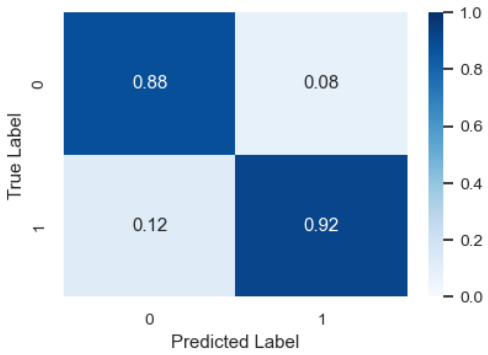
In the results section, all the models are systematically studied, and their performance is compared. First, the cancer detection results are shown, followed by fine-tuning results of CNN with MIL, the baseline model and ViT with MIL, hierarchical ViT architectures used for relapse prediction.

5.1 Pre-training on cancer detection

In this section, the results of the binary classification of cancer on the test dataset, which is used as the pre-training task, are displayed. ROC in Figure 5.1 shows the ViT model with an input image size of 256×256 pixels has achieved 0.97, indicating that the vision transformer model is very effective at distinguishing between the positive and negative classes. High true positives and true negatives in the confusion matrix also demonstrate the model is good at classification of cancer.



(a) ROC on cancer detection with $\text{ViT}_{256} - 16$

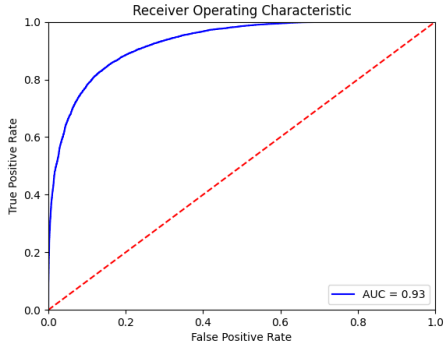


(b) Confusion matrix with $\text{ViT}_{256} - 16$

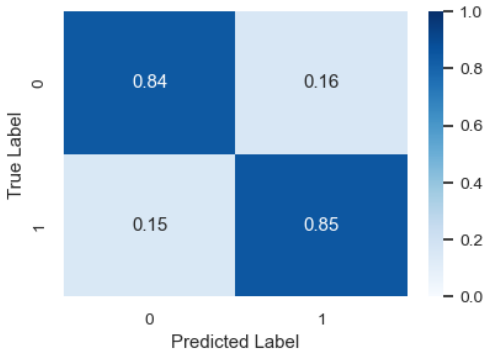
Fig. 5.1: Cancer detection with patch size 256×256 pixels using vision transformer.

For the same classification task, the ViT model has decreased performance when the image size of 1024×1024 pixels is considered. Both true positives and negatives are approximately classified in an equal proportion as depicted in Figure 5.2.

The CNN model, as shown in Figure 5.3, which is selected as the baseline model, achieves an AUROC of 0.88 with a highly predicted negative class in the data with 95 per cent. However, approximately 65 per cent of samples from the positive class are predicted correctly, where as in the ViT, the prediction of a positive class is almost 90 per cent.

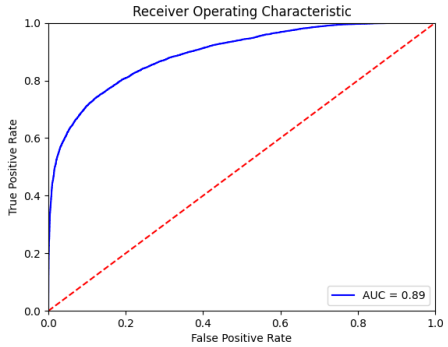


(a) ROC on cancer detection with $\text{ViT}_{1024} - 256$

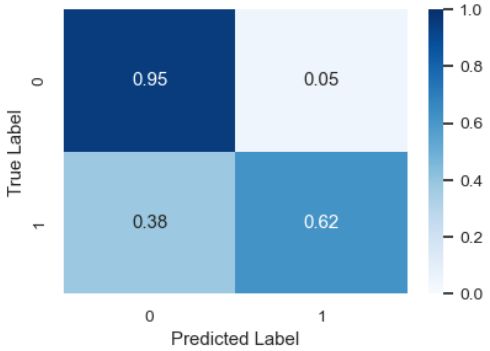


(b) Confusion matrix with $\text{ViT}_{1024} - 256$

Fig. 5.2: Cancer detection with patch size 1024×1024 pixels using vision transformer.



(a) ROC on cancer detection with CNN



(b) Confusion matrix with CNN

Fig. 5.3: Cancer detection with CNN with patch size 256×256 pixels.

5.2 Relapse prediction

In this section, the results of fine-tuning of the vision transformer with already pre-trained weights on ImageNet and PANDA are presented. Both validation and test results of the CNN baseline model, along with two approaches implemented in this thesis, are displayed.

Vision transformers with MIL

The Figure shows 5.4 is the training accuracy curve that shows differences in performance on ImageNet pre-trained models and domain-specific pre-trained models are almost negligible. A dip in the curves of both pre-training datasets with max pooling is observed.

As shown in Table 5.1, vision transformers combined with MIL outperform CNNs in both mean pooling and max pooling with both ImageNet and PANDA pre-trained weights. However, pre-training the models with PANDA, i.e., domain-specific data, did not provide any huge improvement in the performance. When the aggregate layer in multiple instance learning has mean pooling when combined with ViT when pre-trained

Tab. 5.1: Comparision of different methods for relapse prediction with multiple instance learning using CNN and vision transformers. The reported values mentioned here are the mean of three seeds for each experiment on validation (val) and test set. AUROC is abbreviated as the area under the receiver operating curve. Each table corresponds to a different metric.

