SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Biomedical Computing

Deep Learning Based Analysis of Tumor-infiltrating Lymphocytes in H&E Stained Histological Sections for Survival Prediction of Breast Cancer Patients

Margaryta Olenchuk

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Biomedical Computing

Deep Learning Based Analysis of Tumor-infiltrating Lymphocytes in H&E Stained Histological Sections for Survival Prediction of Breast Cancer Patients

Deep Learning basierte Analyse von tumorinfiltrierenden Lymphozyten in H&E gefärbten histologischen Schnitten zur Überlebensvorhersage von Brustkrebspatienten

Author:	Margaryta Olenchuk
Supervisor:	Prof. Dr. Peter Schüffler
Advisor:	Dr. Philipp Wortmann, Ansh Kapil
Submission Date:	15.12.2022

I confirm that this master's thesis in biomedical computing is my own work and I have documented all sources and material used.

Munich, 15.12.2022 Margaryta Olenchuk

Acknowledgments

I would like to say thanks to all who supported me in the last few months, to all who endured my mood, and to all, who encouraged me to stay motivated. Thank you.

I would like to take this opportunity to thank my advisor Peter Schüffler for giving me the opportunity to write my thesis at his chair, allowing me to explore the topic to the fullest, and taking time for numerous meetings and detailed discussions. Thank you and hope we will meet in person someday.

Special thanks also go to all Munich AstraZeneca members for providing amazing resources and constant support along the way. Many thanks to my advisors, Philipp Wortmann and Ansh Kapil for trusting me to work on a remarkable topic. I have learned a lot, thanks to you and your guidance. Particular thanks go to Armin Meier who always found time to discuss frustrating details and provided elegant solutions every time. And thanks to all other sync meetings and lunch break discussions with Guillaume Potdevin, Abdullah Hayran, Gordan Prastalo, Bianca Mocanu, and Srividhya Sathya Narayanan.

Finally, I would like to say thanks for all the non-scientific support. To my family, especially my parents and my grandmother, who repeatedly reminded me not to work too hard. Ukrainian Armed Forces and all defenders and volunteers for protection and the chance to do science. To all Ukrainians for sharing these difficult times with me and reminding me that there should still be room for normal even in war times. And thanks to my wonderful friends Lena Shevtsova and Laczik sisters, Dalma and Dori for being there for me.

Thank you.

Abstract

Breast cancer is the most common type of cancer worldwide with a high mortality rate causing millions of cancer-related deaths annually. Optimization of diagnostic biomarkers is essential to improve breast cancer prognosis and therapeutic outcomes. According to the International Immuno-Oncology Biomarker Working Group, early-stage disease clinicopathological risk stratification is currently performed using a limited set of features such as tumor size and lymph node status, that do not stratify patients with sufficient granularity to permit selection for clinical trials.

The quantification of mononuclear immune cells that infiltrate tumor tissue, named tumor-infiltrating lymphocytes (TILs) is a clinically useful biomarker for breast cancer progression. TILs in tumors and the surrounding microenvironment are thought to reflect the ongoing anti-tumor immune response of the host. TIL analysis is typically performed by pathologists that manually estimate the proportion of TILs in a histological appearance.

The automatic assessment of TILs by computational image analysis is valuable for standardization and potential use in a clinical setting. This thesis shows that TIL evaluation can be automated with deep learning-based techniques. The developed pipeline segments tissue relevant for TIL score, such as stromal and tumorous regions using DeepLabv3+, and detects TILs based on the publicly released TiGER challenge data containing digital pathology images of Her2 positive and Triple Negative breast cancer whole-slide images. The tissue segmentation model scored 0.85 Dice score on the test set and TIL detection was approached as segmentation with a transformer-based architecture - SegFormer, which is novel for computational pathology, and scored 0.66 F1-score on the test set.

The application of deep learning approaches permitted the discovery of image-based features that would be demanding to identify per hand, particularly if they only exist in small groups of patients. It made it feasible not only to show the effect of TIL density in stroma but experiment with various borders of stroma, tumor-associated stroma, and tumor, as well as novel heterogeneity features. Applied to TCGA-BRCA clinical data, TIL features demonstrated their ability to successfully stratify the patients into groups of high and low TIL density. In Cox proportional hazards model TIL density in tumor associated stroma border of 100 µm achieved concordance of 0.59, p-value 0.00018, and 0.78 hazard ratio, supporting the assumption that the high level of TILs plays a role in prolonged survival probability and therefore can be used as a prognostic marker for breast cancer patients.

Kurzfassung

Brustkrebs ist weltweit die häufigste Krebsart mit einer hohen Sterblichkeitsrate, die jährlich Millionen von krebsbedingten Todesfällen verursacht. Die Optimierung diagnostischer Biomarker ist für die Verbesserung der Brustkrebsprognose und der therapeutischen Ergebnisse von essenzieller Bedeutung. Nach Angaben der International Immuno-Oncology Biomarker Working Group erfolgt die klinisch-pathologische Risikostratifizierung im Frühstadium der Erkrankung derzeit anhand einer begrenzten Anzahl von Merkmalen wie Tumorgröße und Lymphknotenstatus, die die Patienten nicht ausreichend genau einteilen, um eine Auswahl für klinische Studien zu ermöglichen.

Die Quantifizierung von mononukleären Immunzellen, die in das Tumorgewebe eindringen und als tumorinfiltrierende Lymphozyten (TILs) bezeichnet werden, ist ein klinisch nützlicher Biomarker für das Fortschreiten von Brustkrebs. Es wird angenommen, dass TILs in Tumoren und der umgebenden Mikroumgebung die laufende Anti-Tumor-Immunreaktion des Wirts widerspiegeln. Die TIL-Analyse wird in der Regel von Pathologen durchgeführt, die den Anteil der TILs in einem histologischen Befund manuell schätzen.

Die automatische Bewertung von TILs durch computergestützte Bildanalyse ist wertvoll für die Standardisierung und den potenziellen Einsatz in einem klinischen Umfeld. Diese Arbeit zeigt, dass die TIL-Bewertung mit Deep-Learning-basierten Techniken automatisiert werden kann. Die entwickelte Pipeline segmentiert Gewebe, das für die TIL-Bewertung relevant ist, wie z. B. stromale und tumoröse Regionen unter Verwendung von DeepLabv3+, und erkennt TILs auf der Grundlage der öffentlich veröffentlichten TiGER-Challenge-Daten, die digitale Pathologiebilder von Her2-positiven und dreifach negativen Brustkrebs-Ganzbildaufnahmen enthalten. Das Gewebesegmentierungsmodell erzielte 0,85 Dice-Wert in der Testreihe und die TIL-Erkennung wurde als Segmentierung mit einer transformatorbasierten Architektur - SegFormer - angegangen, die für die computergestützte Pathologie neu ist und 0,66 F1-Wert in der Testreihe erzielte.

Die Anwendung von Deep-Learning-Ansätzen ermöglichte die Entdeckung von bildbasierten Merkmalen, die per Hand nur schwer zu identifizieren wären, insbesondere wenn sie nur bei kleinen Patientengruppen vorkommen. Dadurch konnte nicht nur die Auswirkung der TIL-Dichte im Stroma aufgezeigt, sondern auch mit verschiedenen Grenzen des Stromas, tumorassoziierten Stromas und Tumors sowie mit neuen Heterogenitätsmerkmalen experimentiert werden. Bei der Anwendung auf die klinischen Daten des TCGA-BRCA zeigten die TIL-Merkmale ihre Fähigkeit, die Patienten erfolgreich in Gruppen mit hoher und niedriger TIL-Dichte zu stratifizieren. Im Cox-Proportional-Hazards-Modell erreichte die TIL-Dichte in der tumorassoziierten Stromagrenze von 100 µm eine Übereinstimmung von 0,59, einen p-Wert von 0,00018 und eine Hazard Ratio von 0,78, was die Annahme stützt, dass die hohe TIL-Dichte eine Rolle bei der verlängerten Überlebenswahrscheinlichkeit spielt und daher als

prognostischer Marker für Brustkrebspatientinnen verwendet werden kann.

Contents

1 Introduction

Breast cancer is the most common form of cancer diagnosed worldwide and the leading cause of cancer-related death among women [1]. It is a heterogeneous disease, consisting of several morphological and molecular subtypes. The molecular subtypes are among the prime factors to characterize breast cancer. There are four common clinically relevant subtypes [2] defined on the status of three receptors, namely the Hormonal Receptor (HR, which is positive if either Estrogen Receptor (ER) or Progesterone Receptor (PR) is positive) and the human epidermal growth factor receptor 2 (Her2):

1. Luminal A (HR positive, Her2 negative)
2. Luminal B (HR positive, Her2 positive)
3. Her2 enriched (HR negative, Her2 positive)
4. Triple Negative (HR negative, Her2 negative)

Regardless of the subtype, breast cancer is primarily classified by its histological appearance. Thus for diagnostic confirmation, a patient's biopsy or surgical resection samples are sectioned onto microscope slides for staining, often with hematoxylin and eosin (H&E), followed by a visual diagnosis by a pathologist. Pathologists examine the tissue for abnormalities that indicate breast cancer. Cancer causes changes in the tissue at the sub-cellular scale, hence an analysis of normal and tumor tissue can provide novel insights into tissue characteristics, lead to a better understanding of mechanisms underlying cancer progression and provide valuable information for medical decision-making such as tumor grading and treatment choices [3].

One of the characteristics of histological images that can be visually assessed by pathologists is lymphocytic infiltration. There are a number of publications that emphasize the prognostic value of tumor-infiltrating lymphocytes (TILs), especially in triple negative (TNBC) and human epidermal growth factor receptor 2 (HER2+) breast cancer [4, 5]. TILs are mononuclear immune cells that infiltrate tumor tissue. They have been detected in almost all solid tumors, including breast cancer [6]. The development and progression of malignant tumors can be characterized by an interaction between the cells in the tumor microenvironment and TILs. In the early stage of HER2+ and TNBC, TILs are detectable in up to 75% of tumors [7]. Studies have shown that an increased degree of lymphocytic infiltration is predictive of better long-term control of the disease [8, 9]. Patients with a high proportion of TILs in the tumor tissue and high immunogenicity of the tumor were shown to respond better to chemotherapy [10]. Accumulating evidence indicates that tumor-infiltrating lymphocytes are clinically useful biomarkers in TNBC and HER2+ and that they play an essential role in cancer progression [11]. Further research and development of TIL related biomarkers would grant clinicians essential prognostic information and promote the research on novel treatments and

therapeutics. For instance, since TILs with exhausted phenotype are associated with loss of antitumor immunity, single-cell RNA Sequencing of TILs has been already performed to search for new immune checkpoint blockade targets that enable the precise definition and even novel development of therapeutic strategies to overcome T-cell exhaustion. Therapeutic approaches to influence T-cell exhaustion have been developed to target proteins CTLA-4, PD-1, and PD-L1 and have proven to be effective in treating melanoma and non-small-cell lung cancer during ongoing trials [12]. TILs in TNBC patients also display immuno-suppressive phenotypes [13] and the number of TILs detected by TNBC patients is one of the highest of all breast cancer subgroups [14] which makes TNBC a valid target for further TILs research.

A valuable contribution to TILs research and any task involving visual analysis of histological images would be method automatization. While manual examination continues to be widely applied in a clinical setting, it is subjective and not scalable to translational and clinical research studies involving large datasets of high-resolution whole slide tissue images (WSIs). Hence, there is a raised demand for reliable and efficient automated methods to complement the traditional manual examination of tissue samples.

With advancing technology and access to a large amount of data, deep learning methods have garnered an interest in computational pathology. There are multiple deep learning-based methods and pipelines that have been proposed for detection and segmentation tasks of WSIs. To stimulate the development of algorithms for automatic TILs evaluation, a special Tumor InfiltratinG lymphocytes in breast cancER (TiGER) [2] challenge was formed. Within this competition, various algorithms were evaluated for the automated assessment of TILs in H&E stained histopathology WSIs that resulted in automatically acquired TILs scores. Those were later internally checked for significance as prognostic values and the concordnace was reported. The clinical focus of the TiGER challenge is on Her2+ and TNBC. It is motivated by research and clinical data that show that Her2+ and TNBC have the worst prognosis making them an intense target of prognostic and predictive biomarker research aimed at improving patient management and prognosis.

This work is closely linked to the TiGER challenge. The goal is to develop a pipeline for HER2 positive and triple negative breast cancer H&E slides that segments tumor and stroma regions, detects TILs, and produces TILs scores as pictured in Figure 1.1 block 1-3.

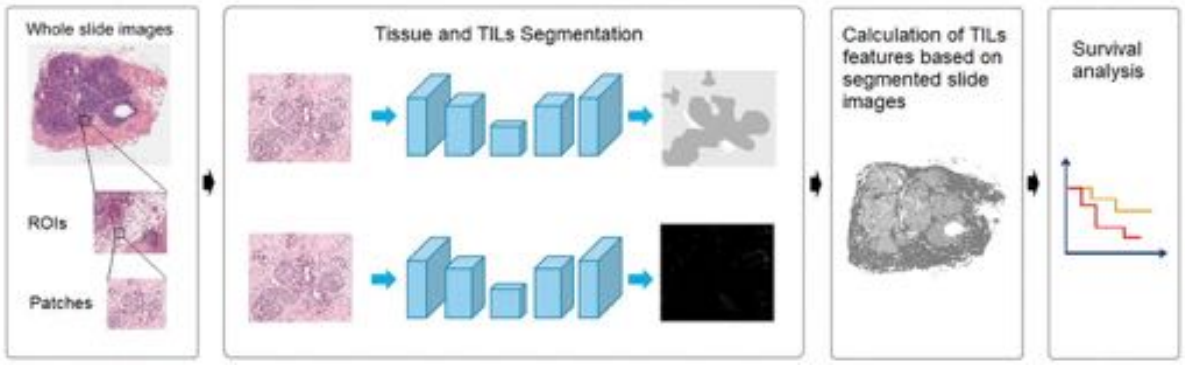


Figure 1.1: Abstract scheme to visually introduce the flow of this thesis work.

This work takes benefit of the annotated region of interests (ROIs) provided by the TiGER challenge for the development of patch-based automated tissue and TIL segmentation. As a step beyond the challenge, the scores based on the degree of lymphocytic infiltration are evaluated on the breast cancer TCGA-BRCA dataset generated by the TCGA Research Network and not on the hidden TiGER dataset which is not available after the end of the competition. The TCGA-BRCA clinical data enables independent broad survival analysis of different experimental TIL characteristics that can be calculated solely based on histological images (Figure 1.1 block 4). As a result, this work aims not only to develop a computational approach to compute TIL score on H&E images of Her2+ and TNBC but also experiment with different TIL scores and show detailed survival analysis based on publically available TCGA-BRCA dataset together with their predictive value for overall patient survival.

2 Related work

2.1 Deep learning-based semantic segmentation

The goal of semantic segmentation is to assign each image pixel to a category label corresponding to the underlying object. Due to the success of deep learning models in a wide range of vision applications, various deep learning-based algorithms have been developed and published in the literature [15]. One of the most prominent deep learning architectures used by the computer vision community include fully convolutional networks (FCNs) [16], encoder-decoders [17], generative adversarial networks (GANs) [18] and recurrent neural networks (RNNs) [19]. As tissue segmentation, TILs detection can also be viewed as a semantic segmentation problem, since detection bounding boxes can be transformed into pseudo-segmentation masks. The focus of the following chapters is to superficially introduce the existing deep learning-based approaches for the histopathological tasks, that can be adapted or extended for tissue and TILs segmentation of breast cancer WSIs.

2.1.1 Fully convolutional networks (FCNs)

FCNs [16] are among the most widely used architectures for computer vision tasks and their general architecture consists of several learnable convolutions, pooling layers, and a final 1×1 convolution. Such models are used on segmentation problems in histology domain such as colon glands segmentation [20], as well as nuclei [21] and TILs [22] segmentation for breast cancer all performed on the Hematoxylin and Eosin (H&E) stained histopathology images. Moreover, the FCN method was applied for semantic segmentation of TCGA [23] breast data set [24], which is also used in this thesis. However, despite its popularity, the conventional FCN model has limitations such as loss of localization and the inability to process potentially useful global context information due to a series of down-sampling and a high sampling rate.

2.1.2 Encoder-decoder networks

A popular group of deep learning models for semantic image segmentation that aims to solve the aforementioned issues of FCNs is based on the convolutional encoder-decoder architecture [17]. Their model consists of two parts, an encoder consisting of convolutional layers and a deconvolution network that consists of deconvolution and unpooling layers that take the feature vector as input and generate a map of pixel-wise class probabilities. An example of such a convolutional encoder-decoder architecture for image segmentation is SegNet [25]. The SegNet's encoder network has 13 convolutional layers with corresponding layers in the decoder. The final decoder output is fed to a multi-class soft-max classifier to

produce class probabilities for each pixel independently. The main feature of SegNet is that the decoder uses pooling indices computed in the max-pooling step of the corresponding encoder to perform non-linear upsampling. This architecture also find a use in histopathology, e.g. colon cancer analysis [26]. There are several encoder-decoder models initially developed for biomedical image segmentation. Ronneberger, O., *et al.* [27] proposed the U-Net model for segmenting biological microscopy images that can train with few annotated images effectively. U-Net has an FCN-like down-sampling part that extracts features with 3×3 convolutions and an up-sampling part. Feature maps from the encoder are copied to the corresponding decoder part of the network to avoid losing pattern information. Besides the segmentation of neuronal structures in electron microscopic recordings demonstrated in the original paper [27], U-Net was applied for numerous histopathology tasks such as nuclei segmentation [28, 29], individual colon glands segmentation [30], epidermal tissue segmentation of skin biopsies [31] and cell segmentation on triple-negative breast cancer patients dataset [32]. A further development of an encoder-decoder model for semantic segmentation of histopathology images is HookNet [33]. The architecture consists of two encoder-decoder branches to extract contextual and fine-grained detailed information and combine it (hook up) for the target segmentation. The model showed improvement compared with single-resolution models and was applied to segment breast cancer tissue sections [33].

Another widely used group of deep learning models for semantic segmentation are the atrous (or dilated) convolutional models that include the DeepLab family [34, 35]. The use of atrous convolutions addresses the decreasing resolution caused by max-pooling and striding and Atrous Spatial Pyramid Pooling analyzes an incoming convolutional feature layer with filters at multiple sampling rates allowing to capture objects and image contexts at multiple scales to robustly segment objects at multiple scales. DeepLabv3+ [36] uses encoder-decoder architecture including atrous separable convolution, composed of a depthwise convolution (spatial convolution for each channel of the input) and pointwise convolution (1×1 convolution with the depthwise convolution as input). Authors [36] demonstrated the effectiveness of DeepLabv3+ model on segmentation of H&E stained breast cancer [37]. Despite all the efforts, even this popular architecture has constraints in learning long-range dependency and spatial correlations due to the inductive bias of locality and weight sharing [38] that may result in the sub-optimal segmentation of complex structures.

2.1.3 Recurrent neural networks (RNNs)

RNNs [19] have proven to be useful in modeling the short/long-term dependencies among pixels to generate segmentation maps. Pixels can be linked together and processed sequentially to model global contexts and improve semantic segmentation. ReSeg [39] is an RNN-based model for semantic segmentation. Each layer is composed of four RNNs that go through the image horizontally and vertically in both directions to provide relevant global information, while convolutional layers extract local features that are then followed by up-sampling layers to recover the predictions at original image resolution. But despite all further developments that showcase the potential for histopathology image segmentation: RACE-net [40] applied for segmentation of the cell nuclei in H&E stained breast cancer slides,

Her2Net [41] segmenting cell membranes and nuclei from human epidermal growth factor receptor-2 (HER2)-stained breast cancer images, etc., an important limitation of RNNs is that, due to their sequential nature, they are comparably slower, since this sequential calculation cannot be easily parallelized.

