Almotion Bavaria

Technische Hochschule Ingolstadt

Fakultät Informatik
Fachgebiet Bildverstehen und medizinische Anwendung der künstlichen Intelligenz

Prediction of Histopathological Markers for Computer-Aided Grading of Mammary Carcinoma

Bachelor Thesis

Laura Klose

Supervisor: Prof. Dr. Marc Aubreville
Second examiner: Prof. Dr. Torsten Schön
Date: May 25, 2023

DECLARATION.

Specimen for declaration in accordance with § 18 Para. 4 Nr. 7 APO THI

I hereby declare that this thesis is my own work, that I have not presented it elsewhere for examination purposes and that I have not used any sources or aids other than those stated. I have marked verbatim and indirect quotations as such.

Ingolstadt, 25.05.2023 Laura Klose

Acknowledgements

I would like to take this opportunity to express my gratitude to several individuals who have contributed to the completion of my bachelor thesis. First and foremost, I am immensely grateful to my professor, Prof. Dr. Marc Aubreville, for his support, guidance, and expertise throughout this journey. Your passion for the subject has been a huge inspiration for me.

I would also like to extend my sincere appreciation to Jonas Ammeling and Jonathan Ganz, whose assistance have been invaluable. Your feedback, discussions, and willingness to quickly help me when I had questions have greatly enriched my research and strengthened the quality of my thesis.

I am grateful to Chloé Puget, who annotated the dataset and answered all my questions about the medical background of this research. Additionally, I would like to thank Frauke Wilm for her open-source repository.

Furthermore, I want to acknowledge the contributions of Sabrina and Chris. Your feedback, encouragement, and friendship have been a source of inspiration and motivation. I would be remiss not to mention my parents and grandparents, who have supported me throughout my academic path, even when it may have been challenging for you to fully understand.

Abstract

An important task of pathologists is to assign a grade during a tumor diagnosis. The grade correlates with a prognosis, that indicates the chances of survival. This prognosis plays a central role in therapy selection and prospects. Breast carcinomas have many histological subtypes, which are quite similar among themselves. Nevertheless, the pathologists need to perform the grading as accurately as possible. There are multiple grading schemes available. All schemes evaluate the 3 criteria: tubule formation, nuclear pleomorphism and mitotic count. During the evaluation, a label between 1-3 is assigned to each criterion. Afterward, the sum of all labels determines the tumor grade. Whereas the mitotic count is a rather objective evaluation, the tubule formation and nuclear pleomorphism labels are subjective decisions and therefore a strong inter-observer variability exists. Having an algorithm that could provide a deterministic evaluation, would improve reliability and could also improve diagnostic accuracy. The goal of this bachelor thesis is to develop a machine learning-based framework, that on the one hand classifies the grade of the subjective criteria (i.e., tubule formation and nuclear pleomorphism) and on the other hand semantically segments important tissue regions used for the prediction. The overall goal of this framework is to support pathologists in clinical decision-making by providing not only a score for each category per image, but also segmentation information and thus providing a more detailed and interpretable decision support. The dataset provided for this thesis contains mammary tumors in dogs. Due to the similarity between canine and human mammary tissue, successes achieved with canine tissue could be directly transferred to humans. Since annotating many images to train a supervised model is time-consuming, a relatively small dataset is provided. According to that limitation, the first research topic is how to overcome this problem by using transfer learning, image augmentation and image preprocessing. Secondly, it will be evaluated whether the outputs of the segmentation and classification tasks are precise enough to be used by a pathologist.

Contents

1 Introduction

In recent years, tumors have become one of the main diseases, causing the deaths of many dogs [1]. Among the most common types of tumors concerning female dogs are mammary tumors, which are mostly diagnosed around the age of 6–18 years [1]. To determine the prognosis and treatment options for these neoplasms at an early stage, the histologic grade plays a significant role. Therefore, the grading schema by Peña et al. [2] evaluates the three criteria: tubule formation (TF), nuclear pleomorphism (NP) and mitotic count (MC). A score between 1 and 3 is assigned to each criterion. The sum of all scores then yields the tumor's malignancy grade. A grade of 1 means "low malignancy" [3], whereas a grade of 3 indicates "high malignancy" [3] of the cells.