Model	**Pre-trained**	**Pooling Layer**	**AUROC(Val)**	**AUROC(Test)**
CNN+MIL	ImageNet	Mean	0.65±0.009	0.64±0.008
		Max	0.64±0.004	0.63±0.008
	PANDA	Mean	0.65±0.004	0.63±0.009
		Max	0.64±0.004	0.64±0.018
ViT+MIL	ImageNet	Mean	0.71±0.008	0.72±0.005
		Max	0.73±0.021	0.73±0.014
	PANDA	Mean	**0.76±0.012**	**0.74±0.009**
		Max	**0.76±0.020**	0.73±0.009

Model	**pre-trained**	**Pooling Layer**	**F1-Score(Val)**	**F1-score(Test)**
CNN+MIL	ImageNet	Mean	0.50±0.009	0.46±0.005
		Max	0.5±0.008	0.44±0.008
	PANDA	Mean	0.5±0.008	0.46±0.009
		Max	0.46±0.012	0.45±0.012
ViT+MIL	ImageNet	Mean	0.57±0.009	0.56±0.009
		Max	0.51±0.008	0.45±0.009
	PANDA	Mean	**0.59±0.012**	**0.59±0.004**
		Max	0.56±0.009	0.56±0.008

Model	**pre-trained**	**Pooling Layer**	**Sensitivity(Val)**	**Sensitivity(Test)**
CNN+MIL	ImageNet	Mean	0.57±0.012	0.53±0.009
		Max	0.52±0.004	0.49±0.004
	PANDA	Mean	0.53±0.008	0.48±0.004
		Max	0.54±0.008	0.55±0.012
ViT+MIL	ImageNet	Mean	0.69±0.004	0.65±0.004
		Max	0.39±0.009	0.35±0.009
	PANDA	Mean	**0.71±0.009**	**0.66±0.009**
		Max	0.54±0.009	0.49±0.004

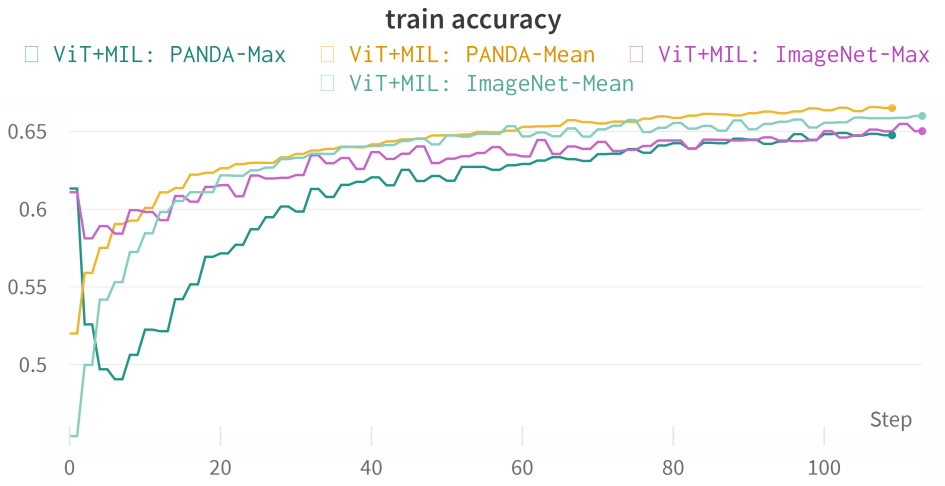
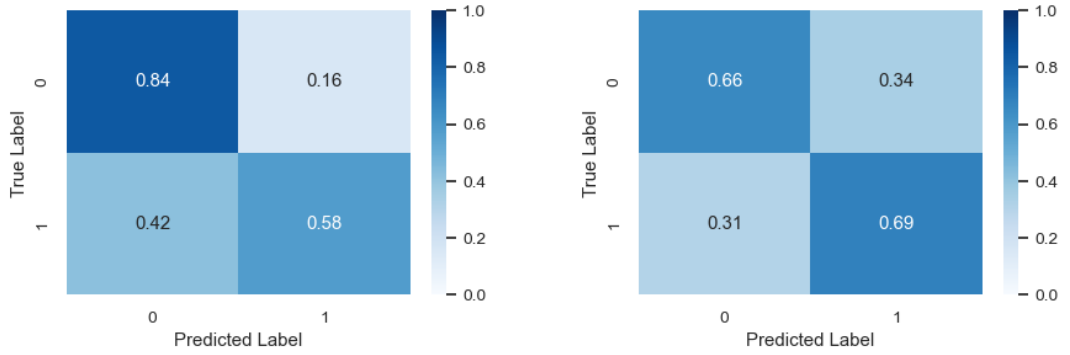


Fig. 5.4: Training accuracy of vision transformers during training.

on PANDA, the model reaches an AUROC of 0.75 in the test set, which is the best performing among all. Similar values are achieved, even with max pooling, when the model is pre-trained on the PANDA dataset. In the mean pooling aggregate layer, there is an improvement of 8 percentage points from ImageNet pre-trained to PANDA pre-trained in the validation sets in ViT+MIL approach. The ImageNet pre-trained with mean pooling has the lowest performance in the test dataset, while all other approaches perform similarly. Even though AUROC values for ViT+MIL: PANDA-mean and ViT+MIL: PANDA-max are the same, a significant difference is observed in the confusion matrix as shown in Figure 5.5. In ViT+MIL: PANDA-mean, negative classes are well classified comparivetly positive classes, whereas for ViT+MIL: PANDA-max, both the classes are approximately classified with similar percentages.

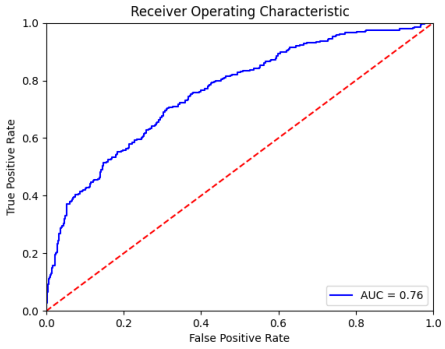


(a) Confusion matrix for ViT+MIL: PANDA-mean.

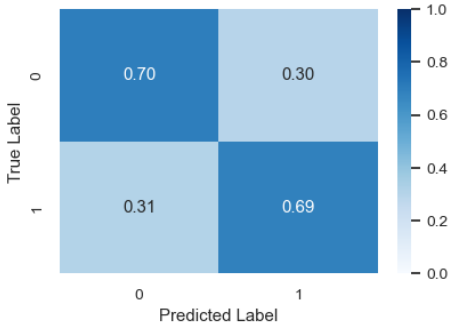
(b) Confusion matrix for ViT+MIL: PANDA-max.