2.1.4 Transformers

A Transformer in Natural Language Processing is an architecture that aims to solve sequence-to-sequence problems. These models rely on self-attention mechanisms and capture long-range dependencies among tokens (words) in a sentence without using RNNs or convolution. Transformers have also emerged in image semantic segmentation. Recent studies have shown that in computer vision the Transformers can achieve superior performance than CNN-based approaches in various semantic segmentation applications [42]. The state-of-the-art Transformer-based semantic segmentation methods can be often applied either as convolution-free models or/and as CNN-Transformer hybrid models. Swin-Transformer [43] for instance is a pure hierarchical Transformer that can serve as a backbone for various computer vision tasks including semantic segmentation. To tokenize the image, it brakes the image into windows that further consist of patches. It constructs a hierarchical representation of an image by starting from small-sized patches and gradually merging neighboring patches into deeper Transformer layers. Swin-Transformer or its slightly modified successors found its application in the medical domain, often as a backbone, for example for colon cancer segmentation in H&E stained histopathology images [44] or gland segmentation [45]. A further popular fully transformer-based model for semantic segmentation is Segmenter [46]. The encoder consists of Multi-head Self Attention and Multi-Layer Perceptron (MLP) blocks, as well as two-layer norms and residual connections after each block and a linear decoder that bilinearly up-samples the sequence into a 2D segmentation mask. While performing well on scene segmentation [46], is not particularly used in the medical domain. In the field of medical image segmentation, TransUNet [47] was the first attempt to establish self-attention mechanisms by combining transformer with U-Net and proved that transformers can be used as powerful encoders for medical image segmentation. A novel positional-encoding-free Transformer SegFormer [48] set new state-of-the-art in terms of efficiency and accuracy in publicly available semantic segmentation datasets and applied for gland and nuclei segmentation [45]. This architecture remains promising also for semantic segmentation in medical applications due to the positional-encoding-free encoder and lightweight MLP decoder.

2.2 TILs as prognostic biomarker

The overall survival (OS) is the primary endpoint for prognostic analysis in this thesis, the survival methods are well established and include the Kaplan–Meier method [49] to estimate OS and Cox proportional hazard models [50] to quantify the hazard ratio (HR) for the effects of biomarker groups. The following chapter focuses on conducted research for the

development of TILs scores as a prognostic biomarker for survival analysis in breast cancer based solely in histological slides.

Amgad, M., *et al.* [51, 52]. assessed three variants of the TILs score:

1. Number of TILs / Stromal area
2. Number of TILs / Number of cells in stroma
3. Number of TILs / Total Number of cells

The results performed on the BCSS and NuCLS breast carcinoma datasets [24, 53] (the source datasets for TCGA part of TiGER dataset) showed the most prognostic TILs score to be the number of TILs divided by the total number of cells within the stromal region. A further breast cancer study [54] showed that the binarized tumor TILs infiltration fraction is predictive of survival, by analyzing the proportion of pixels in the image that were predicted as containing tumor as well as lymphocytes (number of pixels predicted as lymphocyte and tumor divided by the number of pixels predicted as tumor). Bai, Y., *et al.* [55] also found associations of clinical outcomes in breast cancer with TILs scores based on the number of TILs divided by the number of TILs and tumor cells detected.

The stromal TILs (sTILs) have been shown to have prognostic value in HER2+ breast cancer and TNBC [51]. sTIL density was found significantly prognostic for OS not only while applied on H&E slides but IHC as well [56]. Applied on the TCGA-BRCA mixed with non publicaly available dataset, Thagaard, J., *et al.* [57] tried to mimic the approach of the pathologist and therefore defined tumor-associated stroma. Tumor-associated stroma includes a margin of 250µm from the border of the tumor into the surrounding stroma. The sTIL density was calculated as the number of TILs within the tumor-associated stroma per mm^2. The patient cohort was then stratified into two groups: high and low sTIL density by using maximally selected rank statistics for cutpoint selection. As a result sTIL density stratified the patients significantly into two distinct prognostic groups. For continuous variables, the sTIL density was divided by 300 and higher sTILs scores were associated with significantly prolonged overall survival. For the TCGA-BRCA dataset, a further TIL score was found significant as the overlapping area between lymphocyte-dense regions and stromal regions divided by the size of the stromal regions [58]. Whereas a study, that focused on TNBC cases of TCGA, did not observe any differences in OS neither while using a continuous variable of manually annotated TILs (scored by a pathologist and partitioned into eight different groups, e.g. < 1%, 10-20%, etc.) nor after applying the log-rank test [59]. On the other hand, Fassler, D. J., *et al.* [60] confirmed correlation of intratumoral TIL infiltration with increased OS in breast cancer in the TCGA-BRCA cohort. TIL infiltrate percentage was calculated as the number of predicted patches that were classified as positive for tumor and lymphocyte divided by total number of cancer patches. Another used definition of sTILs was the percentage of tumor stroma area containing a lymphocytic infiltrate without direct contact with tumor cells [61]. Furthermore, studies found a three-scale grading system for reporting TILs status to be applicable, instead of continuous or binary grouped TILs densities [62]. More advanced TILs-based features such as the Ball-Hall Index of spatially connected TILs regions (clusters) also showed association with survival, particularly within the BRCA dataset of TCGA [63].

Hence, there is no canonic method for the automatic determination of TILs score based on the H&E breast cancer tissue samples but number of TILs per mm^2 of stromal area is used most frequently.

3 Methods

3.1 Semantic segmentation

3.1.1 DeepLab

One of the challenges in semantic segmentation using standard CNNs is that as the input feature map goes through the network it gets smaller and the information about objects of a smaller scale can be lost. DeepLab family introduces atrous convolutions that extract more dense features which help to preserve the object's information. Compared to standard convolutions, atrous convolutions have an additional parameter, atrous rate, which is the stride at which the input is sampled (Figure 3.1 a). The atrous convolution is used in the last few blocks on features that were extracted from the backbone network (e.g. ResNet [64]).

One of the latest models in this family, DeepLabv3 [35], applies several parallel atrous convolutions with different atrous rates (Atrous Spatial Pyramid Pooling, or ASPP, Figure 3.1 b) to effectively capture multi-scale information. Image-level features, or image pooling, are also applied to incorporate global context information. Those are calculated by applying global average pooling on the last feature map of the backbone. After applying all the operations in parallel, the results of each operation along the channel is concatenated and 1×1 convolution is applied to get the output. The addition of atrous convolutions allows the enlargement of the field of view without increasing the size of the filtering kernel, therefore no increase in the computation time.

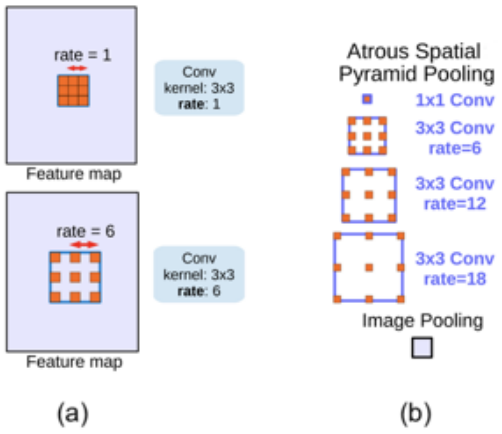
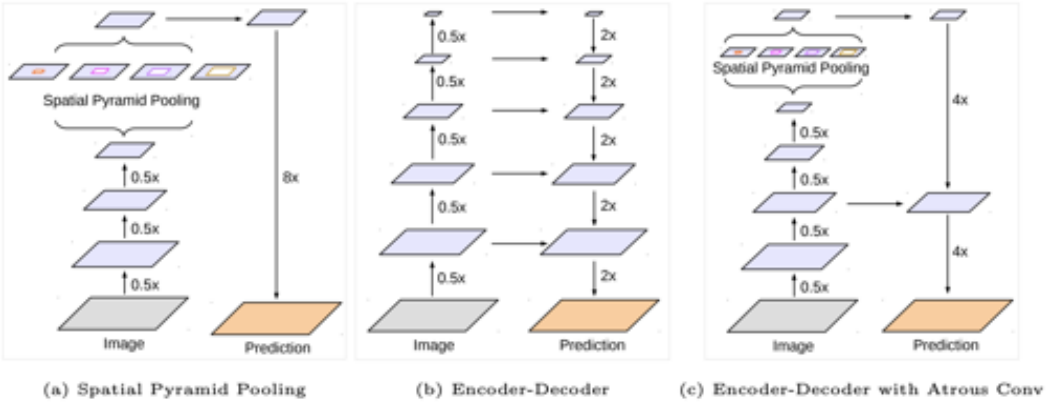


Figure 3.1: (a) Atrous convolution, (b) ASPP augmented with Image Pooling (or Image-level features). Figure taken from [35]

DeepLabv3+

The reproduction of shape contours during semantic image segmentation remained difficult with DeepLabv3 [36]. DeepLabv3 bilinearly upsamples the logits both during training and evaluation (Fig. 3.2 a), hence the improvements were made to employ the encoder-decoder structure (Figure 3.2) to avoid using a naive

decoder.



(a) Spatial Pyramid Pooling (b) Encoder-Decoder (c) Encoder-Decoder with Atrous Conv

Figure 3.2: The spatial pyramid pooling module of DeepLabv3 (a), the encoder-decoder structure (b) and DeepLabv3+ adaptation (c). Figure taken from [36]

DeepLabv3+ [36] adds the decoder module on top of the encoder output, as shown in Fig. 3.3. In the decoder module, the 1×1 convolution reduces the channels of the low-level feature map from the encoder module which is then concatenated with the DeepLabv3 feature map and the 3×3 convolution obtains sharper segmentation results. As a result, DeepLabv3+ holds rich semantic information from the encoder module, while the detailed object boundaries are recovered by the decoder module and the spatial information is retrieved.

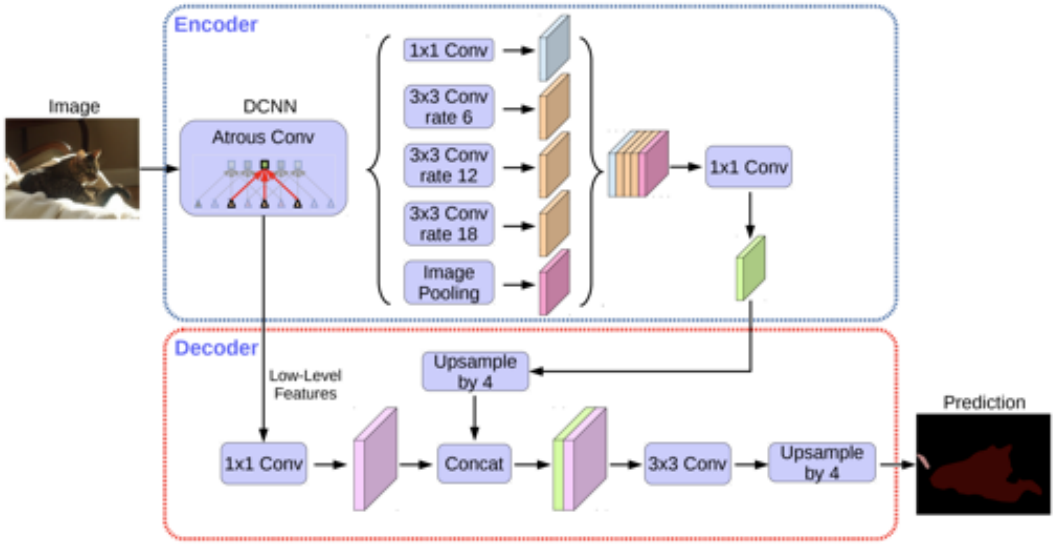


Figure 3.3: DeepLabv3+ architecture. DeepLabv3 as encoder and proposed decoder structure for semantic image segmentation. Figure taken from [36]

3.1.2 Transformers

Transformers [65] were originally designed for the neural machine translation problem in NLP to capture long-range dependencies among words in a sentence. Their architecture converts one sequence into another one based on encoder-decoder architecture, but it differs from the previously existing sequence-to-sequence models because it does not imply any Recurrent Networks.

The input and output are first embedded into an n-dimensional space. Since the network and the self-attention are permutation invariant, the positional encoding is added to create a representation of the position of the word in the sentence. The following modules consist mainly of Multi-Head Attention and Feed Forward layers. Encoder (Figure 3.4, left) and decoder (Figure 3.4, right) are composed of those modules that can be stacked on top of each other $N\times$ times.

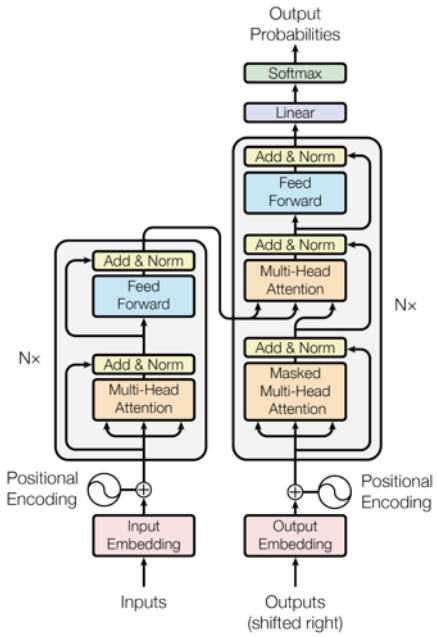


Figure 3.4: Transformer model architecture. Figure taken from [65]

Self-attention is a sequence-to-sequence operation. It takes a weighted average over all the input vectors using dot product. Scaled Dot-Product Attention (Figure 3.5, left) can be described by the following equation:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (3.1)$$

where in the context of the translation problem, Q is a matrix of vector representation of one word in the sequence, K contains vector representations of all the words in the sequence and V contains again the vector representations of all the words in the sequence. For the multi-head attention modules in the encoder and decoder, V consists of the same word sequence as Q. However, for the attention module that is taken into account, the encoder and the decoder sequences, V, and Q are different. Q, K, and V matrices are used to calculate the attention scores. These scores measure how much attention needs to be placed on words of the input sequence with respect to a word at a certain position. The scaling factor $\sqrt{d_k}$ is applied to avoid large values that after applying softmax would lead to vanishing gradients.

While Scaled Dot-Product Attention focuses on the whole sentence, Multi-Head Attention approaches different segments of the words. The word vectors are divided into a fixed number (number of heads) of parts, and then within Multi-Head Attention (Figure 3.5, right) the attention mechanism is repeated multiple times on those separate parts with linear projections of Q, K, and V. Since the Feed-Forward layer is expecting just one matrix, a vector for each word, the outputs are linearly concatenated. This allows the system to learn from different representations of Q, K, and V.

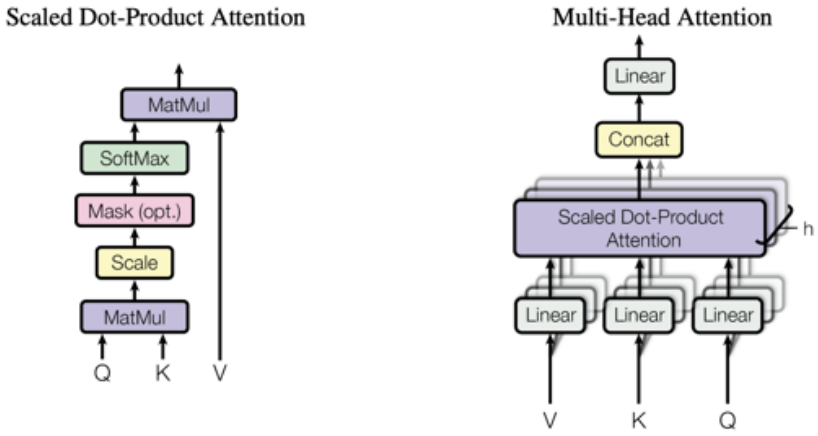


Figure 3.5: Scaled Dot-Product Attention (left). Multi-Head Attention consists of several attention layers running in parallel (right). Figure taken from [65]

To add element-wise non-linearity transformation of incoming vectors, the transformer includes feed-forward networks. It processes the output from one attention layer so that it fits better for the next attention layer. Each of the layers in the encoder and decoder contains a fully connected feed-forward network, which is applied to each position separately and identically. These feed-forward layers can be described as a separate, identical linear transformation of each element from the given sequence.

Naive application of the transformers approach into the image domain would require evaluation of relations between each pixel and every other pixel, which is obviously not scalable. The Visual transformer (ViT) [66] is the first work to prove that a pure Transformer can achieve state-of-the-art performance in image classification. ViT converts the input image into a 1D series by cutting it into patches and feeding it to a linear layer. It yields a patch embedding. Position embeddings are added to the image patch embeddings. Adding the learnable position embeddings to each patch allows the model to learn the structure of the image. The rest of the pipeline is a standard encoder and decoder blocks of the transformer. The decoder learns to map patch-level encodings coming from the encoder to patch-level class scores. Next, these patch-level class scores are upsampled by bilinear interpolation to pixel-level scores.

SegFormer

SegFormer [48] is a positional-encoding-free transformer based semantic segmentation method. As depicted in Figure 3.6, it consists of two main modules: a hierarchical Transformer encoder to generate high-resolution coarse features and low-resolution fine features, and a lightweight All-MLP decoder to fuse these multi-level features and produce the final semantic segmentation mask.

The $H \times W \times 3$ input image is forwarded to the hierarchical Transformer encoder to obtain multi-level features at $\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$ resolution after passing through four transformer blocks.

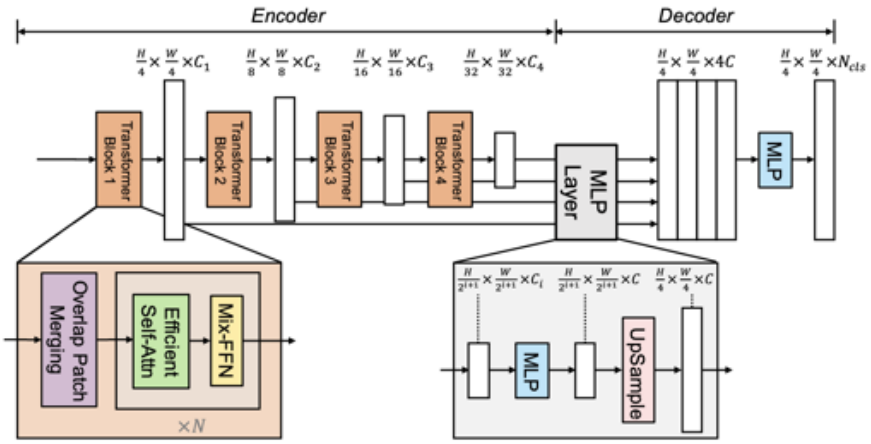


Figure 3.6: SegFormer consists of two main modules: A hierarchical Transformer encoder to extract coarse and fine features; and a lightweight All-MLP decoder to directly fuse these multi-level features and predict the semantic segmentation mask. "FFN" indicates feed-forward network. (modified Figure from [48] according to the official implementation)

Each transformer block consists of three modules: Overlap Patch Merging, and classical transformer building blocks: Self-Attention and Feed-forward network.

The standard transformer receives input as a 1D sequence (such as word embeddings in the previous chapter 3.1.2). To handle images, those need to be reshaped into a sequence of flattened 2D patches. Overlapped Patch Merging produces features given an image and parameters: patch size K, stride between two adjacent patches S, and padding size P. In the original paper [48] those are set to $K = 7$, $S = 4$ and $P = 3$. Therefore the input is split into fixed-size patches, which then go through a linear projection. The result is a hierarchical feature map F_i with a resolution $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$ where $i \in \{1, 2, 3, 4\}$ and C_{i+1} is larger than C_i. By performing this with overlapped patches SegFormer aims to preserve the local continuity around those patches.

The main computation bottleneck of each transformer block in encoder is the self-attention layer. In SegFormer, before applying the self-attention according to the formula 3.1, the sequence K is reduced by ratio R:

$$\hat{K} = Reshape(\frac{N}{R}, C \cdot R)(K)$$

$$K = Linear(C \cdot R, C)(\hat{K})$$

where $N = H \times W$, $Reshape(\frac{N}{R}, C \cdot R)(K)$ refers to reshaping K to the the shape of $\frac{N}{R} \times (C \cdot R)$, and $Linear(C \cdot R, C)(\hat{K})$ refers to a linear layer taking a $(C \cdot R)$-dimensional tensor as input and generating a C-dimensional tensor as output. Therefore, the new K has dimensions $\frac{N}{R} \times C$. In original experiments, R was set to $[64, 16, 4, 1]$ from stage-1 to stage-4 and resulted in a reduction of the complexity of the self-attention mechanism.

Mix-FFN (feed-forward network) can be formulated as:

$$x_{out} = MLP(GELU(Conv3 \times 3(MLP(x_{in})))) + x_{in}$$

where x_{in} is the feature from the self-attention module. By using 3×3 convolution and zero padding in a feed-forward network SegFormer aims to leak pixel location information since it is a positional-encoding-free method.