The schema helps to standardize the diagnoses, but there is still an inter-observer variability between pathologists when assigning the malignancy grade [4]. The study by Ginter et al. [4] compares the grading results of six experts from different institutes on digital whole slide images (WSIs). When assigning the scores, pathologists have the highest agreement rate on the tubule formation score (0.503 kappa), followed by nuclear pleomorphism (0.403 kappa) and mitotic rate with the lowest agreement rate (0.281 kappa) [4]. Regarding the resulting tumor grade, grade 1 had the highest similarity (0.705 kappa) compared to grades 2 (0.375 kappa) and 3 (0.491 kappa) [4]. All in all, it becomes clear that prognoses by different pathologists vary, but in order to find the optimal treatment for the patient, they need to be stable and also replicable [4]. Having an algorithm that could provide a deterministic evaluation on those complex tissue images, would improve reproducibility and might also improve diagnostic accuracy. Although the dataset of this thesis consists of mammary tumors images of dogs, the algorithm can be applied to human breast cancer too, due to the "[similarities] in their pathological, molecular, and visual characteristics"[5].

The problem with neural network algorithms is, that they are mostly not fully interpretable, making it difficult to evaluate and trust them. Other separate studies have already implemented the segmentation of all 3 biomarkers. Especially when assessing the mitotic count, previous studies have demonstrated good results with machine learning models. Hence, this marker was excluded from the present research, which instead concentrates on evaluating tubule formation and nuclear pleomorphism scoring. Later, the mitotic count can be added to the scores in order to compute the final tumor grade.

The main problem that confronted this work was the limited amount of annotated

samples and the uneven distribution of classes in the provided dataset. According to that limitation, the first research topic was how to overcome this problem by using transfer learning, image augmentation, and image preprocessing. Secondly, it should be evaluated whether the outputs of the segmentation and classification tasks were precise enough to be used by a pathologist.

This thesis is structured as follows: In the second chapter, mammary tumors will be introduced. Within, the risk factors are explained and an overview about the different types of mammary carcinomas is provided. Next, the importance of histologic grading is emphasized, and an overview of treatment options provided. Moreover, the field of pathology is introduced, detailing the production of histopathological images and highlighting the recent technological advancement: digital pathology. Additionally, the grading schema used in this study is described and the challenges associated with grading mammary tumors are identified. Chapter 3 discusses the results of previous studies, their materials and approaches. Afterward, we will look at the different datasets, further materials, and applied methods within Chapter 4. Then, in Chapter 5, the results for the different models are summarized and explained. Chapter 6 evaluates and discusses the results of the final models. Finally, in Chapter 7, a conclusion is drawn considering the research questions, and an outlook is provided.

2 Theoretical Background

This chapter, starts with an overview about Canine Mammary Tumors. In the next section, the field of histopathology will be introduced. Within the last section, the grading scheme used in this thesis to perform the histologic grading will be presented. Furthermore, issues that are relevant for either pathologists or algorithms while performing the overall grading will be mentioned. Finally, problems that occur specifically in the tubule formation and nuclear pleomorphism scoring, that were addressed in similar studies, will be outlined.

2.1 Canine Mammary Tumors

This section provides a more in depth overview about Canine Mammary Tumors (CMTs). At first, I will summarize the main risk factors that contribute to such a diagnosis and further introduce the histologic types that occur most commonly. In the final paragraph, the importance of histopathological examination will be highlighted.

2.1.1 Risk Factors

Canine mammary tumors originate from the mammary gland, which is composed of glandular and connective tissue, and can be classified as either benign or malignant [6]. As stated by Sorenmo et al. [7], there are several risk factors that can contribute to the development of CMTs in dogs. 1. Age: Incidences of mammary tumors in dogs increases with age, with most cases occurring in dogs that are over 5 years old [7]. 2. Hormonal Exposure: Hormonal factors, including exposure to estrogen and progesterone, are thought to play a significant role in the development of CMTs [7]. Female dogs that have not been spayed or have been spayed later in life have an increased risk of developing CMTs [7]. 3. Breeds and Genetic Susceptibility: Smaller breeds of dogs are at higher risk of developing CMTs than larger ones [7].

2.1.2 Histologic Types

CMTs are classified into different histologic types based on their cellular characteristics and growth patterns [3]. Most of the mammary tumors are either adenomas (simple or complex), mixed tumors (benign) or carcinomas [3]. Canine mammary carcinomas (CMCs) are the topic of this thesis. For them, the grading scheme can be applied to determine the malignancy [3]. CMCs can be divided into different subtypes again, like ductal tubular

adenocarcinoma, ductal tubulopapillary adenocarcinoma, ductal solid adenocarcinoma, comedocarcinoma, lymphangiosis carcinomatosa, and complex adenocarcinoma. The existence of different subtypes makes grading more difficult for experts. All the mentioned subtypes are part of the dataset, used in this work.