Fig. 5.5: Confusion matrix of relapse prediction using ViT+MIL for best performing model.

Similar to AUROC, F1-score, the sensitivity values also indicate that the ViT+MIL has a considerably better performance than the CNN. Although the differences are



(a) ROC on relapse prediction



(b) Confusion matrix of hierarchical vision transformer

Fig. 5.6: Metrics of Hierarchical vision transformer.

less noticeable on the test set, pre-training on ImageNet appears to usually produce higher scores on the validation set compared to PANDA for both architectures. With the exception of CNN+MIL pre-trained on PANDA, mean pooling typically seems to have produced greater sensitivity values than max pooling across both architectures and pre-training datasets.

Hierarchical vision transformers

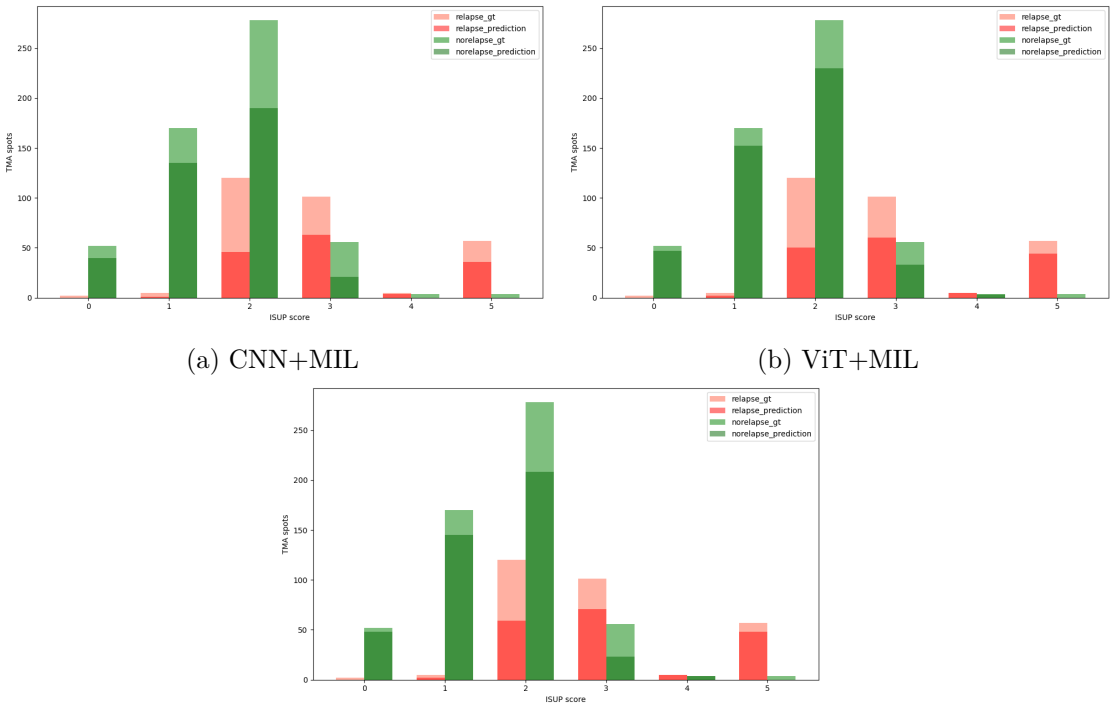
The hierarchical vision transformers have outperformed all other models with certain seeds, indicating better AUROC values, which in turn denotes better differentiability between the positive and negative relapse cases. But overall, with three seeds, HViT does not outperform ViT+MIL as shown in 5.2. From the confusion matrix in Figure 5.6, the model performs best in both the positive and negative classes.

Tab. 5.2: Table displaying results of hierarchical vision transformer.

Datasets	AUROC	F1-score	Sensitivity
Validation	0.75±0.004	0.59±0.004	0.64±0.012
Test	0.75±0.008	0.59±0.004	0.63±0.042

5.3 Analysis based on ISUP grading

For the UKE dataset, pathologists also provided ISUP scores graded for prostate tissues. Figure 5.7 shows the analysis for relapse and no relapse cases based on the ISUP score for CNN+MIL, ViT+MIL and HViT. The plots displayed ground truth and predictions for positive and negative classes of relapse prediction. It can be clearly observed that for ISUP score 2, no relapse cases are well classified with ViT+MIL, followed by HViT. For relapse cases, HViT has better performance than ViT+MIL. Even for the other ISUP scores, ViT+MIL has shown better classification on no relapse cases. From Figure 5.7, it can also be observed that all the models tend to predict lower ISUP grades as no relapse and higher ISUP as relapse cases.



(a) CNN+MIL

(b) ViT+MIL

(c) HViT

Fig. 5.7: Plots displaying relapse and no relapse ground truths and predictions for TMA spots. relapse_gt represents the ground truth count of patients who have cancer relapse within 5 years. relapse_prediction indicates the number of patients who are predicted to have relapse. norelapse_gt represents the ground truth count of patients who have does not have cancer relapse within 5 years. norelapse_prediction indicates the number of patients who are predicted to have no relapse.

6 Discussion

In this chapter, the results presented in chapter 5 are analysed along with different strategies applied. The differences in datasets used are assessed. Finally, the results obtained are compared to previous work and limitations are discussed.

6.1 Datasets

For cancer detection, which is used as a pre-training task for the vision transformers, WSIs are used, and TMA spots are leveraged for relapse prediction. WSIs from the PANDA dataset are high-resolution images that visually represent the tissue sample and have a large set of information, including cell morphology and structures expressing cancer. Though both datasets are similar in structure, the models' performances on cancer detection are highly varied for relapse prediction can be due to the respective label annotations. Cancer detection has achieved an AUROC of 0.97, with a significantly well-distinguishing capacity between cancer and no cancer patches.