The multi-level features are then passed to All-MLP decoder to predict the segmentation mask at $\frac{H}{4} \times \frac{W}{4} \times N_{cls}$ resolution, where N_{cls} is the number of classes. The proposed All-MLP decoder consists of four main steps. First, multi-level features from the encoder go through an MLP layer to unify the channel dimension (3.2). Then, features are up-sampled to $\frac{1}{4}$th of the original image (3.3). Third, an MLP layer is adopted to fuse the concatenated features (3.4). Finally, another MLP layer takes the fused feature to predict the segmentation mask (3.5).

$$\hat{F}_i = MLP(C_i, C)(F_i), \forall i \tag{3.2}$$

$$\hat{F}_i = Upsample(\frac{H}{4} \times \frac{W}{4})(\hat{F}_i), \forall i \tag{3.3}$$

$$F = MLP(4C, C)(MLP(\hat{F}_i)), \forall i \tag{3.4}$$

$$M = MLP(C, N_{cls})(F) \tag{3.5}$$

where F_i is the the feature and M is the final mask.

3.1.3 TILs segmentation postprocessing

As tissue segmentation, TILs detection was approached as a semantic segmentation problem with detection bounding boxes transformed into pseudo-segmentation masks (further details in Data section 4.1). TILs segmentation results were first processed with non-maximum suppression to extract local maxima, then obtain point annotation of each TIL, and compare with ground truth annotations by finding a match with the Hungarian Algorithm.

Non-maximum Suppression (NMS)

Non-maximum Suppression is a class of algorithms to select one entity out of many overlapping entities. In terms of detection models, used to find the best-fitting bounding box out of all predicted bounding boxes for the same object [67]. Each proposal comes with a confidence score. One by one the bounding box with the highest score is selected to keep and compared to all bounding boxes by calculating Intersection Over Union (IoU). If a comparison scores the IoU higher than some defined threshold, those bounding boxes out of the pool are eliminated. The process is repeated by picking the highest confidence bounding box out of the remaining pool of proposals until there are any.

In terms of TILs segmentation, there are no proposed bounding boxes and no scores. Nonetheless, to obtain good centroids for each TILs prediction, the segmentation posteriors need to be condensed to local maxima. The NMS application will in this case appropriately filter the segmentation posteriors with a kappa threshold and the segmentation borders refined to get a clear TIL region annotation with a suitable kernel size used for dilation.

Hungarian algorithm

To evaluate the match of predicted TILs and ground truth annotations, it is not enough to compare the coordinates. Each TIL should be annotated with a point annotation of one pixel, it is aimed to be placed in the center but really placed in the center of a predicted TIL area, which is not perfect. The goal is to find the pairs of coordinates that indicate the same TIL object. Matching between the ground truth set and predictions can be solved as an assignment problem with the Hungarian algorithm [68]. It is a combinatorial optimization algorithm that solves the problem in polynomial time complexity. The distance matrix $n \times m$ between all annotated (n) versus predicted (m) TILs is the cost matrix of each of the prediction coordinates to match any of the ground truth coordinates. The goal is to assign the ground truth coordinates to the prediction coordinates to minimize the total distance. The algorithm can be introduced as step-by-step instruction:

- Step 1: For each row, subtract the smallest element of the row from each of its elements.
- Step 2: For each column, subtract the smallest element of the column from each of its elements.
- Step 3: Cover all zeros with a minimum number of lines

  In the resulting matrix cover all zeros using a minimum number of horizontal and vertical lines. If there are n lines, Step 5.
- Step 4: Find the smallest element that is not covered by a line in Step 3. Subtract it from all uncovered elements, and add it to all elements that are covered twice. Step 3.
- Step 5: Assignment pairs are indicated by the positions of the zeros in the cost matrix.

As a result, the assignment pairs include the matching of ground truth TILs coordinates to the prediction coordinates that minimize the total distance. If desired, the pairs can be filtered by minimal allowed distance. This work uses the `scipy.optimize.linear_sum_assignment` implementation of the Hungarian algorithm.

3.2 Survival Analysis

The overall survival (OS) is the primary endpoint for prognostic analysis in this thesis, hence time to the event (death) is of interest. Survival data are generally described and modeled in terms of two related probabilities, namely survival and hazard. [69] This thesis focuses on non-parametric models to avoid making any additional assumptions about the distributions. Throughout the analysis, the python library `lifelines` [70] was used. The survival probability

$S(t)$ is the probability that an individual survives from the time origin (in our case diagnosis of breast cancer) to a specified future time t. It can be denoted as:

$$S(t) = Pr(T > t) = 1 - F(t) = \int_t^\infty f(x)dx = \text{ Probability of surviving past time } t$$

where T is a random variable that indicates the time until the event of interest (death). $F(t)$ and $f(t)$ are the cumulative distribution function and probability density function of T. The hazard is the probability that an individual who is under observation at a time t has an event at that time:

$$h(t) = lim_{\delta t \to 0} \frac{Pr(t \leq T \leq t + \delta t | T > t)}{\delta t} = \frac{f(t)}{1 - F(t)}$$

In contrast to the survival function, which focuses on not having an event, the hazard function focuses on the event occurring. So if hazard probability describes the intensity of death [71] at the time t given that the individual has already survived past time t, then the cumulative hazard is the cumulative amount of hazard up to time t. The cumulative hazard $H(t)$, defined as the integral of the hazard, can be calculated using the survival probability with help of the Laplace transform:

$$H(t) = \int_0^t h(x)dx = \int_0^t \frac{f(x)}{1 - F(x)}dx = -ln(1 - F(t)) = -\log(S(t))$$

The cumulative hazard can be interpreted as the number of events that would be expected for each individual by time t if the event was a repeatable process. [69]

3.2.1 Kaplan–Meier estimator

The survival probability can be estimated nonparametrically from observed survival times, both censored and uncensored, using the Kaplan–Meier method. The estimated probability of surviving past time t is calculated as:

$$\hat{S}(t) = \prod_{i; t_i \leq t} (1 - \frac{d_i}{n_i})$$

where n_i is the number of patients alive before t_i (and not censored) and d_i is the number of observed events at t_i. $t_0 = 0$ and $S(0) = 1$. The estimated probability is a step function that changes value only at the time of an event. To characterize the survival in a homogeneous group often the empirical survival function is visualized with Kaplan–Meier plot.

3.2.2 Cox model

Additionally to the event time, there is often access to other covariates of individuals (e.g. age, gender, BMI, etc.). Often the goal is to understand how the covariates affect the survival function of the event. [71] Let C denote those covariates. The conditional survival function can be formulated as followed:

$$S(t|c) = Pr(T > t | C = c) = \text{ Probability of surviving past time } t \text{ given } c$$

Hence, the conditional hazard function and conditional cumulative hazard are:

$$H(t|c) = -\log(S(t|c)), \text{ hence } h(t|c) = -\frac{\partial \log S(t|c)}{\partial t}$$

The Cox proportional hazard model models the hazard function $h(t|C = c)$ as:

$$h(t|C = c) = h_0(t)\exp(c^T\beta)$$

where β is the vector of coefficients for each of the covariates and $h_0(t)$ is the baseline hazard function. The hazard ratio, or ***risk***, is the exponential of β_i value $\eta_i = \exp(\beta_i)$ and the baseline hazard describes how the risk of event per time unit changes over time at baseline levels of covariates. The Cox model assumes that the covariates have a linear multiplication effect on the hazard function and the effect stays the same over time.

$$\frac{h(t|c_i)}{h(t|c_j)} = \frac{h_0(t)\exp(c_i^T\beta)}{h_0(t)\exp(c_j^T\beta)} = \frac{\exp(c_i^T\beta)}{\exp(c_j^T\beta)} = \exp((c_i - c_j)^T\beta)$$

The ratio of the hazard function between two individuals with different covariates c_i and c_j is a constant over time since $h_0(t)$ was canceled out. Hence the name, proportional hazard model. The conditional hazard function is:

$$H(t|c) = \exp(c^T\beta)\int_0^t h_0(s)ds = \exp(c^T\beta)H_0(t)$$

It yields a conditional survival function:

$$S(t|c) = \exp(-H(t|c)) = \exp(-\exp(c^T\beta)H_0(t)) = \exp(-H_0(t))^{\exp(c^T\beta)} = S_0(t)^{\exp(c^T\beta)}$$

Estimation of the parameter β is often done by maximizing the partial likelihood function $\hat{\beta}_n = argmax_\beta \hat{L}_n(\beta)$, where:

$$\hat{L}(\beta) = \prod_{i=1}^{n} \frac{h(T_i|C_i)}{\sum_{j:T_j \geq T_i} h(T_j|C_j)} = \prod_{i=1}^{n} \frac{\exp(C_i^T\beta)}{\sum_{j:T_j \geq T_i} \exp(C_j^T\beta)}$$

A positive sign of β_i indicates a higher risk of an event, hence the probability for the event for that particular subject is higher. Likewise for a negative signed β_i, lower risk, and lower probability. The actual value of β_i plays a role as well. Values less than one will reduce the hazard and values greater than one, increase it.

A model's accuracy can be quantified based on concordance. [72] It is a measure of the rank correlation between predicted risks and observed time points. It is defined as the ratio of correctly ordered (concordant) patient pairs to all concordant and discordant patient pairs. Let i, j be a patient pair. If a model predicts a higher risk for the first patient ($\eta_i > \eta_j$), for it to be a concordant pair first patient should have a shorter survival time in comparison with the other patient ($T_i < T_j$) and similarly if lower risk then longer survival time, $\eta_i < \eta_j$ & $T_i > T_j$. If both patients are censored the pair is discarded. If only one patient is censored, the pair is not discarded only if the other patient experienced the event before the censoring time. By

construction, concordance must be between 0 and 1, with 1 representing the perfect agreement between model and observation and 0.5 representing random guesses.

Additionally, to estimate the goodness-of-fit the p-value is determined. The Wald test is typically used to evaluate the significance of a variable in the model estimated with the maximum likelihood function. The null hypothesis is that the model does not fit the data well. The Wald statistic tests, whether β_i coefficient is statistically significantly different from 0 and is defined as:

$$W = \frac{(\hat{\beta}_n - \beta_0)^2}{var(\hat{\beta}_n)}$$

If the true coefficient was β_0, then the sampling distribution of the Wald test statistic should be approximate $\mathcal{N}(0,1)$. The p-value gives the probability of observing a test statistic as extreme as the one observed if the sampling distribution was $\mathcal{N}(0,1)$. If the p-value is small, the observed test statistic is very unlikely under the null hypothesis. And the significance level of 0.05 indicates that there is a 5% risk of being wrong by concluding that the model fits the data well when it doesn't.

4 Datasets used

To give an overview of the data used as well as extend the scheme presented in the Introduction chapter (Figure 1.1), refer to Figure 4.1. The following sections provide more details about the data.

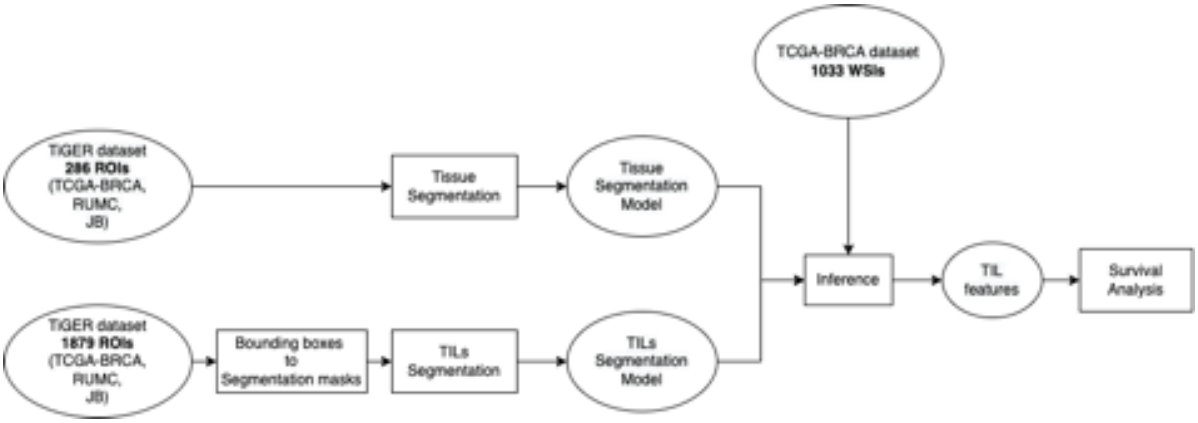


Figure 4.1: Workflow in form of a Petri net showing completed tasks in this thesis complemented with the data sources and counts. For training, validation, and testing of segmentation models only TiGER challenge data was used, originating from three medical institutes (TCGA-BRCA, RUMC, and JB). Calculation of TIL features was performed on TCGA-BRCA diagnostic slides.

4.1 Segmentation

The data for tissue and TILs segmentation originates from publicly available Tumor Infiltrating lymphocytes in breast cancER (TiGER) challenge dataset containing digital pathology images of Her2 positive (Her2+) and Triple Negative (TNBC) breast cancer whole-slide images (WSIs), regions of interest (ROIs), and manual annotations. More specifically, the WSIROIS dataset was used for model training, validation, and testing (see Table 4.1). It includes 195 WSIs of breast cancer core-needle biopsies and surgical resections with pre-selected ROIs and manual annotations. TiGER data, both at WSI and ROI level, was released at a spacing (pixel size) of approximately 0.5 µm/px (at 20× magnification), more information is available on the original challenge website[1]. The TiGER tissue annotations include eight labels that were reduced to three (see Table 4.2). The training masks were generated using available

[1] `https://tiger.grand-challenge.org/Data/`

Source	Tissue			TILs		
	slides	ROIs	median ROI size #pixels [k]	slides	ROIs	median ROI size #pixels [k]
TCGA-BRCA	151	151	4 983	124	1744	20
RUMC	26	81	1 312	26	81	1 312
JB	18	54	1 465	18	54	1 465
	195	286		168	1879	

Table 4.1: TiGER data overview. Sources: Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA), Radboud University Medical Center (RUMC) and Jules Bordet Institute (JB). Tissue slides and ROIs refer to the segmentation images and annotations whereas TILs prefix specifies the data for TILs detection provided by the challenge. TCGA-BRCA dataset statistics showcase a variation in ROIs sizes for tissue and TILs tasks as well as compared to two other datasets.

XML files. Within provided mask images, in certain cases, regions not included in ROIs and non-annotated regions were marked with the same label, which could not be directly used for training (see ground truth in Figure 5.5). While for tissue segmentation the images and their

TiGER Tissue Label	Share	ID	new ID	new Tissue Label
Invasive tumor	0.283	1	1	Tumor
In-situ tumor	0.029	3	1	Tumor
Tumor-associated stroma	0.286	2	2	Stroma
Inflamed stroma	0.096	6	2	Stroma
Necrosis not in-situ	0.048	5	0	Rest
Healthy glands	0.0008	4	0	Rest
Background	0.231	0	0	Rest
Rest	0.026	7	0	Rest

Table 4.2: Reduction of labels provided in TiGER challenge dataset. Resulting labels include three classes: Tumor (1), Stroma (2) and Rest (0) with shares of 0.312, 0.382 and 0.306. Shares were calculated by dividing the number of pixels belonging to some label by the number of the pixel in the current image and averaged over all images.

masks could be used directly as extracted from the dataset, the data for TILs segmentation required some preprocessing. The TiGER fixed-size bounding box annotation for lymphocytes and plasma cells (see Table 4.3) was adapted for segmentation by transforming each bounding box into an annotation of the center pixel with a dilatation of three.

For the inferences the breast cancer TCGA-BRCA dataset was used, generated by the TCGA Research Network: https://www.cancer.gov/tcga. While there are a large number of images available in The Cancer Genome Atlas (TCGA), only the diagnostic slide were downloaded (Formalin-Fixed Paraffin-Embedded (FFPE) slides). FFPE slides are the gold standard in diagnostic medicine [73]. They are prepared by fixing a specimen in formaldehyde and then

Source				Number of cells per ROI		
	slides	ROIs	cells	min	max	median
TCGA-BRCA	124	1 744	19 115	0 (44.3%)	206	1
RUMC	26	81	4 728	0 (7.4%)	657	19
JB	18	54	5 523	0 (7.4%)	608	51.5
	168	1 879	29 366			

Table 4.3: Data overview for TILs detection. Sources: Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA), Radboud University Medical Center (RUMC) and Jules Bordet Institute (JB). Number of cells here refers to the number of bounding boxes that were assigned for lymphocytes and plasma cells, further named TILs. Increased number of ROIs compared to tissue segmentation task, only due to TCGA-BRCA dataset and its considerably smaller ROIs as summarized in Table 4.1.

placing it in a paraffin block to section it. After generating a manifest file on TCGA dataset website, which includes 1133 diagnostic slides of BRCA breast cancer patients, the files were downloaded with the GDC Data Transfer Tool (`gdc-client` with `--manifest` option).

4.2 Survival Analysis

The TiGER challenge aims to assess the prognostic significance of computer-generated TILs scores for predicting survival by applying the Cox proportional hazards model. The survival analysis was done using a large independent dataset that includes cases from both clinical routine and from a phase 3 clinical trial. 200 WSIs and their clinicopathological variables, including recurrence and survival data were used as the experimental set, and 707 cases for the test set, both not directly accessible to participants.

The survival analysis within this thesis is done exclusively on publicly available TCGA-BRCA data. The first twelve characters of the TCGA barcode identifiers are saved under `case_submitter_id` as a unique patient identifier. In the dataset, death (`vital_status` = 1) is considered as an event, and the time until the event or censoring is either `days_to_death` (number of days to death from the first diagnosis) or `days_to_followup` (number of days to last follow-up from the first diagnosis). The patients that had missing values, that are essential for survival analysis (`case_submitter_id`, `vital_status` and time to event) were filtered out. The resulting dataset includes clinical variables for 1076 patients with 150 events. Originally the time is recorded in days, but it was converted to years (by simply dividing by 365), due to the fact that the mean duration until censoring is 3.3 years and 4.4 years until death, and the maximum duration until censoring reaches 23.6 years and 20.4 years until death. The final number of patients that are included in the survival analysis and have corresponding diagnostic slide(s) is 1015.

Additionally, a further file (nationwidechildrens.org_clinical_patient_brca.txt) was downloaded from the TCGA portal. It includes Her2, estrogen, and progesterone receptor levels measured in a primary tumor or metastases with immunohistochemistry. 115 TNBC and 159

Her2+ patients were identified, by filtering columns `er_status_by_ihc`, `pr_status_by_ihc` and `her2_status_by_ihc`. This dataset also includes the earlier mentioned columns that are essential for survival analysis, but the durations and vital statuses are not kept up-to-date, thus the previous survival data was used with extracted patient identifiers depending on the receptor levels.

5 Results & Discussion

5.1 Tissue Segmentation

Three models were trained to segment the RGB input of H&E stained image at 20× magnification (resolution of 0.5 micron-per-pixel) into three prediction maps: tumor, stroma, and rest (not white space, but tissue that is neither tumor nor stroma, for details refer to Table 4.2). The data was separated on the patient level (or slide level, since there is one slide per patient present) into training, validation, and test with approximately 80%, 10%, and 10% rates accordingly. In order to keep the distribution of the resulting patch numbers (see Table 5.2) and dataset sources fair, the patient separation in Table 5.1 was introduced.

	Train	Validation	Test
TCGA-BRCA	120	16	15
RUMC	20	3	3
JB	16	1	1
	156	20	19

Table 5.1: Split of patients across different medical sources into train, validation, and test sets for segmentation tasks. That resulted an 80%, 10%, and 10% split of patches for training, validation and testing, as showed in Table 5.2.

	slides	ROIs	patches	Number of patches that include					
				Tumor	Stroma	Rest	1 class	2 classes	3 classes
Train	156	228	220 567	154 734	172 046	107 506	63 919	99 577	57 071
				(70%)	(78%)	(49%)	(29%)	(45%)	(26%)
Validation	20	25	29 465	16 884	22 187	13 954	11 171	13 028	5 266
				(57%)	(75%)	(47%)	(38%)	(44%)	(18%)
Test	19	33	30 248	18 194	25 630	14 787	8 548	15 037	6 663
				(60%)	(85%)	(49%)	(28%)	(50%)	(22%)
	195	286	280 280						

Table 5.2: Overview of patches that were split into train, validation, and test sets for tissue segmentation. The percentages indicate the fraction of a specific conditioned group of patches to the number of all patches in the "patches" column.

The patches were created using a sliding window approach with 256×256 sized patches and stride equals 128. The additional rotation augmentation was applied, by rotating each

patch 5 times at 9 degrees each.

The models were trained with the mmsegmentation toolbox [74]. All models were trained for 160K iterations with Cross Entropy Loss and standard data augmentation techniques including resizing at a random sample scale in the range of (0.5, 2.0), cropping with the maximum 0.75 of a single category present, flipping with 0.5 probability, and application of photometric distortion which includes 0.5 probability for each of the following transformations: random brightness, contrast, saturation, hue and color adjustments. The DeepLabv3+ model was taken as a baseline and trained with ResNet50 backbone, Adam optimizer, learning rate equals 10^{-4} and batch size of 64. Whereas the transformer-based SegFormer-B5 (further referred to as SegFormer) architecture was trained once with the same setup of Adam optimizer, 10^{-4} learning rate and batch size of 64, and additionally, as in original paper [48], using AdamW optimizer, the learning rate set to an initial value of $6 \cdot 10^{-5}$ and then used a poly learning rate schedule with factor 1.0 by default.

Model	FLOPs	Params	Iterations	Runtime	mDice			
					Overall	Tumor	Stroma	Rest
DeepLabv3+	44.16	43.58 M	160 K	1d 12h 2m	85.25	85.13	88.07	83.66
SegFormer	12.96	81.97 M	160 K	3d 4h 30m	83.44	83.83	86.32	80.16
SegFormer, AdamW	12.96	81.97 M	160 K	3d 4h 25m	84.93	85.40	87.40	82.00

Table 5.3: Overview of the trained tissue segmentation models. The runtime is given for a training on one GPU NVIDIA A100 SXM4. Describes comparable performance, but severe runtime difference of SegFormer-based methods compared to the DeepLabv3+ complemented by doubled number of parameters.

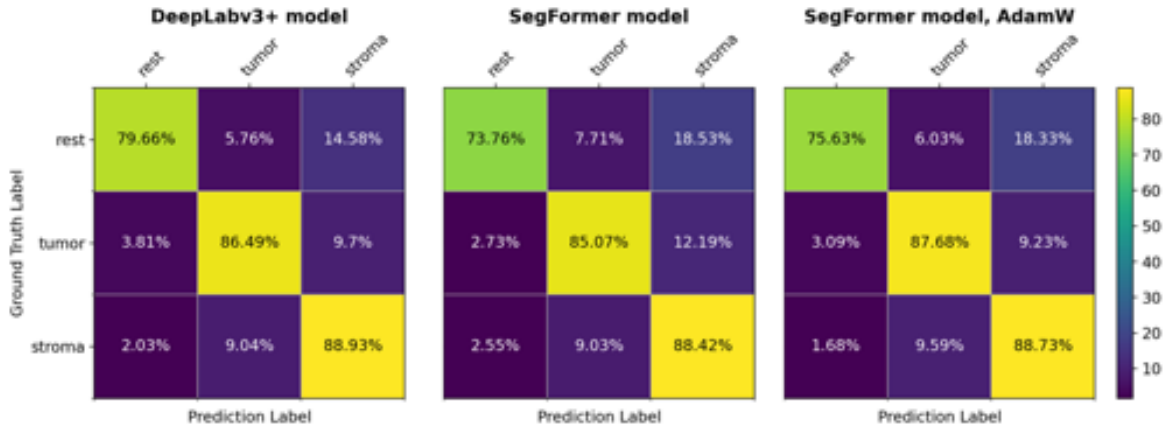


Figure 5.1: Confusion matrices on pixel level for DeepLabv3+, SegFormer and SegFormer with AdamW optimizer based on test set of 32 ROIs. Across all models, the performance is comparable but there is some tendency to misinterpret rest pixels for stroma.

During a close investigation of test data, one slide (with the prefix TCGA-OL-A5RW-01Z-00-DX1) was excluded from the test set due to an obvious image-mask mismatch. The overall performance of the models, the number of parameters, and the resulting Dice score after testing on 32 test images can be found in Table 5.3. The first thing that catches the eye is the severe runtime difference of SegFormer-based methods compared to the DeepLabv3+ accompanied by doubled number of parameters. None of the SegFormer approaches outperform the DeepLabv3+, but the performance is comparable, which can be also observed in the confusion matrices in Figure 5.1. Both SegFormer-based methods show difficulty correctly segmenting rest regions, whereas SegFormer AdamW slightly outperforms DeepLabv3+ in the number of true positive detected tumor pixels.

As previously mentioned, the dataset originates from three medical institutions which make it reasonable to characterize the performance separately. The boxplots in Figure 5.2 indicate that the performance of the DeepLabv3+ and SegFormer AdamW models in TCGA-BRCA and JB groups are fairly invariant. Whereas the RUMC group accounts not only for the lower performance of SegFormer-based methods in segmenting rest regions but also for an improvement of tumor region segmentation by SegFormer AdamW model both observed in Figure 5.1.

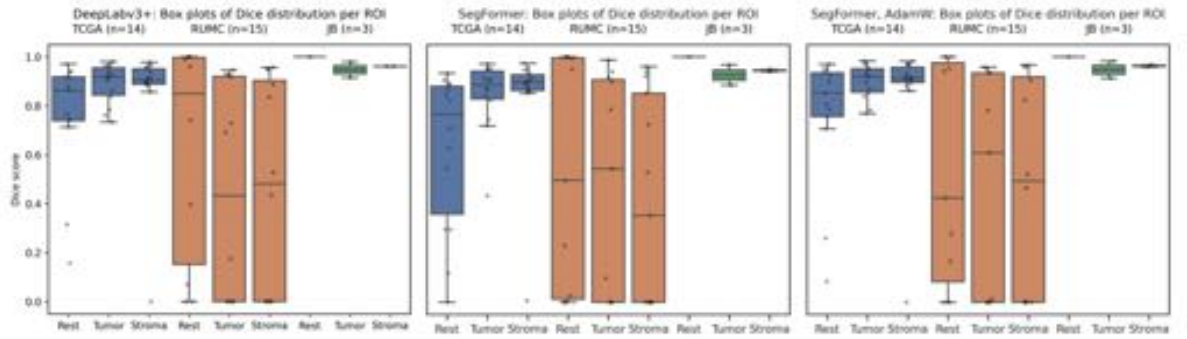


Figure 5.2: Boxplots of pixel-wise calculated Dice score across three datasets (TCGA-BRCA, RUMC, JB) and three segmentation labels. Each boxplot represents the results of one model. Boxplot color helps to distinguish the Dice scores between different datasets and finally, each dataset is represented by three boxes for each of the class labels - rest, tumor, and stroma.

The additional specialty that Figure 5.2 brings to light is the considerable number of RUMC ROIs that have been evaluated with Dice scores close to zero across all models. Due to the nature of the Dice score, those can be originated from significant numbers of either false positives, false negatives, or both. Additionally, not only true positives need to be taken into account, but the false negatives as well, for instance in case of a low number of true TILs in the image. Hence further exploration is needed. According to the boxplots of precision and recall in Figure 5.4 the precision across all models has more close zero values that indicate more frequent false positives. There are clear examples, such as Figure 5.3, where the ground truth includes exclusively rest but all trained models provide multiple class predicitons. Even

though there are also opposite examples of regions that were solely annotated as rest and predicted as such (which then lead to occasional Dice, precision, and recall equal to one, see Figure 5.2 and 5.4), the issue of false positives needed to be further explored.

Out of the original 81 RUMC ROIs, 14% of ROIs carry the annotation of only class label 7 - rest. According to the organizers, that class contains regions of several tissue compartments that are not specifically annotated in the other categories, such as healthy stroma, erythrocytes, adipose tissue, skin, nipple, etc. There are none of such annotations in TCGA-BRCA or JB datasets. After forming new masks with only three classes, the number of RUMC ROIs annotated completely as rest grew to 17% due to an additional ROI that represented only necrosis not in-situ. Even though the number of all-rest-ROIs in the new TCGA-BRCA is 0.2% and even 22% in JB, RUMC ROIs originally annotated as rest (label 7) need to be minded. Those ROIs may not be the consequence of a bad annotation, but in future experiments, it should be addressed by assigning them a reduced sampling rate, lower weights or excluding those completely.

The patch size and the sampling stride define the overlap between consecutively extracted patches from the WSI or ROI image. The stitching of the segmented patches introduced tiling artifacts visible in Figure 5.3 and 5.5. Since DeepLabv3+ and transformer-based networks experience it, the reason is the training setup. A probable reason is overfitting due to small patches of 256×256 and 50% overlap used for training and validation. Hence the problem can be tackled by applying less spacing during patch formation and additional augmentations. During inference, Khened, M *et al.* [75] addressed a similar problem by increasing patch size by a factor of 4 (from 256×256 to 1024×1024) and keeping 50% overlap. Due to time constraints and difficulties in migrating transformer based models to accept generalized input, this experiment was not further pursued.

A close look at the prediction also revealed that at some cases Dice scores might suffer due to some inaccuracies in the annotations. Figure 5.5 showcases that all models were penalized for detecting a rest region inside of the tumor, which was probably learned with some dependency to the presence of white space, which also present in the same slide (the bubles in the lower part of the image) and was annotated as rest.

Nonetheless, there are positive segmentation results present, such as JB ROIs depicted in Figure 5.6 and 5.7 where the performance of SegFormer AdamW is either very close or slightly better then DeepLabv3+. Curiously, in Figure 5.7 the SegFromer AdamW performs slightly better partly because this model, in contrast to the rest, did not attempt to annotate the small possibly tumorous region in the upper left corner (it is not annotated in ground truth but visually highly similar to the tumor regions in this example). A further experience could be conducted by filtering out those small island areas that contributed to a better result in this case, and evaluating the performance without them. But due to time constraints, this experiment was not executed. To decide whether this is an encouraging model behavior the pathologist must be consulted, but in some cases of larger ROIs, SegFromer AdamW manages to outperform DeepLabv3+, by predicting more smooth and visually coherent prediction maps, as visualized in Figure 5.8.

For tissue segmentation, the TiGER challenge evaluated the Dice score for stroma, i.e.

tumor-associated stroma grouped with inflamed stroma versus all other classes and invasive tumor against all other classes (see Table 4.2). The motivation for such a metric is that regions of invasive tumors and tumor-associated stroma play a central role in the definition of the TILs score. The resulting values are registered in Table 5.4. The final TiGER best result was achieved by a segmentation model that used UperNet [76] with visual attention network [77] as the backbone, hence there are some parallels to both of the trained models: use of hierarchical representation and attention. The TiGER challenge results are based on an experimental set of 26 WSIs and a final test of 38 WSIs and different class definitions (current tumor class includes extra in-situ tumor, see Table 4.2), the results in Table 5.4 can not be directly compared. It prohibits making the statement that trained models in this work outperform the models developed within the TiGER challenge, but their results make them promising for further experiments to acquire real comparison.

	DeepLabv3+	SegFormer	SegFormer, AdamW	TiGER best (experimental)	TiGER best (final)
Dice score	0.866	0.851	0.864	0.794	0.812

Table 5.4: Dice score for stroma compared between the TiGER challenge leaderboard results [2] versus three models developed in this work. Pixel-wise Dice score was calculated for stroma and tumor regions. The results were obtained from different data.

The observation of similar performance between the developed methods complies with the literature. As Dosovitskiy, A *et al.* [66] witnessed, transformer based models yield modest accuracies of a few percentage points below ResNets of comparable size. It is claimed that this outcome is expected, since Transformers lack some of the inductive biases inherent to CNNs, such as translation equivariance and locality, and therefore do not generalize well when trained on insufficient amounts of data. Liu, Y at al. [78] came to a similar conclusion and showed that the performance of visual transformers largely varies when trained with small and medium datasets. A typical fine-tuning scenario that results in major improvements in the performance is pre-training the model on a big dataset (e.g., ImageNet) which is also done in SegFromer paper [48] and repeated strong data augmentations [79]. An alternative is the use of a principled optimizer to avoid excessive demands of data, computing, and hyperparameters tuning [80].The importance of the optimizer choice is also seen in the presented experiments of SegFromer with Adam and AdamW optimizers. Adam and AdamW were selected according to literature reporting that these adaptive gradient methods do not underperform momentum or gradient descent optimisers [81]. AdamW improves regularization compared to Adam optimizer by decoupling the weight decay from the gradient-based update [82]. The evidence that AdamW is a better choice, agrees with the prior works that use AdamW to optimize visual transformer models from random initialization [83]. Generally using AdamW to optimize deep transformers is extensively practiced in the community [84].

Due to time constraints, there were no further experiments to improve the performance

of transformer based models, which could involve the previously noticed pre-training on ImageNet, increasing the number of argumentations, or applying different optimizers. Hence, while the SegFormer AdamW model remains promising, due to overall better performance and runtime, this thesis will use the DeepLabv3+ model for further inferences and analysis. For model inference, the model with the best pixel-wise Dice score of all three regions on the validation set was used. Patches of size 256×256 were extracted from the tissue region at 20× magnification with a stride of 184 (for faster runtime). The whitespace was extracted by using thresholding and inference of non-background pixels was then performed.

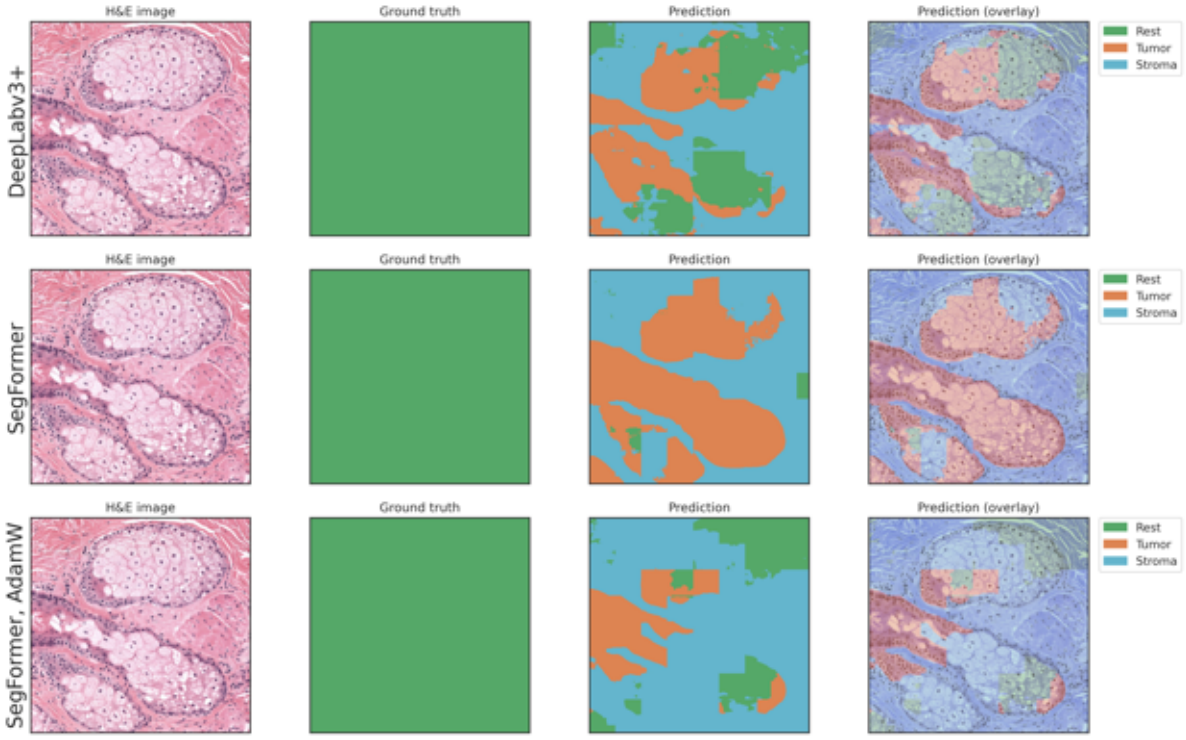


Figure 5.3: TC_S01_P000159_C0001_B108_[18565, 51594, 19649, 52655] ROI. Example of rich false positive segmentation RUMC ROI that contributes to the cases of close zero Dice scores. Each row includes H&E image, ground truth, and prediction with one of three developed models. This example showcases the tiling artifacts that are present across all models. Resulted Dice scores: DeepLabv3+ 0.132, SegFormer 0.009 and SegFormer, AdamW 0.093.

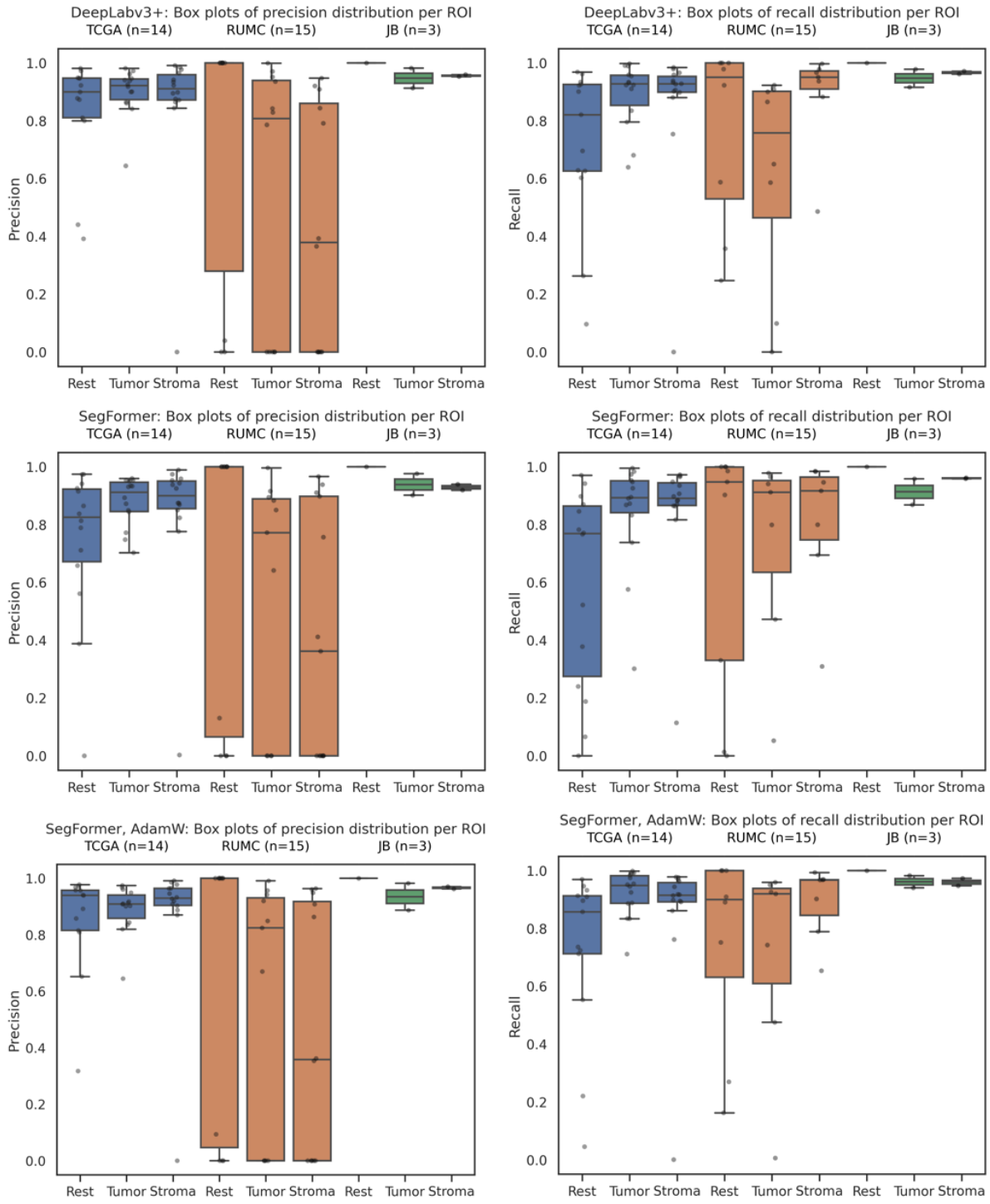


Figure 5.4: Boxplots of pixel wise calculated precision and recall across three datastes (TCGA-BRCA, RUMC, JB) and three segmentation labels. Columns: precision, recall. Rows: DeepLabv3+, SegFormer, and SegFormer, AdamW. Boxplot color helps to distinguish the values between different datasets and each dataset is represented by three boxes with the scores for each of the class labels - rest, tumor, and stroma.