Cancer detection is a relatively simpler and more straightforward task, which thus enables the vision transformer to have better performance. In contrast, relapse prediction using the images is more challenging in general as the relapse depends on various additional parameters rather than only on structures in images. The availability of data also plays a significant role in performance. For cancer detection, there is a large dataset of about 4.5 million patches available for this thesis, whereas, for relapse prediction, it is comparatively tiny, with 6000 images. The data accessibility and labelling for cancer classification is easy but human expertise is required, while it is quite cumbersome for relapse as it requires long-term follow-up and has to be integrated int the clinical work flow. With more data availability, the models perform better, especially vision transformers, as they are data-hungry.

Regarding the labelling of the data for PANDA, the structures representing the presence and absence of cancer are clearly defined with well-established markers. Ground truth for relapse is harder to obtain and also noisy as it is not directly indicated through the images but obtained through BCR. The factors of relapse cannot be directly seen in the image, but can include multiple factors, leading to classification challenges. The TMA spot used is actually a small excerpt of the whole prostate and is also smaller than biopsies on the WSIs. That means that not all representative areas for the whole cancer can be seen. Also there can be external factors that do not directly relate to the tissue. This could e.g. be that surgery has been more successful in removing all prostate cancer tissue in one case, whereas in another case, some small amounts of prostate cancer tissue remaining could grow again and cause an increase in PSA (= BCR). Furthermore, there could already be metastases as well.

The image size for relapse has been chosen as 512×512 pixels in UKE instead of 256×256

pixels like PANDA due to the difference in resolution of PANDA and UKE datasets. The resolution of PANDA images is twice the resolution of UKE, which could be a reason for less predictive capacity. Furthermore, the staining differences in the two image types can be another reason for the considerable difference in cancer detection and relapse prediction, even though pre-trained on similar domain data.

6.2 Evaluation of cancer detection

In this section, pre-training implementations of CNN and ViT are discussed. For cancer detection, CNN has a poorer performance than ViTs when using the same image size by almost eight percentage, which can be due to the ViTs' capacity to capture global information and the relationship between the various regions instead of just local information as in CNNs. By dynamically distributing attention to different patches depending on their importance, ViTs might concentrate on the finer details, spotting small patterns suggestive of early malignancy that CNNs could miss. Also, ViTs on pre-training for cancer detection with a smaller image size and a high number of patches with a small patch size has better performance than with a larger image size, small patch count and high patch size. The smaller image sizes naturally enable the model to focus on fine details, which are essential for cancer identification. The smaller image sizes reduce the noise or unnecessary information that does not contribute to the task. There could be a loss of information in the large image sizes due to broader focus, with the possibility of essential structures contributing to detection, might be overlooked. In the case of larger image patches taken from the WSIs, there can be a higher chance of having pixels that are white spaces in the patches in comparison to smaller images, which can be a factor for better performance in smaller images.

Another reason for better performances in smaller images can be due to patch sizes and the effect of patch sequence lengths obtained through the number of patches. Similar to the smaller image sizes, the smaller patch sizes would have more focus than the larger patch sizes which is observed in pre-trainings of cancer detection. Smaller patch sizes resulting in longer sequences, giving a rich contextual environment for the self-attention mechanism to operate on and potentially leading to a better understanding of the spatial relationships in the image, which in turn helps in prediction. Reduced sequence length with bigger patches might have reduced the capacity to capture complex information between different regions of the image.

6.3 Research questions

R1: Are vision transformers suitable for relapse prediction of in-house pathology data?

The research question of whether vision transformers are effective for relapse prediction on in-house UKE pathology data is discussed below. For comparison, EfficientNet, one of the best-performing models on the in-house data, is chosen. As the image sizes are large, multiple instance learning in combination is set as the baseline model for comparison. Vision transformers have significantly outperformed CNNs on the in-house data because of their self-attention mechanism, which helps in capturing detailed structures in the

histopathology images, unlike CNNs, which focus on local patterns. With CNNs, it can be challenging for determining overall contextual relationships in images.

R2: Can the performance of ViTs in relapse prediction be enhanced through the incorporation of additional domain-related data for pre-training?

This thesis explores the possibility of pre-training on extra domain-related data to improve the performance of vision transformers (ViTs) in relapse prediction. The results in Table 5.1 show that pre-training of ViTs on another histopathology prostate cancer dataset did not yield better results when compared to pre-trained models on ImageNet data. The two reasons for the marginal performance improvement could be the relevance of the data and the size of the datasets. Though the PANDA dataset is one of the largest datasets for prostate cancer, it is comparatively small in comparison to the millions of images on ImageNet. As ViTs are data hungry, the data used for pre-training might not have been sufficiently large to lead to an improvement with pre-training.

R3: How does leveraging multiple image sizes in histopathology images impact prediction using hierarchical ViTs in relapse prediction?

One of the methods investigated is the use of Hierarchical vision transformers (ViTs) and the leveraging of multiple image sizes in histopathology image processing, which aimed to better capture varying scales of characteristics that may be relevant for relapse prediction. Relationships at various scales and regions of the images are intended to be more successfully modelled using the hierarchical method. However, from the research results shown in the 5.2, hierarchical models did not seem to have this effect on relapse prediction, which can be due to the complexity of the model implemented. The complexity of the hierarchical models is increased by additional encoder layers, which attempt to learn features on different scales. This additional complexity does not help the model to learn more features as they might have already been learnt in ViT, which can lead to overfitting. The image sizes are also small, so the model seems to be overfitting with the available dataset

The same hyperparameters used for both ViT+MIL and hierarchical ViT could be an issue as different hyperparameter tuning might be required depending on the model's complexity. Finally, the hierarchical model is trained with a batch size of one, which could also lead to overfitting and underperformance of the model.

6.4 Model training

CNNs have shown remarkable progress in classifying images over the last few years. The primary factor in their success is their operational structure. Convolutional layers are used by CNNs to apply filters to tiny, local parts of an image, allowing the network to recognise regional patterns and spatial information. However, ViTs are a current architecture that has grown in favour recently and can perform better or equally to CNNs (Dosovitskiy et al., 2020). The training parameters of both models vary greatly. In this thesis, the base version of the ViT is employed for pre-training and fine-tuning, which has 330M trainable parameters, while for CNN, there are only 20M parameters as

shown in Table 4.3. The computational resources needed to train these two model types vary largely but could not be compared precisely in this work due to the limitations in parallel processing of the input data. The training on other versions of ViT, ViT-Large and ViT-Huge has not been not attempted as they could lead to overfitting due to limited amounts of data. Several experiments were conducted on a subset of the PANDA dataset to determine the optimal configuration of weight decay and learning rate to train the model.

6.5 Limitations

Despite the fact that AI applications have made considerable advancements and contributions to digital pathology, there are still a number of challenges to be overcome. Several technological and computational challenges must be solved before computer-assisted image analysis of digital pathology becomes a common clinical diagnostic method. In this current work, the most important limitations concern to datasets and computational resources.

Data

Generally, for ViTs, datasets required for training are very large. A large number of parameters enables the model to learn complicated representations, but it also demands a lot of data to appropriately adjust these parameters and prevent overfitting. Large datasets generally have a greater range of data distribution, which can help a model become more resilient and better at generalising to new data. Furthermore, ViT models usually perform better when trained and assessed on big datasets than small datasets, where performance saturation may occur quickly.

Computational resources

In the current thesis, all the experiments were conducted on a Kubernetes cluster with limited CPU resources, which constrained the capacity to execute computationally intensive vision transformer training. This limitation slowed down the experimentation process. Also, smaller batch sizes were used, which are generally not suited for generalization compared to models trained on large datasets.

7 Conclusion and Outlook

Computational pathology provides a promising path for augmenting and enhancing the skills of traditional pathology, notably in the domain of prostate cancer. Instead of relying on Gleason grading of prostate cancer, which is highly variable from pathologist to pathologist, this thesis proposes directly predicting cancer relapse after radical prostatectomy.

The literature review reveals that methodologies for cancer detection and survival prediction using histopathology images are diverse. Due to variations in the datasets, and their availability or problem conceptualization, none of the available models can be used to compare the problem presented in this thesis. Recently, vision transformers have shown remarkable performance in computer vision tasks, including medical imaging. In this work, the potential of vision transformers and their variant hierarchical vision transformers for relapse prediction is explored. However, vision transformers require huge datasets to achieve performance comparable to CNNs. Unlike models for classifying ordinary objects, pre-trained models on histopathological prostate cancer detection or relapse prediction are not open-source. This is also addressed in this thesis by pre-training the models with domain-related data. Later, performance of the hierarchical vision transformers is explored.

Due to the large image sizes, multiple instance learning combined with model architectures is leveraged to evaluate the performances. The results highlilght that ViT combined with MIL outperforms CNN with MIL, showing that ViT are superior in classifying the occurrence of relapse on our in-house data.

To investigate whether pre-training on domain-related data has an influence on relapse prediction, firstly, the ViT model is trained for the cancer detection task with the PANDA dataset. Later, this pre-trained model is fine-tuned with the smaller in-house UKE dataset for relapse prediction within five years of treatment. Several experiments were conducted to choose the hyperparameters. The experiments presented that the pre-training models on PANDA, did not have a significant impact on the relapse prediction. It could be concluded that more images are required pre-training to improve the performance.

Furthermore, the performance of the hierarchical vision transformer was examined. For this task, two ViTs were pre-trained for the cancer detection task with different image sizes, and then combined hierarchically to obtain the relapse prediction for a single TMA spot. The experiments indicated that there is not much performance increase compared to ViT+MIL, this could mean that HViT are not necessarily apt for this task given, which might be due to the relatively small image sizes of the image size 2048×2048 pixels.

7.1 Future work

This thesis provided a foundation that ViTs can perform better when compared to CNNs for relapse prediction of prostate cancer using histopathology images. There are several aspects that can be added to a ViT to achieve more accurate performance in distinguishing relapse or relapse-free conditions. Firstly, ViT can be pre-trained with more and varied histopathology data, not only for prostate cancer but any other disease, which could aid in the performance increase of ViT. Secondly, architectural changes in ViT, like a decrease in encoder layers, could also benefit the tasks if the model is overfitting. In addition, different pre-training approaches like self-supervised learning, e.g., DINO, can be implemented rather than task-specific pre-training (Chen et al., 2022b). For HViT, rather than pre-training ViTs in two stages, combined pre-training could be beneficial. Also, a reduction in the number of attention layers for multi-staged ViT could lead to less overfitting and better training of the data.

Bibliography

Esther Abels, Liron Pantanowitz, Famke Aeffner, Mark D Zarella, Jeroen van der Laak, Marilyn M Bui, Venkata NP Vemuri, Anil V Parwani, Jeff Gibbs, Emmanuel Agosto-Arroyo, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *The Journal of pathology*, 249(3):286–294, 2019.

Abeer Aljuaid and Mohd Anwar. Survey of supervised learning for medical image processing. *SN Computer Science*, 3(4):292, 2022.

William C Allsbrook Jr, Kathy A Mangold, Maribeth H Johnson, Roger B Lane, Cynthia G Lane, and Jonathan I Epstein. Interobserver reproducibility of gleason grading of prostatic carcinoma: general pathologist. *Human pathology*, 32(1):81–88, 2001.

Nasar Yousuf Alwahaibi, Azza Sarhan Alkhatri, and Johanes Selva Kumar. Hematoxylin and eosin stain shows a high sensitivity but sub-optimal specificity in demonstrating iron pigment in liver biopsies. *International Journal of Applied and Basic Medical Research*, 5(3):169, 2015.

Eirini Arvaniti, Kim S Fricker, Michael Moret, Niels Rupp, Thomas Hermanns, Christian Fankhauser, Norbert Wey, Peter J Wild, Jan H Rueschoff, and Manfred Claassen. Automated gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific reports*, 8(1):12054, 2018.

Ayesha S Azam, Islam M Miligy, Peter KU Kimani, Heeba Maqbool, Katherine Hewitt, Nasir M Rajpoot, and David RJ Snead. Diagnostic concordance and discordance in digital pathology: a systematic review and meta-analysis. *Journal of Clinical Pathology*, 74(7):448–455, 2021.

Pierre Baldi and Peter J Sadowski. Understanding dropout. *Advances in neural information processing systems*, 26, 2013.