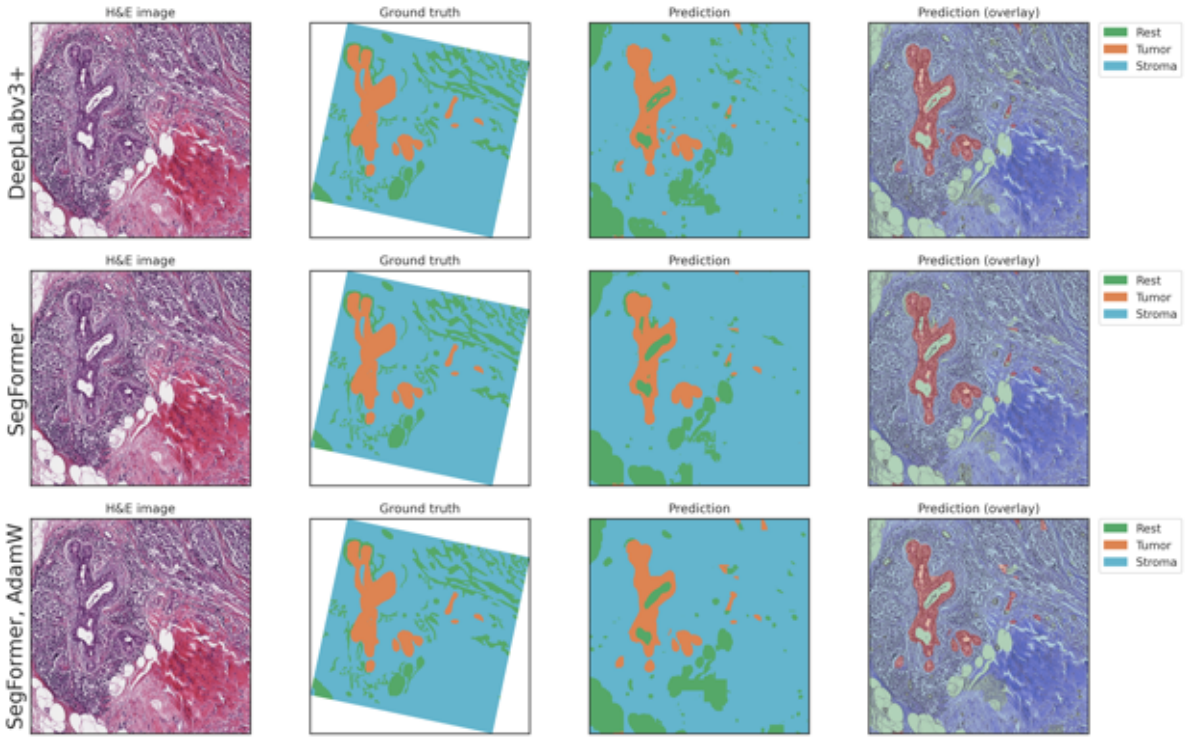


Figure 5.5: TCGA-GM-A2DF-01Z-00-DX1.CD0BE6D7-2DB3-4193-84CC-F9BE7BF18CC2_[25322, 21890, 27778, 24293] ROI. Example of a slightly devalued Dice score due to some annotation inaccuracies. Each row includes H&E image, ground truth, and prediction with one of three developed models. This example showcases the tilted ground truth. The whitespace was annotated as class zero in the original mask image. Resulted Dice scores: DeepLabv3+ 0.684, SegFormer 0.68, and SegFormer, AdamW 0.664.

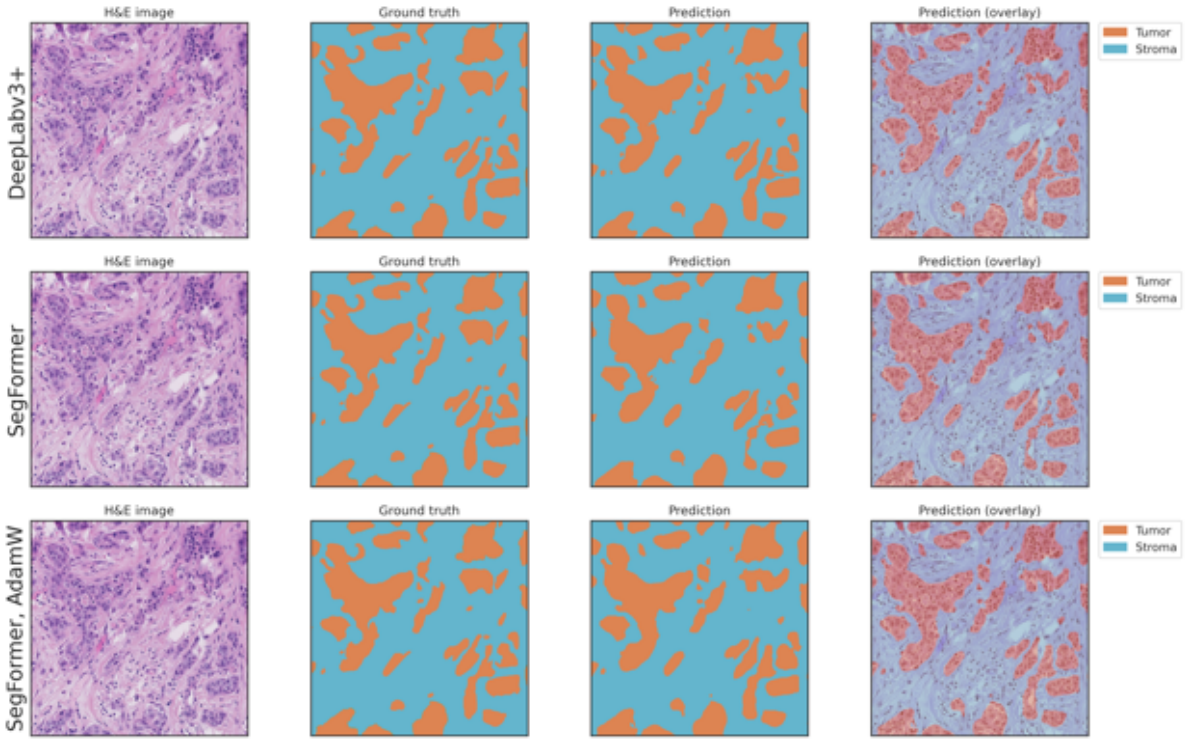


Figure 5.6: 250B_[16272, 25312, 17441, 26473] JB ROI. Each row includes H&E image, ground truth, and prediction with one of three developed models. Resulted Dice scores: DeepLabv3+ 0.937, SegFormer 0.917, and SegFormer, AdamW 0.936.

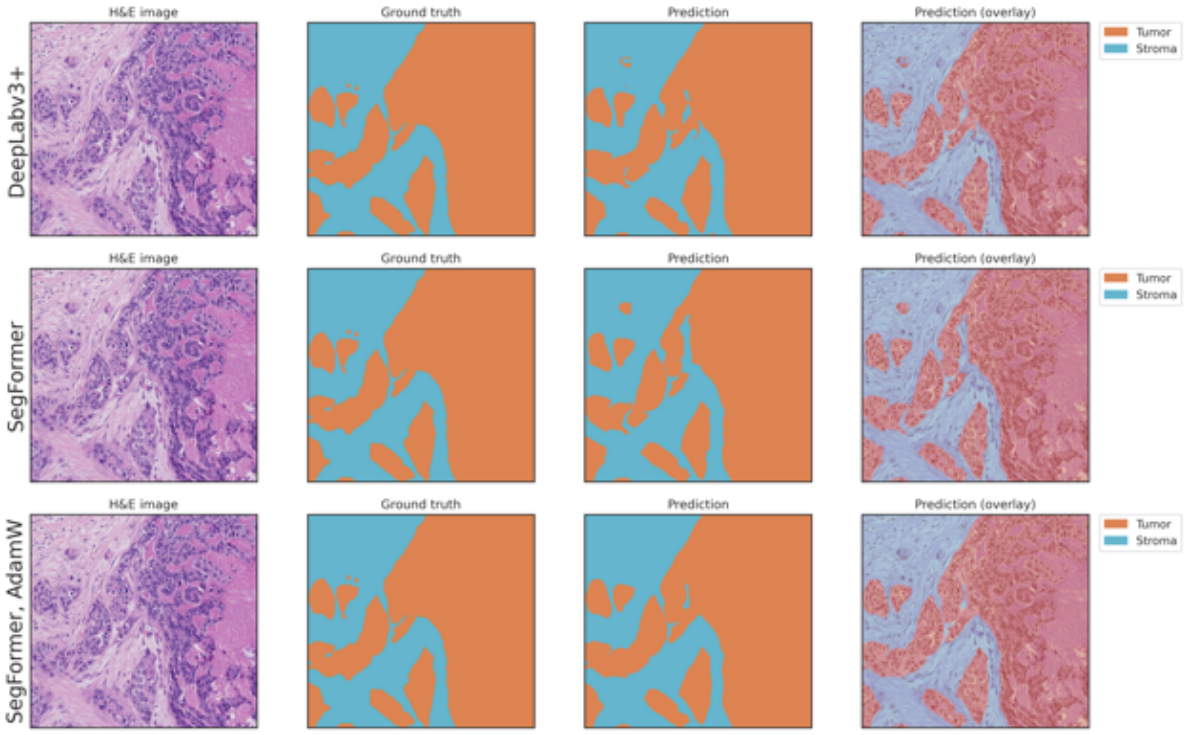


Figure 5.7: 250B_[29477, 26912, 30725, 28086] JB ROI. Each row includes H&E image, ground truth, and prediction with one of three developed models. Resulted Dice scores: DeepLabv3+ 0.971, SegFormer 0.953, and SegFormer, AdamW 0.975.

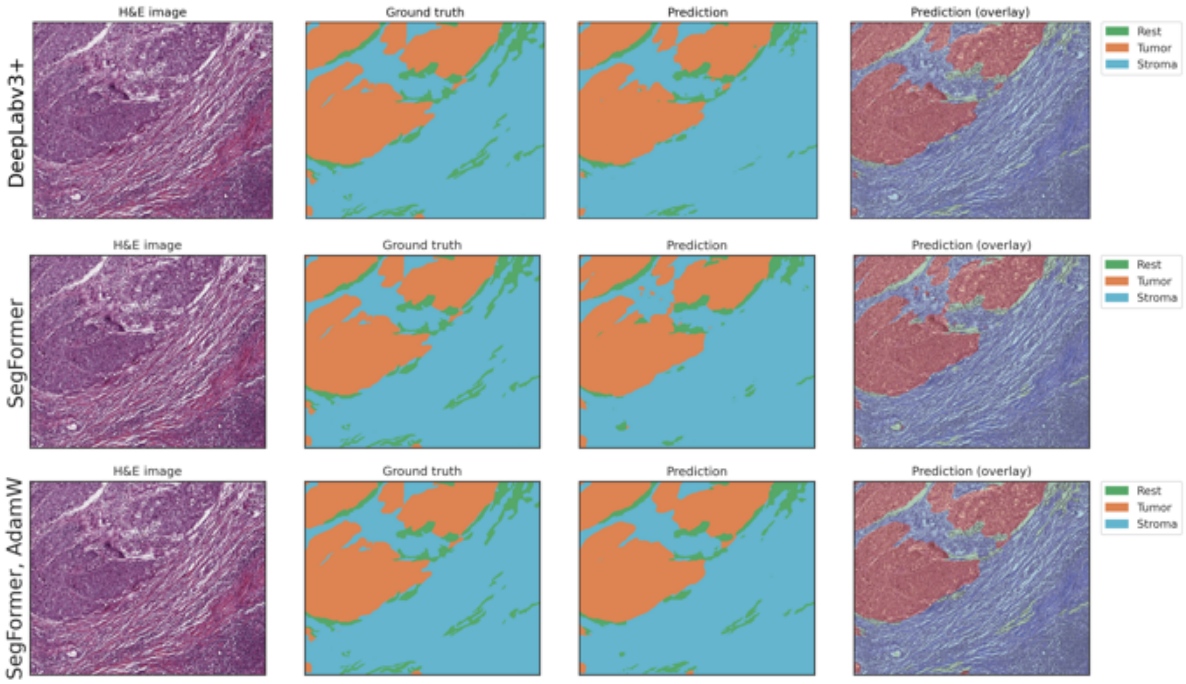


Figure 5.8: TCGA-EW-A1P4-01Z-00-DX1.3E9AE553-83D4-4B09-AB7F-D096BCE3BC4D_[8630, 17717, 11173, 19809] ROI. Each row includes H&E image, ground truth, and prediction with one of three developed models. Resulted Dice scores: DeepLabv3+ 0.88, SegFormer 0.842, and SegFormer, AdamW 0.917.

5.2 TIL Segmentation

TIL segmentation task aimed to segment the RGB input of H&E stained image at 20× magnification (resolution of 0.5 micron-per-pixel) into two prediction maps: TILs and rest. The split of the patients was used similarly to the previously described in Tabel 5.1. Even though the same patients are present in this data set, the annotations originate from a different study which results in a different ROI and patch statistics shown in Table 5.5. The patches were created using a sliding window approach with 128×128 sized patches and stride equals 100. The ROIs that were smaller than 128×128 were padded with 255. The additional rotation augmentation was applied, by rotating each patch 5 times at 9 degrees each. The column "Number of patches that include rest" in Table 5.5 seems excessive, but was still added for better data comprehension: the ROIs for TILs segmentation are either completely annotated as rest or include occasional TILs masks.

The model architectures and their parameters were used as described in 5.1: DeepLabv3+, SegFormer with Adam optimizer, and SegFormer with AdamW. An important change is an increased batch size of 128, which was allowed due to smaller patch sizes of 128×128. In the model overview in Table 5.6 the number of parameters and iterations remain the same.

	slides	ROIs	patches	Number of patches that include			
				TILs	Rest	1 class	2 classes
Train	156	1 552	106 974	35 433 (33%)	106 974 (100%)	71 541 (67%)	35 433 (33%)
Validation	20	154	15 518	14 164 (27%)	15 518 (100%)	5 638 (60%)	3 734 (40%)
Test	19	173	9 372	3 734 (40%)	9 372 (100%)	11 354 (73%)	4 164 (27%)
	195	1 879	131 864				

Table 5.5: Overview of patches that were partitioned into train, validation, and test sets for TILs segmentation following the patient split presented in Table 5.1. The percentages indicate the fraction of a specific conditioned group of patches to the number of all patches in the "patches" column.

FLOPs values are dependent on input shape, therefore it was expected to drop since the input is smaller. It is also noticed that for TILs segmentation the runtime between all three models becomes comparable in contrast to tissue segmentation (Table 5.3).

Model	FLOPs	Params	Iterations	Runtime	F1 score	Precision	Recall
DeepLabv3+	11.04	43.58 M	160 K	2d 8h 36m	0.49	0.58	0.43
SegFormer	3.24	81.97 M	160 K	2d 11h 36m	0.62	0.62	0.33
SegFormer, AdamW	3.24	81.97 M	160 K	2d 15h 19m	0.66	0.64	0.69

Table 5.6: Overview of the trained TILs segmentation models. The runtime is given for training on one GPU NVIDIA A100 SXM4. Doubled number of model parameters, as in tissue segmentation task (Table 5.3), but comparable training run time and significantly better F1 score of SegFormer model with AdamW optimizer.

For proper evaluation the predicted TILs segmentation needed to be reduced to TILs centers (one pixel) that can be then further matched to ground truth. To get optimal centers of predicted TILs, non-maximum suppression was applied on posterior images that were clipped between 0 and 255. The search for the best-fitted parameter of kappa (threshold) and kernel size was completed on the validation set. As pictured in Figure 5.9 there were multiple experiments performed with kernel sizes in [1, 3, 5, 7, 9, 11, 13] and multiple kappas. The highest value of kappa was the median value over all posteriors. The consecutive values were the two power fractions of the median. Figure 5.9 includes two images for DeepLabv3+ (first in the first row and first in the second row) to provide a zoomed look of the tighter range. As a result, the best parameters on the validation set were chosen as kappa=21, kernel size=9 for DeepLabv3+ and kappa=64, kernel size=5 for SegFormer-based methods. The resulting centers were matched to the ground truth by applying the Hungarian algorithm that found the best assignment to match ground truth TILs with predicted ones. The allowed maximum

distance for a match of predicted with ground truth TILs was set to 5 µm.

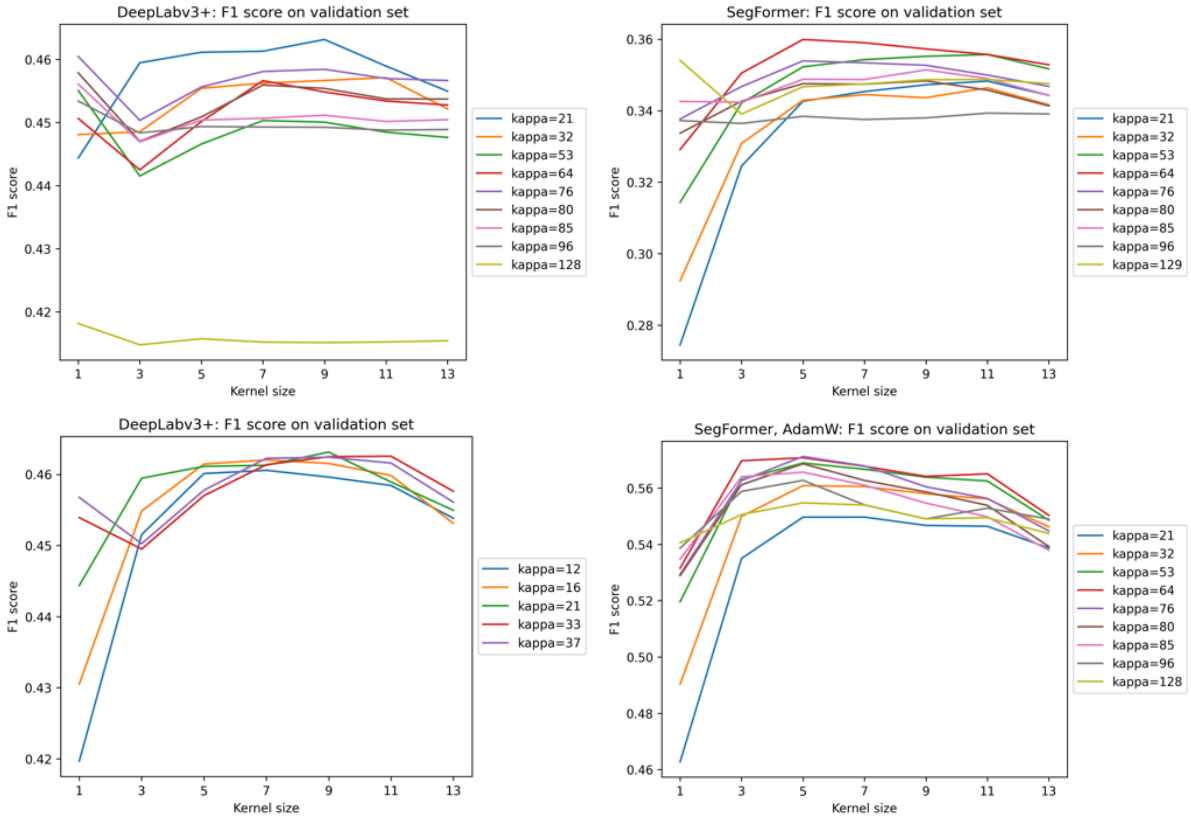


Figure 5.9: Determination process for best kappa and kernel size for DeepLabv3+ (first in the first row and first in the second row), SegFormer (second in the first row), and SegFormer with AdamW (second in the second row).

The final results represented in Table 5.6 show that SegFormer model with AdamW optimizer strongly outperforms DeepLabv3+ and simple SegFormer. Furthermore, while distinguishing the F1 scores between ROIs originating from different medical centers in Figure 5.10, SegFormer AdamW results show significantly better results in all subgroups. Interestingly the precision boxplots in Figure 5.11 do not show such an unequivocal superiority, where in the JB subgroup simple SegFormer manages to predict fewer false positives and (n=2) indicates that the model managed to predict empty ROI as a complete rest region, which is not the case for any other model. In more detailed Figure 5.12 one can see the intermediate steps of how the posteriors are simplified into point segmentations and later the misted, falsely annotated and correct TILs can be compared. Even on the level of posteriors (overlayed with H&E image), it is visible that SegFormer AdamW manages to detect more regions, especially closer to the border of the image. The grids in Figure 5.12 were added for readers' convenience and are not artifacts.