AD Belsare and MM Mushrif. Histopathological image analysis using image processing techniques: An overview. *Signal & Image Processing*, 3(4):23, 2012.

Subrata Bhattacharjee, Kobiljon Ikromjanov, Yeong-Byn Hwang, Rashadul Islam Sumon, Hee-Cheol Kim, and Heung-Kook Choi. Detection and classification of prostate cancer using dual-channel parallel convolution neural network. In *Proceedings of the Future Technologies Conference (FTC) 2021, Volume 2*, pages 66–83. Springer, 2022.

Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*, 21(2):233–241, 2020.

Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022.

Nikolay Burlutskiy, Nicolas Pinchaud, Feng Gu, Daniel Hägg, Mats Andersson, Lars Björk, Kristian Eurén, Cristina Svensson, Lena Kajland Wilén, and Martin Hedlund. Segmenting potentially cancerous areas in prostate biopsies using semi-automatically annotated data. *arXiv preprint arXiv:1904.06969*, 2019.

Hongbin Cai, Xiaobing Feng, Ruomeng Yin, Youcai Zhao, Lingchuan Guo, Xiangshan Fan, and Jun Liao. Mist: multiple instance learning network based on swin transformer for whole slide image classification of colorectal adenomas. *The Journal of Pathology*, 2022.

Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

American Cancer society Cancer.org. Key statistics of prostate cancer. `https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html`, 2021. Accessed: 2023-01-04.

Haoyuan Chen, Chen Li, Ge Wang, Xiaoyan Li, Md Mamunur Rahaman, Hongzan Sun, Weiming Hu, Yixin Li, Wanli Liu, Changhao Sun, et al. Gashis-transformer: A multi-scale visual transformer approach for gastric histopathological image detection. *Pattern Recognition*, 130:108827, 2022a.

Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4025, 2021.

Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022b.

En Cheng, Fang-Shu Ou, Chao Ma, Donna Spiegelman, Sui Zhang, Xin Zhou, Tiffany M Bainter, Leonard B Saltz, Donna Niedzwiecki, Robert J Mayer, et al. Diet-and lifestyle-based prediction models to estimate cancer recurrence and death in patients with stage iii colon cancer (calgb 89803/alliance). *Journal of Clinical Oncology*, 40(7): 740–751, 2022.

Stephanie Cheon, Arnav Agarwal, Marko Popovic, Milica Milakovic, Michael Lam, Wayne Fu, Julia DiGiovanni, Henry Lam, Breanne Lechner, Natalie Pulenzas, et al. The accuracy of clinicians' predictions of survival in advanced cancer: a review. *Ann Palliat Med*, 5(1):22–29, 2016.

Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. In *Machine learning techniques for multimedia: case studies on organization and retrieval*, pages 21–49. Springer, 2008.

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

Luca Deininger, Bernhard Stimpel, Anil Yuce, Samaneh Abbasi-Sureshjani, Simon Schönenberger, Paolo Ocampo, Konstanty Korski, and Fabien Gaire. A comparative study between vision transformers and cnns in digital pathology. *arXiv preprint arXiv:2206.00389*, 2022.

Oscar Jiménez del Toro, Manfredo Atzori, Sebastian Otálora, Mats Andersson, Kristian Eurén, Martin Hedlund, Peter Rönnquist, and Henning Müller. Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score. In *Medical Imaging 2017: Digital Pathology*, volume 10140, pages 165–173. SPIE, 2017.

Howard B Demuth, Mark H Beale, Orlando De Jess, and Martin T Hagan. *Neural network design*. Martin Hagan, 2014.

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

Anastasiia Petrivna Denysenko, Taras Ruslanovych Savchenko, Anatolii Stepanovych Dovbysh, Anatolii Mykolaiovych Romaniuk, and Roman Andriiovych Moskalenko. Artificial intelligence approach in prostate cancer diagnosis: Bibliometric analysis. 2022.

Esther Dietrich. *Deep learning-based discrete-time survival prediction on prostate cancer histopathology images*. PhD thesis, Universität Hamburg, 2022.

Esther Dietrich, Patrick Fuhlert, Anne Ernst, Guido Sauter, Maximilian Lennartz, H Siegfried Stiehl, Marina Zimmermann, and Stefan Bonn. Towards explainable end-to-end prostate cancer relapse prediction from h&e images combining self-attention multiple instance learning with a recurrent neural network. In *Machine Learning for Health*, pages 38–53. PMLR, 2021.

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.

Lourdes Duran-Lopez, Juan P Dominguez-Morales, Antonio Felix Conde-Martin, Saturnino Vicente-Diaz, and Alejandro Linares-Barranco. Prometeo: A cnn-based computer-aided diagnosis system for wsi prostate cancer detection. *IEEE Access*, 8:128613–128628, 2020.

David M Dutton and Gerard V Conroy. A review of machine learning. *The knowledge engineering review*, 12(4):341–367, 1997.

Lars Egevad, Roberta Mazzucchelli, and Rodolfo Montironi. Implications of the international society of urological pathology modified gleason grading system. *Archives of pathology & laboratory medicine*, 136(4):426–434, 2012.

Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, and Peter A Humphrey. The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. *The American journal of surgical pathology*, 40(2):244–252, 2016.

Lei Fan, Arcot Sowmya, Erik Meijering, and Yang Song. Learning visual features by colorization for slide-consistent survival prediction from whole slide 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 592–601. Springer, 2021.

M. Furihata and T. Takeuchi. *Gleason grading*, pages 1904–1907. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017.

Matt W Gardner and SR Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

Donald F Gleason and George T Mellinger. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *The Journal of urology*, 111(1):58–64, 1974.

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

Jennifer Gordetsky and Jonathan Epstein. Grading of prostatic adenocarcinoma: current state and prognostic implications. *Diagnostic pathology*, 11:1–8, 2016.

David J Grignon. Prostate cancer reporting and staging: needle biopsy and radical prostatectomy specimens. *Modern Pathology*, 31:96–109, 2018.

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

Jun Han and Claudio Moraga. The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International workshop on artificial neural networks*, pages 195–201. Springer, 1995.