To evaluate the results of TILs detection TiGER challenge performed a Free Response Operating Characteristic (FROC) analysis, computing sensitivity versus average false positives

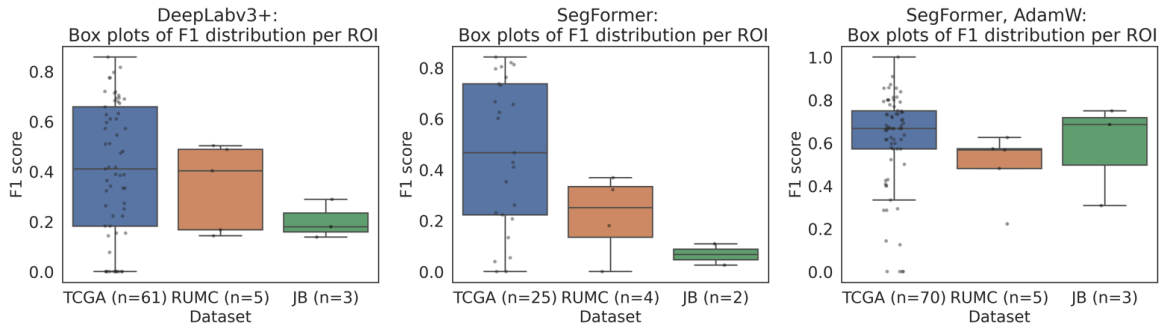


Figure 5.10: Boxplots of F1 score across three datastes (TCGA-BRCA, RUMC, JB) with maximum allowed distance between ground truth and prediction equals 10 pixels (5 µm). Each boxplot represents the result of one of three models: DeepLabv3+, SegFormer, SegFormer with AdamW optimizer. Within each boxplot, F1 scores per ROI are also distinguished between different source datasets.

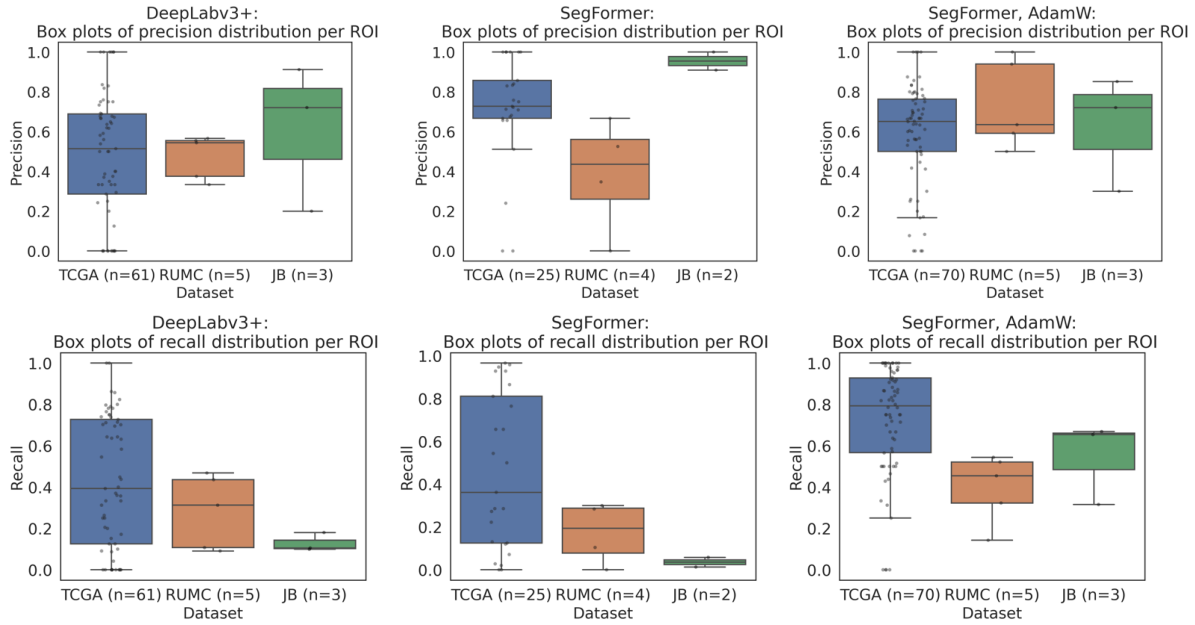


Figure 5.11: Boxplots of precision and recall across three datastes (TCGA-BRCA, RUMC, JB) with maximum allowed distance between ground truth and prediction equals 10 pixels (5 µm). Rows: precision, recall. Columns: DeepLabv3+, SegFormer, and SegFormer, AdamW. Within each boxplot, F1 scores per ROI are also distinguished between different source datasets.

per mm^2 over all test slides. The experimental set included 26 WSIs and 38 WSIs in the final dataset. The FROC ratio for three developed models was calculated on 173 test ROIs and revealed significant differences between the trained model in the thesis. The results can be viewed in Table 5.7. Even though the values are not directly comparable, since they are based

on different data sets, SegFormer, AdamW yield a FROC value close to 0.7, outperforming the TiGER results. Both the experimental and final TiGER best models belong to the Bio-totem team. They based the model on a modified SFCN-OPI network [85] which is a sibling fully convolutional network that performs nuclei detection and classification using weak labels. The resulting SFCN-OPI first contains a detection FCN branch, then the regions with high confidence of TILs existence from detection output are gathered with thresholding (objectness prior) and fed into the false positive suppressing branch for the TILs detection task.

	DeepLabv3+	SegFormer	SegFormer, AdamW	TiGER best (experimental)	TiGER best (final)
FROC score	0.432	0.326	0.693	0.600	0.5504

Table 5.7: Free Response Operating Characteristic (FROC) analysis for TILs detection. The TiGER challenge leaderboard results [2] versus three models developed in this work. The results were obtained from different data.

Once again, just as for tissue segmentation AdamW optimizer for transformer based method proved to be a better choice. But the statements that were done in the previous chapter 5.1 regarding tissue segmentation models comparison do not apply here, since SegFormer with AdamW optimizer manages to outperform DeepLabv3+ even though no pre-training or hyperparameter tuning was used. As Naseer, M *et al.* [86] discovered, when presented with the texture and shape of the same object, in this case typically dark, round to ovoid TIL, CNN models often make decisions based on texture. In contrast, transformers perform better than CNNs on shape recognition. This indicates the robustness of transformers to deal with significant distribution shifts and might be the reason for superior performance, due to for instance better recognition of TILs shapes in non-homogeneously stained slides. Taking into account all discussions above and the overall better performance, the SegFromer AdamW was considered the best model, and the model with the highest pixel-wise Dice score on the validation set was taken to the next step. For model inference, it was planned to use the model with the best pixel-wise Dice score on the validation set. But due to model conversion problems (.pth to .pt) that were discovered late by observing rather bad performance after conversion, DeepLabv3+ was used instead. Patches of size 128×128 were extracted from the tissue region at 20× magnification with a stride of 100. The whitespace was extracted by using thresholding and inference of non-background pixels was then performed.

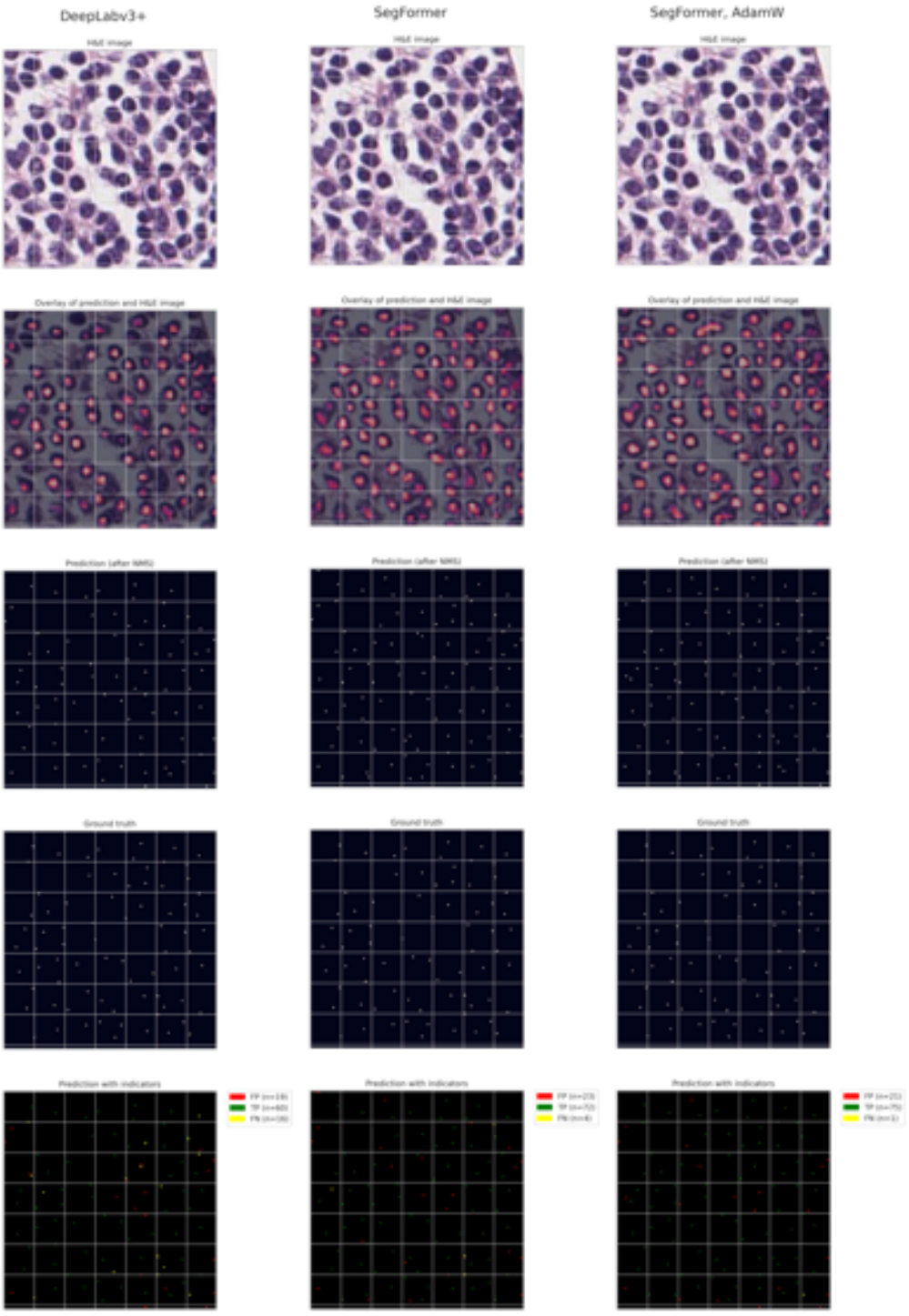


Figure 5.12: Visualisation of TCGA-D8-A142-01Z-00-DX12 ROI TILs detection. Columns: DeepLabv3+, SegFormer, and SegFormer, AdamW. Rows: H&E image, prediction posterior, segmentation mask after applying non-maximum suppression, ground truth, and prediction overlayed with the ground truth. All grids were added for easier visual navigation. Resulted F1 score: DeepLabv3+ 0.713, SegFormer 0.804, and SegFormer, AdamW 0.812.

5.3 Survival Analysis

The patients from TCGA-BRCA clinical data with negative or not complete event times were removed. The remaining 1015 patients were used for survival analysis. 1133 segmented diagnostic slides were saved at 5× magnification (resolution of 2 micron-per-pixel). Each slide has determined regions of tumor, stroma, and rest, as well as a list of all selected TILs. Based on this information four groups of TILs densities were calculated: in stroma, stroma border, tumor associated stroma border, and tumor border. The description of the regions is visualized in Figure 5.13. Additionally, to measure the heterogeneity of TIL density in the stroma, the densities of separate not connected stroma components, as two stroma regions depicted in Figure 5.13 were calculated. The resulting collection of stroma densities for every slide was transformed into a value of its mean, standard deviation, and variance. The idea behind it was not to only measure the densities but also to capture TIL distribution within a slide. The small stroma components less than 100 or 300 pixels, were filtered out. The features for patients with multiple slides were mean aggregated. The overall TILs density in stroma was considered a baseline feature. All densities were calculated as $\frac{\text{Number of TILs}}{\text{Area in } mm^2}$. For each border region, different widths were used. Tumor border experiments included lower border widths of 10-40 µm. Whereas stroma borders ranged between 50 and 350µm. Additionally, tumor associated stroma border of 125 µm was included, since this is the optimal border found for breast cancer patients [57]. It resulted in 20 features that were first evaluated on correlation as depicted on a heatmap Figure 5.14. There are three coherent groups visible in Figure 5.14. There is a clear link between different groups and the baseline, such as a high correlation of TIL density in stroma with TIL densities in stroma borders, which increases the bigger the border gets. And on a contrary, a lower correlation of baseline with TIL density in tumor borders. The values of pair-wise Pearson correlations in each group were compared and the columns with a correlation higher than 90% were removed. If all columns in a group displayed a correlation above 90%, the most similar column of a group was kept, which should represent a group best. The resulting features and their correlations can be seen on the right heatmap in Figure 5.14. For further comparison, features were evaluated at different prevalences. For every feature, the patients were sorted descending according to a current feature. For every prevalence in the range of 0.2 until 0.9, the patients were separated into according groups and the Kaplan Meier method was used on both of them. The Log-Rank test was applied to measure whether the two resulting event series are statistically different. The resulting p-values are visualized in Figure 5.16, both as a line plot per each prevalence, as well as a boxplot. In the boxplot, it can be seen that the median p-value of the density in 10 µm tumor and heterogeneity

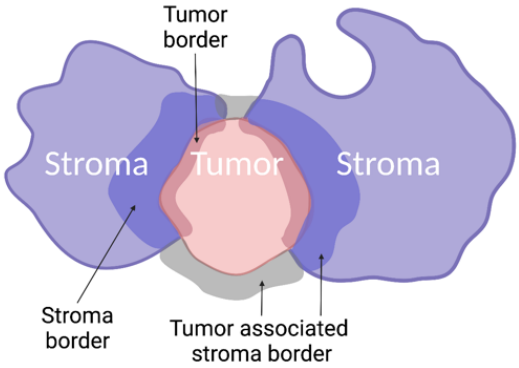


Figure 5.13: Schematic border explanation. Created with BioRender.com

lay higher than the other features. Those densities are also found more frequently over the statistical significance line in the line plot. Those features were decided to exclude. And for visibility, the remaining features were plotted again separately in Figure 5.17 with an additional boxplot depicting differences in median survival time at the same prevalence range. TILs density in tumor associated stroma achieving lower p-values in the lower prevalences. But TILs density in stroma border shows a better performance overall. It is also a feature that stays under the significance line across all prevalences. Additionally, according to the Median survival time boxplot, the median value of TILs densities in stroma border is 1.5 years higher.

For fitting the data into a Cox model, the TILs densities were divided by 100. The result revealed 0.577 concordance, 0.007 p-value, and 0.94 hazard ratio for baseline feature of TIL density in overall stroma. That means that the coefficient is negative (-0.06) which supports the assumption that the high level of TILs plays a role in longer survival probability. For the TILs density in stroma border, the result is identical: 0.579 concordance, 0.004 p-value, and 0.94 hazard ratio. And TILs density in tumor associated stroma reaches 0.588 concordance, 0.006 p-value, and 0.88 hazard ratio. Regression models generally give more reliable results with normally-distributed variables. Since the TILs densities are not normally-distributed (c.f. Figure 5.15), the experiments were repeated with log features. The TILs densities in stroma scored 0.577 concordance, p-value 0.00064 and 0.8 hazard ratio, for TILs desnities in stroma border - 0.579 concordance, p-value 0.00032 and 0.79 hazard ratio, and TILs densities in tumor associated stroma - concordance of 0.588, p-value 0.00018 and 0.78 hazard ratio. It was expected for concordance to stay indifferent since it was not influenced by taking the log of all values.

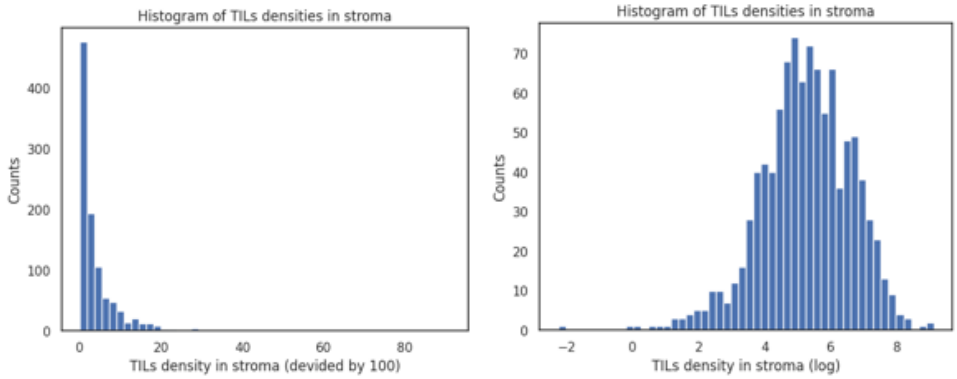


Figure 5.15: Histograms of baseline TILs density in stroma features versus its log distribution.

The Kaplan Meier curves in Figure 5.18 at low, median and high prevalences (0.33, 0.5 and 0.66), matched the cases when TILs density in tumor associated stroma features stratified patients better in lower range and median. Whereas at 0.66 the TILs density in stroma border scores a more significant p-value. The observation that a border of tumor associated stroma features in some prevalences perform better than the stroma border may come from the observed behavior that the tissue segmentation model leans to confuse rest with stroma (Figure 5.1). Hence, theoretically taking into consideration only regions close to tumorous

regions, that are better segmented, includes falsely missed stroma regions that therefore lead to a higher significance of a feature.

To evaluate TILs scores TiGER challenge was assessed in comparison to a baseline Cox regression model [2]. The TiGER baseline survival model was based on clinical variables that included age, morphology subtype, grade, molecular subtype, stage, surgery, and adjuvant therapy. The concordance of a baseline model reached 0.63. The submitted TIL scores were evaluated on the impact they had on the baseline model as an additional feature. The best result on the experimental set is 0.7194 [87]. The TILs score was calculated by dividing the total predicted TILs area by the stromal area and multiplying by 100. The score was further constrained to be an integer between 0 and 100. On the final dataset, the concordance barely improved and reached 0.6388. According to the final webinar, the winning team detected lymphocytes cells in the peritumoral stroma area and averaged it over all patches with an additional step of averaging with an accumulated ratio of detected lymphocytes cells in peritumoral stroma area not averaged over all patches.

To define a similar baseline model for TCGA-BRCA data, the following features were used: AJCC staging criteria identifying the extent of cancer (`ajcc_pathologic_stage`), gender, and age at which breast cancer was first diagnosed. The Cox model with these features showed an outstanding concordance of 0.77285. The addition of TIL density of stroma improved the value to 0.77795. Whereas the stroma border features had a bit less of an impact: stroma border with 0.77673 and tumor associated stroma border with 0.77719.

Lastly, 261 patients that were diagnosed with either Her2+ or TNBC breast cancer were extracted, if the respective diagnostic images were provided. As depicted in Figure 5.19 the Log-Rank test was applied to measure statistical differences in range prevalences with three features: stroma as the baseline, TIL density in 100µm tumor associated stroma border and TILs density in 300 µm stroma border. In this case, baseline feature is performing considerably better at the median split. While both border features have a rather similar course in the line plot, the median p-value shown in the boxplot below depicts the superiority of tumor associated stroma over stroma border, and even some improvement against the baseline. The Cox survival model for TIL density in stroma reports concordance of 0.59, p-value 0.033, and 0.92 hazard ratio. In contrast to the complete dataset, the stroma border shows comparatively worse results of 0.58 concordance, 0.04 p-value, and 0.93 hazard ratio. The tumor associated stroma border scores a slightly better concordance of 0.593, while the p-value reports a 0.08 and 0.89 hazard ratio.