Matthew G Hanna, Anil Parwani, and Sahussapont Joseph Sirintrapun. Whole slide imaging: technology and applications. *Advances in Anatomic Pathology*, 27(4):251–259, 2020.

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

Axel Heidenreich. Guidelines and counselling for treatment options in the management of prostate cancer. *Prostate Cancer*, pages 131–162, 2007.

Céline N Heinz, Amelie Echle, Sebastian Foersch, Andrey Bychkov, and Jakob Nikolas Kather. The future of artificial intelligence in digital pathology–results of a survey across stakeholder groups. *Histopathology*, 80(7):1121–1127, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Yaoshiang Ho and Samuel Wookey. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE access*, 8:4806–4813, 2019.

Yajie Hu, Feng Su, Kun Dong, Xinyu Wang, Xinya Zhao, Yumeng Jiang, Jianming Li, Jiafu Ji, and Yu Sun. Deep learning system for lymph node quantification and metastatic cancer identification from whole-slide pathology images. *Gastric Cancer*, 24:868–877, 2021.

Wei Huang, Ramandeep Randhawa, Parag Jain, Samuel Hubbard, Jens Eickhoff, Shivaani Kummar, George Wilding, Hirak Basu, and Rajat Roy. A novel artificial intelligence–powered method for prediction of early recurrence of prostate cancer after prostatectomy and cancer drivers. *JCO Clinical Cancer Informatics*, 6:e2100131, 2022.

Tim Hulsen. An overview of publicly available patient-centered prostate cancer datasets. *Translational andrology and urology*, 8(Suppl 1):S64, 2019.

Kobiljon Ikromjanov, Subrata Bhattacharjee, Yeong-Byn Hwang, Rashadul Islam Sumon, Hee-Cheol Kim, and Heung-Kook Choi. Whole slide image analysis and detection of prostate cancer using vision transformers. In *2022 international conference on artificial intelligence in information and communication (ICAIIC)*, pages 399–402. IEEE, 2022.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

Robert A Jacobs. Increased rates of convergence through learning rate adaptation. *Neural networks*, 1(4):295–307, 1988.

Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. *arXiv preprint arXiv:1702.05659*, 2017.

Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

Luiz Carlos Uchôa Junqueira and José Carneiro. Basic histology: text & atlas. 2005.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 1(8), 2017.

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

Marc D Kohli, Ronald M Summers, and J Raymond Geis. Medical image data and datasets in the era of machine learning—whitepaper from the 2016 c-mimi meeting dataset session. *Journal of digital imaging*, 30:392–399, 2017.

Jan Kukačka, Vladimir Golkov, and Daniel Cremers. Regularization for deep learning: A taxonomy. *arXiv preprint arXiv:1710.10686*, 2017.

Neeraj Kumar, Ruchika Verma, Ashish Arora, Abhay Kumar, Sanchit Gupta, Amit Sethi, and Peter H Gann. Convolutional neural networks for prostate cancer recurrence prediction. In *Medical Imaging 2017: Digital Pathology*, volume 10140, pages 106–117. SPIE, 2017.

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

Jake Lever, Martin Krzywinski, and Naomi Altman. Points of significance: model selection and overfitting. *Nature methods*, 13(9):703–705, 2016.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021a.

Chunyuan Li, Xinliang Zhu, Jiawen Yao, and Junzhou Huang. Hierarchical transformer for survival prediction using multimodality whole slide images and genomics. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4256–4262. IEEE, 2022.

Jiayun Li, Wenyuan Li, Anthony Sisk, Huihui Ye, W Dean Wallace, William Speier, and Corey W Arnold. A multi-resolution model for histopathology image classification and localization with multiple instance learning. *Computers in biology and medicine*, 131:104253, 2021b.

Zhongxiao Li, Yuwei Cong, Xin Chen, Jiping Qi, Jingxian Sun, Tao Yan, He Yang, Junsi Liu, Enzhou Lu, Lixiang Wang, et al. Vision transformer-based weakly supervised histopathological image analysis of primary brain tumors. *IScience*, 26(1), 2023.

Bernard Lobel. Does localized prostate cancer exist? *Prostate Cancer*, pages 101–107, 2007.

Stacy Loeb, Marc A Bjurlin, Joseph Nicholson, Teuvo L Tammela, David F Penson, H Ballentine Carter, Peter Carroll, and Ruth Etzioni. Overdiagnosis and overtreatment of prostate cancer. *European urology*, 65(6):1046–1055, 2014.

Henk B Luiting and Monique J Roobol. 10 prostatakrebs-früherkennung: Stand und evidenz der methoden. 2019.

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

Tom M Mitchell and Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

Emmanuel Montagnon, Milena Cerny, Alexandre Cadrin-Chênevert, Vincent Hamilton, Thomas Derennes, André Ilinca, Franck Vandenbroucke-Menu, Simon Turcotte, Samuel Kadoury, and An Tang. Deep learning workflow in radiology: a primer. *Insights into imaging*, 11:1–15, 2020.

Kunal Nagpal, Davis Foote, Yun Liu, Po-Hsuan Cameron Chen, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L Smith, Arash Mohtashamian, James H Wren, et al. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *NPJ digital medicine*, 2(1):48, 2019.

Soo Jeong Nam, Yosep Chong, Chan Kwon Jung, Tae-Yeong Kwak, Ji Youl Lee, Jihwan Park, Mi Jung Rho, and Heounjeong Go. Preference and demand for digital pathology and computer-aided diagnosis among korean pathologists: A survey study focused on prostate needle biopsy. *Applied Sciences*, 11(16):7380, 2021.

Sridhar Narayan. The generalized sigmoid activation function: Competitive supervised learning. *Information sciences*, 99(1-2):69–82, 1997.

Mustafa Umit Oner, Mei Ying Ng, Danilo Medina Giron, Cecilia Ee Chen Xi, Louis Ang Yuan Xiang, Malay Singh, Weimiao Yu, Wing-Kin Sung, Chin Fong Wong, and Hwee Kuan Lee. An ai-assisted tool for efficient prostate cancer diagnosis. *bioRxiv*, pages 2022–02, 2022.