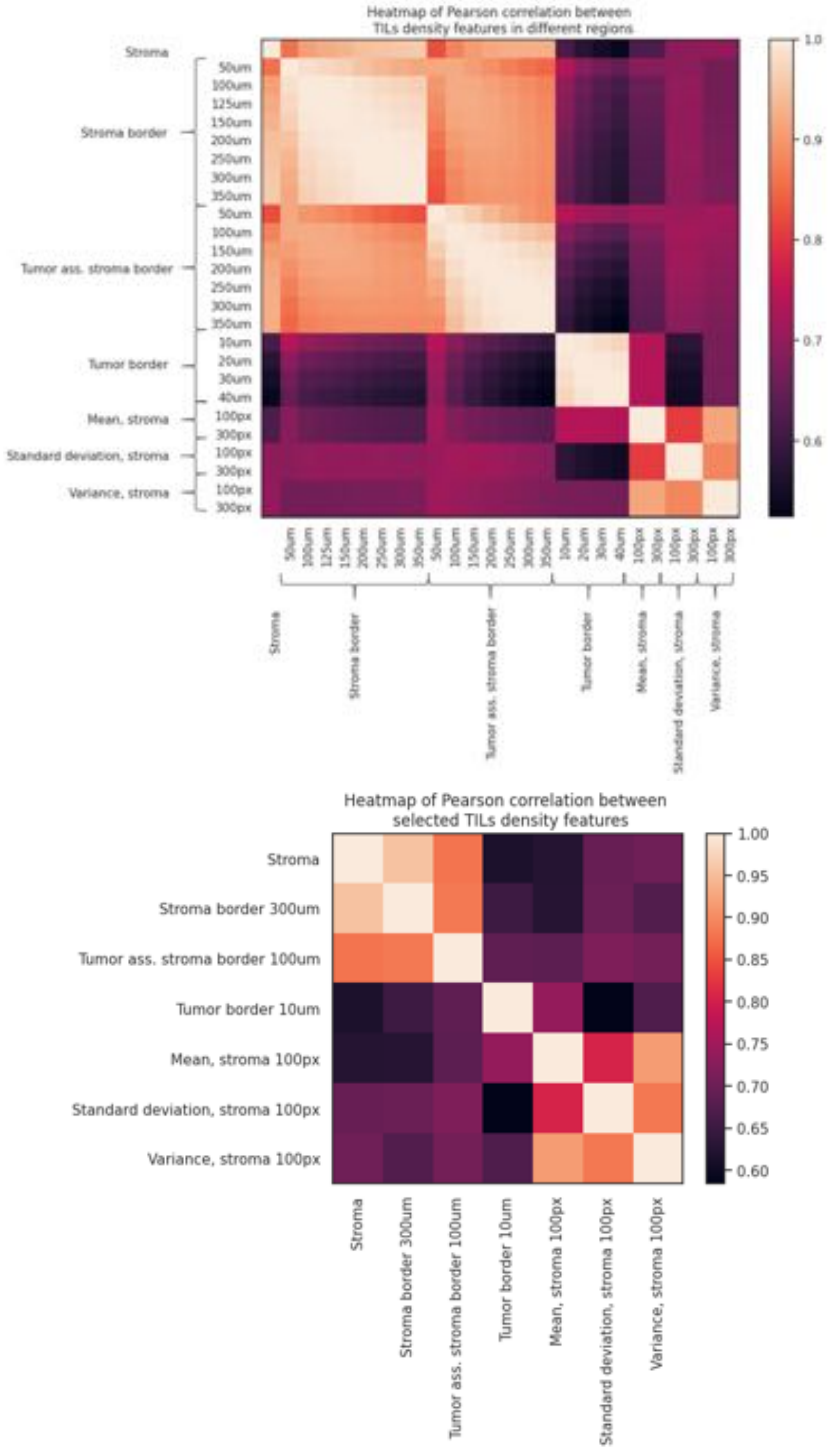


Figure 5.14: Heatmaps representing pair-wise Pearson correlation between TIL density features in different areas: stroma border, tumor associated stroma border, tumor border, complete stroma, and its heterogeneity. The smaller heatmap at the bottom includes only group representatives.

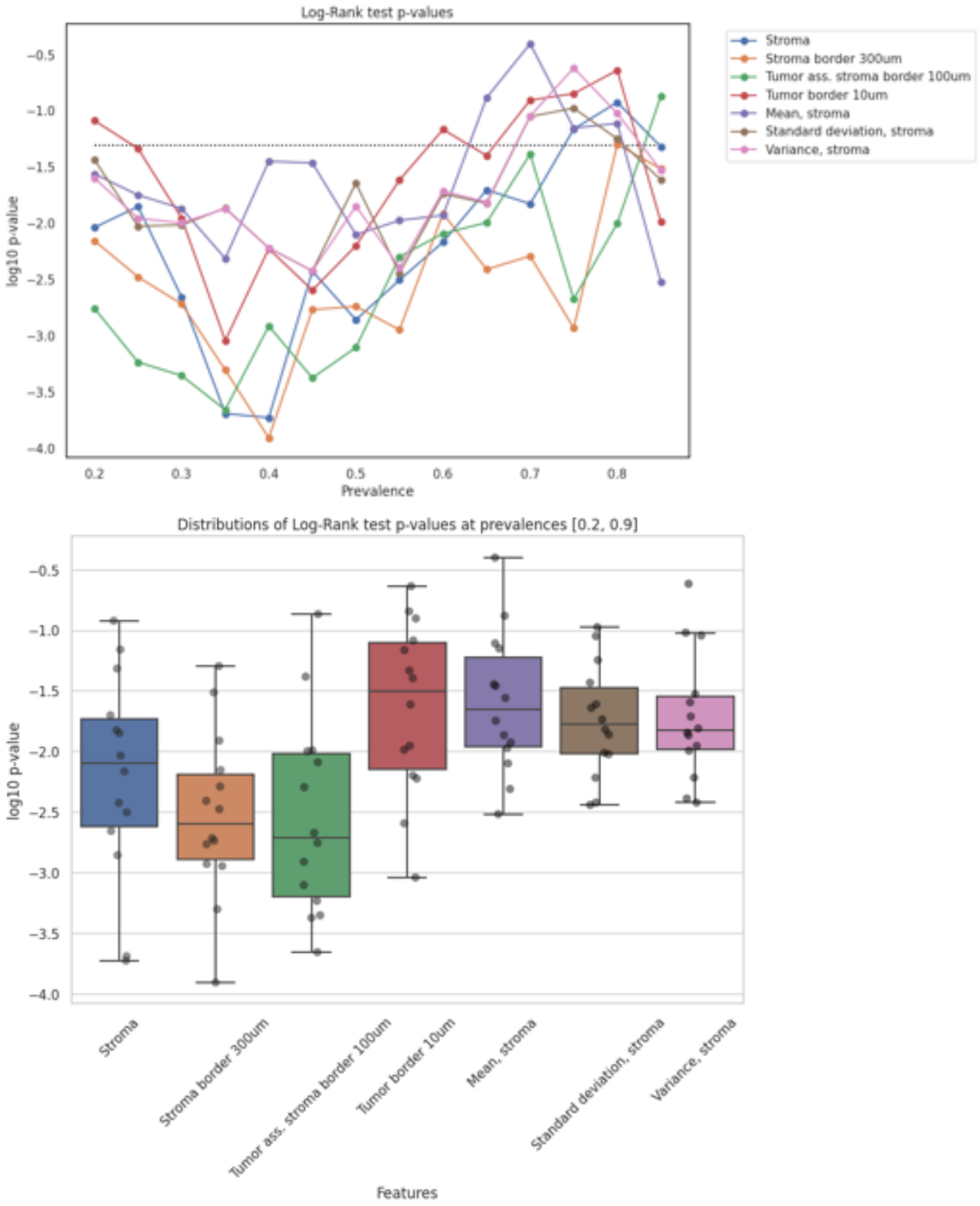


Figure 5.16: Log-Rank p-values distribution by different patient separations. Prevalance range between 0.2 and 0.9 for both plots. The TIL density in stroma (blue) was considered a baseline feature. The dotted line in the line plot represents the significance level of 0.05 ($\log_{10} 0.05$).

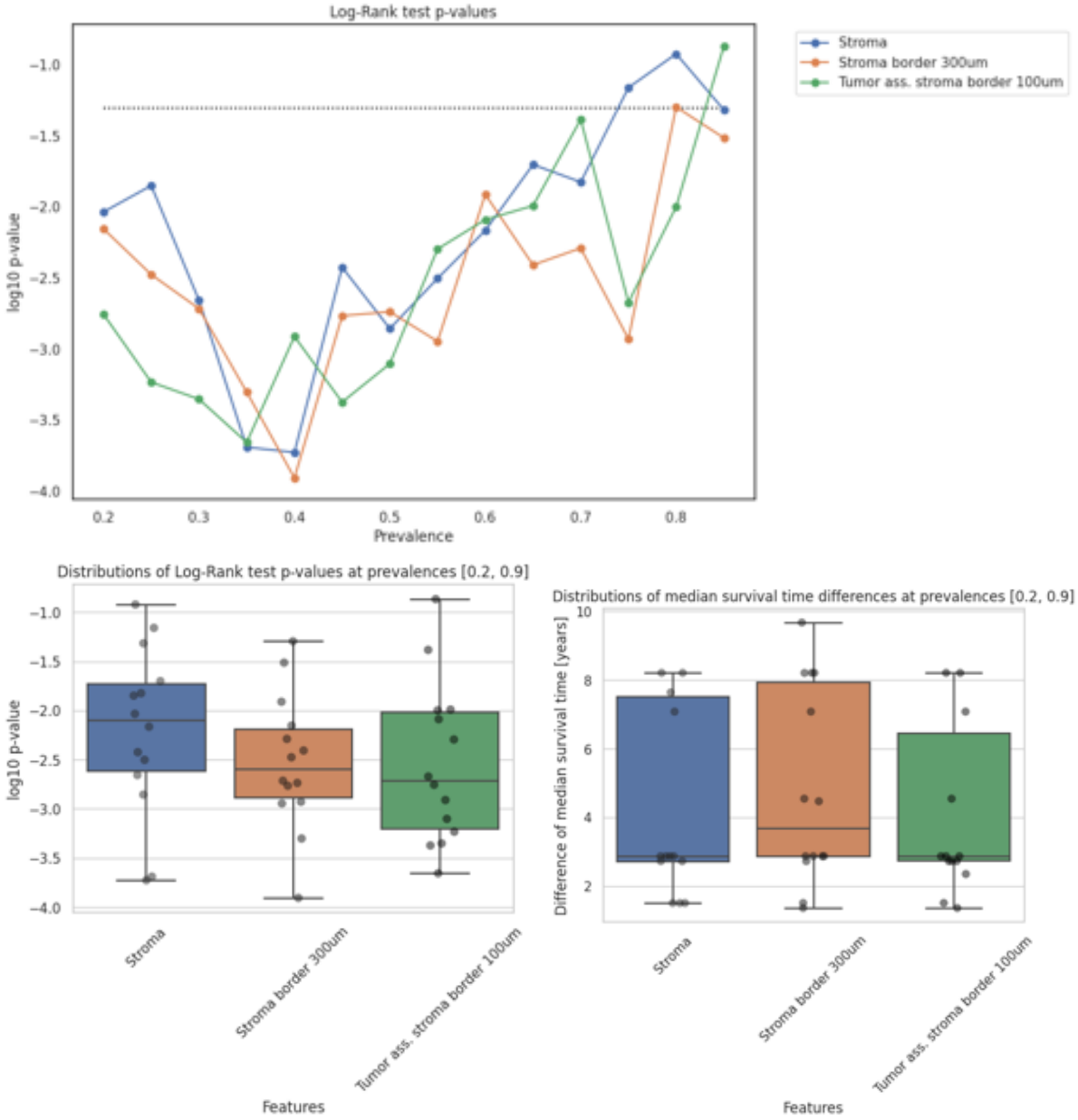


Figure 5.17: Log-Rank p-values distribution by different patient separations with an additional boxplot with median survival time differences. Prevalance range between 0.2 and 0.9 for both Log-Rank p-values plots. The TIL density in stroma (blue) was considered a baseline feature. The dotted line in the line plot represents the significance level of 0.05 ($\log_{10} 0.05$).

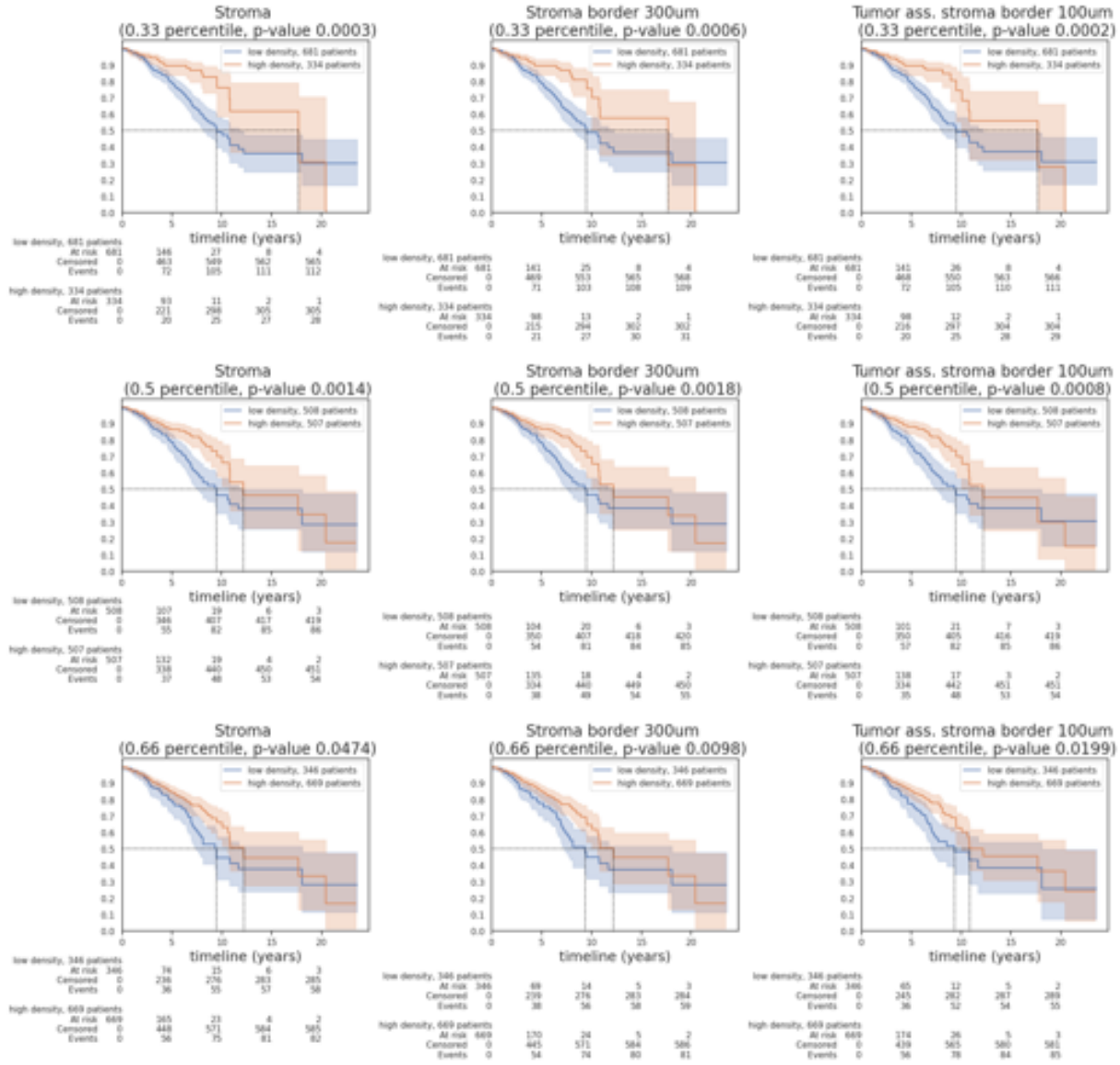


Figure 5.18: Kaplan Meier curves. Columns correspond to denisty features in: stroma, 300 µm stroma border, and 100 µm tumor associated stroma border. Rows refer to following prevalences: 0.33, 0.5, and 0.66.

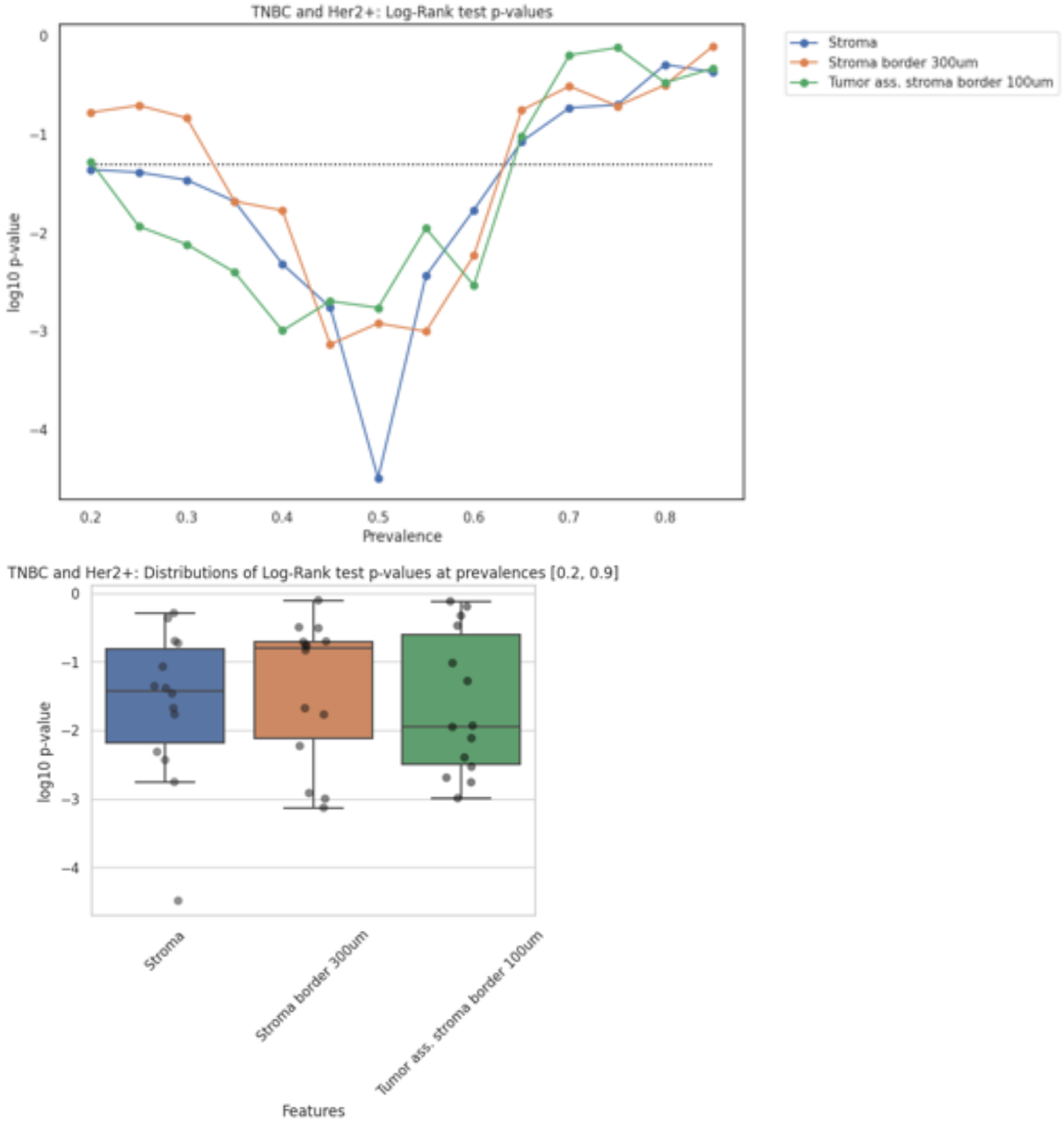


Figure 5.19: Her2+ and TNBC patient cohort. Log-Rank p-values distribution by different patient separations. Prevalance range between 0.2 and 0.9 for both plots. The TIL density in stroma (blue) was considered a baseline feature. The dotted line in the line plot represents the significance level of 0.05 ($\log_{10} 0.05$).

6 Conclusion

In this thesis, a new pipeline for TILs scoring in breast cancer patients was presented. The tasks of tissue segmentation and TILs detection for WSI were approached with DeepLabv3+ which is actively used in computational pathology, as well as with a novel transformer-based approach, SegFormer. As a result, the SegFormer was found to be superior in the TILs segmentation task compared to DeepLabv3+. The analysis of TCGA-BRCA slides with the resulting models, allowed for experiments with different TILs density based features. 100 µm tumor associated border and 300 µm stroma border showed the ability to successfully stratify patients into high and low TIL density groups at multiple prevalences. This proves signal exsitence of TILs in the TCGA-BRCA dataset, which was not previously demonstrated while analyzing exclusively this publically available dataset.

Undoubtedly the conversion issue for SegFormer models needs to be resolved so that TIL feature calculation can be based on the best-performing model. The fact that survival analysis was done based on DeepLabv3+ segmentation of tissue and TILs is a drawback.

As an outlook, one of the main missing prospects of this work is the absence of benchmarking with the existing challenge entries. Due to the late start of this work the submission portal was closed by the time the models were present. The TiGER challenge team that achieved the overall highest rank on the preliminary leaderboards for both segmentation of tissue and TILs published their approach before the challenge ended. The authors [87] trained Efficient-UNet-B0 for both tasks using Jaccard loss. The model was trained via a stratified 5-fold cross-validation framework with 512×512 patches at 10× magnification, with a stride of 256 pixels. Whereas for TILs segmentation, the patches of 128×128 were extracted at 20× magnification, with a stride of 100 pixels. The selected model with the best F1-score across each experiment on 5-fold cross-validation achieved a dice score of 0.762 for tumor segmentation, 0.718 for stroma segmentation, and 0.702 for TILs detection. In the meantime any of the results in this paper or any future papers of the TiGER challenge can be compared to the results in this thesis. Due to the arrangement of the challenge, the reported results are based on hidden test set and internal clincal data. Despite the fact that direct comparion is not possible, validation metrics detemined for the approches investigated in this work suggest that performace is comparible with the challenge winner. DeepLabv3+ tissue segmentation model scored 0.851 for tumor segmentation, 0.88 for stroma segmentation, and transformer based model SegFormer score slightly lower F1-score of 0.66. The final results on the challenge portal are summarized in Table 6.1. As for the survival analysis, the C-index of a baseline survival model was evaluated as 0.63. The added TILs score feature increased the concordance to 0.719 for predicting survival as part of a Cox proportional hazards model. Within this work the Cox model was attempted to be defined in the same way, but the baseline alone resulted a concordance of 0.77. As long as the submission server is closed,

	DeepLabv3+	SegFormer	SegFormer, AdamW	TIAger paper [87]
Tissue segmentation, Dice score	0.866	0.851	0.864	0.791
TILs detection, FROC score	0.432	0.326	0.693	0.572
Survival abalysis, C-index	0.78			0.719

Table 6.1: Tumor-stroma dice score for tissue segmentation, Free Response Operating Characteristic (FROC) analysis for TILs detection and concordance (C-index) for predicting survival as part of a Cox proportional hazards model. The TIAger publication [87] results versus three models developed in this work. The results were obtained from different data.

those models and ideas need to be reimplemented. Besides benchmarking, those challenge entries can provide valuable ideas for improvements to the current pipeline. Such as applying transfer-learning, by using this model with pre-trained weights from ImageNet for tissue segmentation which sould also improve SegFormer results according to the discussion in chapter 5.1 or an on-the-fly under-sampling approach to alleviate class imbalance of TILs segmentation task. There are also some parallels in approaches such as, similarly to this thesis, the authors [87] viewed TILs detection problem as TILs segmentation, but additionally, applied stain augmentation. Even though the importance of the stain addition could be suspected from the tissue segmentation results, this is a valuable experiment that should be performed. This thesis provides a better understanding and test possibilities for TILs features development. Instead of blindly submitting a value, adapting the TCGA-BRCA data set enables it to actually view the data specifics which could have helped the participants of the challenge (applicable if of course TCGA-BRCA data is not included in the hidden clinical data). Finally, the results depicted in Table 6.1 showcase that the selected DeepLabv3+ model for tissue segmentation and SegFormer for TILs detection are meritorious candidates for the challenge and let alone an efficient tool to gain TILs biomarker features for efficient patient prognosis analysis.

List of Figures

List of Tables

Bibliography

[1] B. S. Chhikara and K. Parang. "Global Cancer Statistics 2022: the trends projection analysis". In: (2022).

[2] *Tiger - Grand Challenge*. URL: https://tiger.grand-challenge.org/Home/.

[3] Q. D. Vu, S. Graham, T. Kurc, M. N. N. To, M. Shaban, T. Qaiser, N. A. Koohbanani, S. A. Khurram, J. Kalpathy-Cramer, T. Zhao, et al. "Methods for segmentation and classification of digital microscopy tissue images". In: *Frontiers in bioengineering and biotechnology* (2019), p. 53.

[4] R. Salgado, C. Denkert, S. Demaria, N. Sirtaine, F. Klauschen, G. Pruneri, S. Wienert, G. Van den Eynden, F. L. Baehner, F. Pénault-Llorca, et al. "The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014". In: *Annals of oncology* 26.2 (2015), pp. 259–271.

[5] C. Denkert, G. von Minckwitz, S. Darb-Esfahani, B. Lederer, B. I. Heppner, K. E. Weber, J. Budczies, J. Huober, F. Klauschen, J. Furlanetto, et al. "Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy". In: *The lancet oncology* 19.1 (2018), pp. 40–50.

[6] A.-V. Laenkholm, G. Callagy, M. Balancin, J. Bartlett, C. Sotiriou, C. Marchio, M. Kok, C. H. Dos Anjos, and R. Salgado. "Incorporation of TILs in daily breast cancer care: how much evidence can we bear?" In: *Virchows Archiv* (2022), pp. 1–16.

[7] M. V. Dieci, N. Radosevic-Robin, S. Fineberg, G. Van den Eynden, N. Ternes, F. Penault-Llorca, G. Pruneri, T. M. D'Alfonso, S. Demaria, C. Castaneda, et al. "Update on tumor-infiltrating lymphocytes (TILs) in breast cancer, including recommendations to assess TILs in residual disease after neoadjuvant therapy and in carcinoma in situ: a report of the International Immuno-Oncology Biomarker Working Group on Breast Cancer". In: *Seminars in cancer biology*. Vol. 52. Elsevier. 2018, pp. 16–25.

[8] S. Adams, R. J. Gray, S. Demaria, L. Goldstein, E. A. Perez, L. N. Shulman, S. Martino, M. Wang, V. E. Jones, T. J. Saphner, et al. "Prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancers from two phase III randomized adjuvant breast cancer trials: ECOG 2197 and ECOG 1199". In: ASCO. 2014.

[9] S. Loi, N. Sirtaine, F. Piette, R. Salgado, G. Viale, F. Van Eenoo, G. Rouas, P. Francis, J. Crown, E. Hitre, et al. "Prognostic and predictive value of tumor-infiltrating lymphocytes in a phase III randomized adjuvant breast cancer trial in node-positive breast cancer comparing the addition of docetaxel to doxorubicin with doxorubicin-based chemotherapy: BIG 02-98". In: *J Clin Oncol* 31.7 (2013), pp. 860–867.

[10] C. Denkert, S. Loibl, A. Noske, M. Roller, B. Muller, M. Komor, J. Budczies, S. Darb-Esfahani, R. Kronenwett, C. Hanusch, et al. "Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer". In: *J Clin Oncol* 28.1 (2010), pp. 105–113.

[11] G. Gao, Z. Wang, X. Qu, and Z. Zhang. "Prognostic value of tumor-infiltrating lymphocytes in patients with triple-negative breast cancer: a systematic review and meta-analysis". In: *BMC cancer* 20.1 (2020), pp. 1–15.

[12] M. A. Postow, M. K. Callahan, and J. D. Wolchok. "Immune checkpoint blockade in cancer therapy". In: *Journal of clinical oncology* 33.17 (2015), p. 1974.

[13] W. Chung, H. H. Eum, H.-O. Lee, K.-M. Lee, H.-B. Lee, K.-T. Kim, H. S. Ryu, S. Kim, J. E. Lee, Y. H. Park, et al. "Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer". In: *Nature communications* 8.1 (2017), pp. 1–12.

[14] A. Schneeweiss, C. Denkert, P. A. Fasching, C. Fremd, O. Gluz, C. Kolberg-Liedtke, S. Loibl, and H.-J. Lück. "Diagnosis and therapy of triple-negative breast cancer (TNBC)–recommendations for daily routine practice". In: *Geburtshilfe und Frauenheilkunde* 79.06 (2019), pp. 605–617.

[15] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos. "Image segmentation using deep learning: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* (2021).

[16] J. Long, E. Shelhamer, and T. Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[17] H. Noh, S. Hong, and B. Han. "Learning deconvolution network for semantic segmentation". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1520–1528.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[19] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[20] A. BenTaieb and G. Hamarneh. "Topology aware fully convolutional networks for histology gland segmentation". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, pp. 460–468.

[21] V. A. Natarajan, M. S. Kumar, R. Patan, S. Kallam, and M. Y. N. Mohamed. "Segmentation of nuclei in histopathology images using fully convolutional deep neural architecture". In: *2020 International Conference on computing and information technology (ICCIT-1441)*. IEEE. 2020, pp. 1–7.

[22] M. Amgad, A. Sarkar, C. Srinivas, R. Redman, S. Ratra, C. J. Bechert, B. C. Calhoun, K. Mrazeck, U. Kurkure, L. A. Cooper, et al. "Joint region and nucleus segmentation for characterization of tumor infiltrating lymphocytes in breast cancer". In: *Medical Imaging 2019: Digital Pathology*. Vol. 10956. SPIE. 2019, pp. 129–136.

[23] D. A. Gutman, J. Cobb, D. Somanna, Y. Park, F. Wang, T. Kurc, J. H. Saltz, D. J. Brat, L. A. Cooper, and J. Kong. "Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data". In: *Journal of the American Medical Informatics Association* 20.6 (2013), pp. 1091–1098.

[24] M. Amgad, H. Elfandy, H. Hussein, L. A. Atteya, M. A. Elsebaie, L. S. Abo Elnasr, R. A. Sakr, H. S. Salem, A. F. Ismail, A. M. Saad, et al. "Structured crowdsourcing enables convolutional segmentation of histology images". In: *Bioinformatics* 35.18 (2019), pp. 3461–3467.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.

[26] A. B. Hamida, M. Devanne, J. Weber, C. Truntzer, V. Derangère, F. Ghiringhelli, G. Forestier, and C. Wemmert. "Deep learning for colon cancer histopathological images analysis". In: *Computers in Biology and Medicine* 136 (2021), p. 104730.

[27] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241.

[28] A. Lagree, M. Mohebpour, N. Meti, K. Saednia, F.-I. Lu, E. Slodkowska, S. Gandhi, E. Rakovitch, A. Shenfield, A. Sadeghi-Naini, et al. "A review and comparison of breast tumor cell nuclei segmentation performances using deep convolutional neural networks". In: *Scientific Reports* 11.1 (2021), pp. 1–11.

[29] Z. Zeng, W. Xie, Y. Zhang, and Y. Lu. "RIC-Unet: An improved neural network based on Unet for nuclei segmentation in histology images". In: *Ieee Access* 7 (2019), pp. 21420–21428.

[30] H. Pinckaers and G. Litjens. "Neural ordinary differential equations for semantic segmentation of individual colon glands". In: *arXiv preprint arXiv:1910.10470* (2019).

[31] K. R. Oskal, M. Risdal, E. A. Janssen, E. S. Undersrud, and T. O. Gulsrud. "A U-net based approach to epidermal tissue segmentation in whole slide histopathological images". In: *SN Applied Sciences* 1.7 (2019), pp. 1–12.

[32] M. E. Bagdigen and G. Bilgin. "Cell segmentation in triple-negative breast cancer histopathological images using U-Net architecture". In: *2020 28th Signal Processing and Communications Applications Conference (SIU)*. IEEE. 2020, pp. 1–4.

[33] M. Van Rijthoven, M. Balkenhol, K. Siliņa, J. Van Der Laak, and F. Ciompi. "HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images". In: *Medical Image Analysis* 68 (2021), p. 101890.

[34] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.

[35] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. "Rethinking atrous convolution for semantic image segmentation". In: *arXiv preprint arXiv:1706.05587* (2017).

[36] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818.

[37] B. M. Priego-Torres, D. Sanchez-Morillo, M. A. Fernandez-Granero, and M. Garcia-Rojo. "Automatic segmentation of whole-slide H&E stained breast histopathology images using a deep convolutional neural network architecture". In: *Expert Systems With Applications* 151 (2020), p. 113387.

[38] Y. Xie, J. Zhang, C. Shen, and Y. Xia. "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2021, pp. 171–180.

[39] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville. "Reseg: A recurrent neural network-based model for semantic segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016, pp. 41–48.

[40] A. Chakravarty and J. Sivaswamy. "RACE-net: a recurrent neural network for biomedical image segmentation". In: *IEEE journal of biomedical and health informatics* 23.3 (2018), pp. 1151–1162.

[41] M. Saha and C. Chakraborty. "Her2Net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation". In: *IEEE Transactions on Image Processing* 27.5 (2018), pp. 2189–2200.

[42] C. Nguyen, Z. Asad, R. Deng, and Y. Huo. "Evaluating transformer-based semantic segmentation networks for pathological image segmentation". In: *Medical Imaging 2022: Image Processing*. Vol. 12032. SPIE. 2022, pp. 942–947.

[43] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10012–10022.

[44] Z. Qian, K. Li, M. Lai, E. I. Chang, B. Wei, Y. Fan, Y. Xu, et al. "Transformer based multiple instance learning for weakly supervised histopathology image segmentation". In: *arXiv preprint arXiv:2205.08878* (2022).

[45] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang. "Ds-transunet: Dual swin transformer u-net for medical image segmentation". In: *IEEE Transactions on Instrumentation and Measurement* (2022).

[46] R. Strudel, R. Garcia, I. Laptev, and C. Schmid. "Segmenter: Transformer for semantic segmentation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 7262–7272.

[47] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. "Transunet: Transformers make strong encoders for medical image segmentation". In: *arXiv preprint arXiv:2102.04306* (2021).

[48] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. "SegFormer: Simple and efficient design for semantic segmentation with transformers". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12077–12090.

[49] E. L. Kaplan and P. Meier. "Nonparametric estimation from incomplete observations". In: *Journal of the American statistical association* 53.282 (1958), pp. 457–481.

[50] D. R. Cox. "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202.

[51] Z. Kos, E. Roblin, R. S. Kim, S. Michiels, B. D. Gallas, W. Chen, K. K. van de Vijver, S. Goel, S. Adams, S. Demaria, et al. "Pitfalls in assessing stromal tumor infiltrating lymphocytes (sTILs) in breast cancer". In: *NPJ breast cancer* 6.1 (2020), pp. 1–16.

[52] M. Amgad, R. Salgado, and L. A. Cooper. "MuTILs: Explainable, multiresolution computational scoring of Tumor-Infiltrating Lymphocytes in breast carcinomas using clinical guidelines". In: *medRxiv* (2022).

[53] M. Amgad, L. A. Atteya, H. Hussein, K. H. Mohammed, E. Hafiz, M. A. Elsebaie, A. M. Alhusseiny, M. A. AlMoslemany, A. M. Elmatboly, P. A. Pappalardo, et al. "Nucls: A scalable crowdsourcing, deep learning approach and dataset for nucleus classification, localization and segmentation". In: *arXiv preprint arXiv:2102.09099* (2021).

[54] H. Le, R. Gupta, L. Hou, S. Abousamra, D. Fassler, L. Torre-Healy, R. A. Moffitt, T. Kurc, D. Samaras, R. Batiste, et al. "Utilizing automated breast cancer detection to identify spatial distributions of tumor-infiltrating lymphocytes in invasive breast cancer". In: *The American journal of pathology* 190.7 (2020), pp. 1491–1504.

[55] Y. Bai, K. Cole, S. Martinez-Morilla, F. S. Ahmed, J. Zugazagoitia, J. Staaf, A. Bosch, A. Ehinger, E. Nimeus, J. Hartman, et al. "An Open-Source, Automated Tumor-Infiltrating Lymphocyte Algorithm for Prognosis in Triple-Negative Breast CancerMachine-Read TIL Variables in TNBC". In: *Clinical Cancer Research* 27.20 (2021), pp. 5557–5565.

[56] A. Kapil, A. Meier, A. Shumilov, S. Haneder, H. Angell, and G. Schmidt. "Breast cancer patient stratification using domain adaptation based lymphocyte detection in HER2 stained tissue sections". In: (2021).

[57] J. Thagaard, E. S. Stovgaard, L. G. Vognsen, S. Hauberg, A. Dahl, T. Ebstrup, J. Doré, R. E. Vincentz, R. K. Jepsen, A. Roslind, et al. "Automated quantification of stil density with h&e-based digital image analysis has prognostic potential in triple-negative breast cancers". In: *Cancers* 13.12 (2021), p. 3050.

[58] P. Sun, J. He, X. Chao, K. Chen, Y. Xu, Q. Huang, J. Yun, M. Li, R. Luo, J. Kuang, et al. "A computational tumor-infiltrating lymphocyte assessment method comparable with visual reporting guidelines for triple-negative breast cancer". In: *EBioMedicine* 70 (2021), p. 103492.

[59] K. E. Craven, Y. Gökmen-Polar, and S. S. Badve. "CIBERSORT analysis of TCGA and METABRIC identifies subgroups with better outcomes in triple negative breast cancer". In: *Scientific reports* 11.1 (2021), pp. 1–19.

[60] D. J. Fassler, L. A. Torre-Healy, R. Gupta, A. M. Hamilton, S. Kobayashi, S. C. Van Alsten, Y. Zhang, T. Kurc, R. A. Moffitt, M. A. Troester, et al. "Spatial Characterization of Tumor-Infiltrating Lymphocytes and Breast Cancer Progression". In: *Cancers* 14.9 (2022), p. 2148.

[61] S. Meng, L. Li, M. Zhou, W. Jiang, H. Niu, and K. Yang. "Distribution and prognostic value of tumor-infiltrating T cells in breast cancer". In: *Molecular medicine reports* 18.5 (2018), pp. 4247–4258.

[62] V. Kotoula, K. Chatzopoulos, S. Lakis, Z. Alexopoulou, E. Timotheadou, F. Zagouri, G. Pentheroudakis, H. Gogas, E. Galani, I. Efstratiou, et al. "Tumors with high-density tumor infiltrating lymphocytes constitute a favorable entity in breast cancer: a pooled analysis of four prospective adjuvant trials". In: *Oncotarget* 7.4 (2016), p. 5074.

[63] J. Saltz, R. Gupta, L. Hou, T. Kurc, P. Singh, V. Nguyen, D. Samaras, K. R. Shroyer, T. Zhao, R. Batiste, et al. "Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images". In: *Cell reports* 23.1 (2018), pp. 181–193.

[64] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[66] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[67] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. "Soft-NMS–improving object detection with one line of code". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5561–5569.

[68] H. W. Kuhn. "The Hungarian method for the assignment problem". In: *Naval research logistics quarterly* 2.1-2 (1955), pp. 83–97.

[69] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman. "Survival analysis part I: basic concepts and first analyses". In: *British journal of cancer* 89.2 (2003), pp. 232–238.

[70] C. Davidson-Pilon. "lifelines: survival analysis in Python". In: *Journal of Open Source Software* 4.40 (2019), p. 1317.

[71] Y.-C. Chen. "Lecture 5: Survival Analysis". In: *STAT 425: Introduction to Nonparametric Statistics, University of Washington* (Winter 2018).

[72] T. Therneau and E. Atkinson. "1 The concordance statistic". In: (2020).

[73] C. Smith. *FFPE or frozen? working with human clinical samples*. Nov. 2014. URL: https://www.biocompare.com/Editorial-Articles/168948-FFPE-or-Frozen-Working-with-Human-Clinical-Samples/.

[74] M. Contributors. *MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark*. https://github.com/open-mmlab/mmsegmentation. 2020.

[75] M. Khened, A. Kori, H. Rajkumar, G. Krishnamurthi, and B. Srinivasan. "A generalized deep learning framework for whole-slide image segmentation and analysis". In: *Scientific reports* 11.1 (2021), pp. 1–14.

[76] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. "Unified perceptual parsing for scene understanding". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 418–434.

[77] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu. "Visual attention network". In: *arXiv preprint arXiv:2202.09741* (2022).

[78] Y. Liu, E. Sangineto, W. Bi, N. Sebe, B. Lepri, and M. Nadai. "Efficient training of visual transformers with small datasets". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 23818–23830.

[79] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. "Training data-efficient image transformers & distillation through attention". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10347–10357.

[80] X. Chen, C.-J. Hsieh, and B. Gong. "When vision transformers outperform ResNets without pre-training or strong data augmentations". In: *arXiv preprint arXiv:2106.01548* (2021).

[81] D. Choi, C. J. Shallue, Z. Nado, J. Lee, C. J. Maddison, and G. E. Dahl. "On empirical comparisons of optimizers for deep learning". In: *arXiv preprint arXiv:1910.05446* (2019).

[82] I. Loshchilov and F. Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

[83] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick. "Early convolutions help transformers see better". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 30392–30400.

[84] Anonymous. "Applying Second Order Optimization to Deep Transformers with Parameter-Efficient Tuning". In: *Submitted to The Eleventh International Conference on Learning Representations*. under review. 2023. URL: https://openreview.net/forum?id=4Fi-5Jiyy5w.

[85] Y. Zhou, Q. Dou, H. Chen, J. Qin, and P.-A. Heng. "Sfcn-opi: Detection and fine-grained classification of nuclei using sibling fcn with objectness prior interaction". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[86] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. Shahbaz Khan, and M.-H. Yang. "Intriguing properties of vision transformers". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 23296–23308.

[87] A. Shephard, M. Jahanifar, R. Wang, M. Dawood, S. Graham, K. Sidlauskas, S. A. Khurram, N. Rajpoot, and S. E. A. Raza. "TIAger: Tumor-Infiltrating Lymphocyte Scoring in Breast Cancer for the TiGER Challenge". In: *arXiv preprint arXiv:2206.11943* (2022).