Mike Parsons and Heike Grabsch. How to make tissue microarrays. *Diagnostic histopathology*, 15(3):142–150, 2009.

Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. Tunability: Importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research*, 20(1):1934–1965, 2019.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.

Emad A Rakha, Michael Toss, Sho Shiino, Paul Gamble, Ronnachai Jaroensri, Craig H Mermel, and Po-Hsuan Cameron Chen. Current and future applications of artificial intelligence in pathology: a clinical perspective. *Journal of clinical pathology*, 74(7): 409–414, 2021.

Prashanth Rawla. Epidemiology of prostate cancer. *World journal of oncology*, 10(2):63, 2019.

Jian Ren, Ilker Hacihaliloglu, Eric A Singer, David J Foran, and Xin Qi. Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 201–209. Springer, 2018.

RKI Zentrum RKI. Krebs in deutschland für 2017/2018 (13. ausgabe). *Robert Koch-Institut: Berlin, Germany*, 2021.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Sagar Sharma, Simone Sharma, and Anidhya Athaiya. Activation functions in neural networks. *Towards Data Sci*, 6(12):310–316, 2017.

Milad Sikaroudi, Maryam Hosseini, Ricardo Gonzalez, Shahryar Rahnamayan, and HR Tizhoosh. Generalization of vision pre-trained models for histopathology. *Scientific reports*, 13(1):6065, 2023.

Ronald Simon, Martina Mirlacher, and Guido Sauter. Tissue microarrays. *Biotechniques*, 36(1):98–105, 2004.

Bhaswati Singha Deo, Mayukha Pal, Prasanta K Panigrahi, and Asima Pradhan. Supremacy of attention based convolution neural network in classification of oral cancer using histopathological images. *medRxiv*, pages 2022–11, 2022.

Nitin Singhal, Shailesh Soni, Saikiran Bonthu, Nilanjan Chattopadhyay, Pranab Samanta, Uttara Joshi, Amit Jojera, Taher Chharchhodawala, Ankur Agarwal, Mahesh Desai, et al. A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies. *Scientific reports*, 12(1):3383, 2022.

Emily Sohn. Screening: diagnostic dilemma. *Nature*, 528(7582):S120–S122, 2015.

René Sotelo, Juan Arriaga, Raed A Azhar, and Inderbir S Gill. A patient's guide.

Maximilian Springenberg, Annika Frommholz, Markus Wenzel, Eva Weicken, Jackie Ma, and Nils Strodthoff. From cnns to vision transformers–a comprehensive evaluation of deep learning models for histopathology. *arXiv preprint arXiv: 2204.05044*, 2022.

Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. *Optimization for machine learning*. Mit Press, 2012.

Peter Ström, Kimmo Kartasalo, Henrik Olsson, Leslie Solorzano, Brett Delahunt, Daniel M Berney, David G Bostwick, Andrew J Evans, David J Grignon, Peter A Humphrey, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology*, 21(2):222–232, 2020.

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

Anusha Maria Thomas, G Adithya, AS Arunselvan, and R Karthik. Detection of breast cancer from histopathological images using image processing and deep-learning. In *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*, pages 1008–1015. IEEE, 2022.

Hamid Reza Tizhoosh and Liron Pantanowitz. Artificial intelligence and digital pathology: challenges and opportunities. *Journal of pathology informatics*, 9(1):38, 2018.

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.

Peter Walhagen, Ewert Bengtsson, Maximilian Lennartz, Guido Sauter, and Christer Busch. Ai-based prostate analysis system trained without human supervision to predict patient outcome from tissue samples. *Journal of Pathology Informatics*, 13:100137, 2022.

Jing Wei, Xuan Chu, Xiang-Yu Sun, Kun Xu, Hui-Xiong Deng, Jigen Chen, Zhongming Wei, and Ming Lei. Machine learning in materials science. *InfoMat*, 1(3):338–358, 2019.

Frederik Wessels, Max Schmitt, Eva Krieghoff-Henning, Tanja Jutzi, Thomas S Worst, Frank Waldbillig, Manuel Neuberger, Roman C Maron, Matthias Steeg, Timo Gaiser, et al. Deep learning approach to predict lymph node metastasis directly from primary tumour histology in prostate cancer. *BJU international*, 128(3):352–360, 2021.

Bethany Jill Williams, Jessica Lee, Karin A Oien, and Darren Treanor. Digital pathology access and usage in the uk: results from a national survey on behalf of the national cancer research institute's cm-path initiative. *Journal of clinical pathology*, 71(5): 463–466, 2018.

Ellery Wulczyn, David F Steiner, Zhaoyang Xu, Apaar Sadhwani, Hongwu Wang, Isabelle Flament-Auvigne, Craig H Mermel, Po-Hsuan Cameron Chen, Yun Liu, and Martin C Stumpe. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PloS one*, 15(6):e0233678, 2020.

Yan Xu, Zhipeng Jia, Liang-Bo Wang, Yuqing Ai, Fang Zhang, Maode Lai, Eric I Chang, et al. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC bioinformatics*, 18(1): 1–17, 2017.

Yun Xu and Royston Goodacre. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of analysis and testing*, 2(3): 249–262, 2018.

Yoichiro Yamamoto, Toyonori Tsuzuki, Jun Akatsuka, Masao Ueki, Hiromu Morikawa, Yasushi Numata, Taishi Takahara, Takuji Tsuyuki, Kotaro Tsutsumi, Ryuto Nakazawa, et al. Automated acquisition of explainable knowledge from unannotated histopathology images. *Nature communications*, 10(1):5642, 2019.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26:289–315, 2007.

Magdy Abd-Elghany Zeid, Khaled El-Bahnasy, and SE Abo-Youssef. Multiclass colorectal cancer histology images classification using vision transformers. In *2021 tenth international conference on intelligent computing and information systems (ICICIS)*, pages 224–230. IEEE, 2021.

Jingwei Zhang, Ke Ma, John Van Arnam, Rajarsi Gupta, Joel Saltz, Maria Vakalopoulou, and Dimitris Samaras. A joint spatial and magnification based attention framework for large scale histopathology classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3776–3784, 2021.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.