2.1.3 Importance of Histopathological Examination and Treatment Options

The first step in tumor identification is physical examination, during which each mammary gland is examined and palpated [6]. Furthermore, a fine needle aspiration (FNA), in which tissue fluid is removed with a needle, can be performed to check for cancer cells [6]. Physical examination and FNA are low-cost options to ensure that a tumor is present, but in order to define the type, grade, and stage of CMTs, the histopathological examination is a critical component [6]. The parameters for type, grade, and stage of the tumor are crucial in determining the appropriate treatment plan for each individual case. Several treatment options are available, including surgery, radiation therapy, and chemotherapy [6].

2.2 Histopathology

In the previous section, the importance of histopathological examination for canine mammary tumor diagnosis and research has been emphasized. Now, the generation of histopathological images with the commonly used Hematoxylin and Eosin staining technique will be described. Afterward, the recently emerged alternative to a manual evaluation of these images by a pathologist, the field of digital pathology, will be introduced.

2.2.1 Histopathological Images

Histopathological images are generated through a process that involves the collection, processing, and staining of tissue samples. Hematoxylin and Eosin (H&E) staining is the most commonly used staining technique, producing images with high contrast and detail [8]. The process of generating histopathological images begins with the collection of tissue samples from patients, which are typically obtained through biopsy or surgical resection [9]. Once the tissue has been collected, it is processed through a series of steps to prepare it for staining and imaging, which are described by Slaoui et al. [10]:

1. Fixation: First, the tissue is fixed in a solution, typically formalin, which preserves the tissue structure and prevents decay. 2. Trimming: Afterward, samples are trimmed to

a size suitable for the further steps. 3. Pre-embedding and 4. Embedding: The samples are then soaked and embedded in paraffin wax to remove water and prevent them from moving. 5. Sectioning: The embedded tissue is then sliced into thin sections, typically 4-5 μm-thick, using a microtome. 6. Staining and Mounting: Finally, these almost transparent sections are mounted on glass slides and stained using the H&E technique. [10]

In the H&E staining process, the tissue sections are first stained with hematoxylin, which binds to the acidic components of the tissue, such as nuclei, and stains them in a blue tone [11] [12]. Afterward, the sections are stained with eosin, which binds to the basic components of the tissue, such as proteins, and stains them purple to pink [11] [12]. The resulting histopathological images show the detailed structures within the sample, allowing medical professionals to identify abnormalities and make a diagnosis.

2.2.2 Digital Pathology

Microscopic evaluation of tissue images is an important procedure for diagnostic evaluation [13]. Currently, this is mostly done manually by a pathologist and requires "long training [...], quality control by peer reviews, and personal experience" [13]. However, in recent years, there has been great technical progress in the field, so that preparations can now be digitized with the aid of a scanner. The virtual microscopy method is also known as digital pathology and provides whole slide images that can now be analyzed digitally [13]. Another advantage, besides the high-resolution digital images, is that the patients' metadata is also available digitally. This and the recent progress in image analysis algorithms are paving the way for the automation of the analysis of histopathological images.

2.3 Histologic Grading of Canine Mammary Carcinomas

Histopathological examination of mammary tumors is crucial for accurate diagnosis, prognosis, and treatment planning. A common grading system is the Elston and Ellis method [14], also known as the Nottingham method [2]. The Elston and Ellis method is primarily used for humans [2]. The grading scheme developed by Peña et al. [2] is a minor modification of this method to be used in canine mammary tumors. Peña et al. [2] state that an adjustment of the criteria is required due to a lower spread of cancer cells, different probability distributions of tumor types and different markers to define the grade of malignancy for dog tumors.

Table 1. Criteria for Histologic Malignant Grade adapted from Peña et al. [2].

Score	Tubule Formation	Nuclear Pleomorphism	Mitoses per 10 HPF[1]
1 point	TF > 75% of the specimen	Uniform or regular small nucleus and occasional nucleoli	0-9 mitoses/10 HPF
2 points	Moderate TF 10-75% of the specimen admixed with areas of solid tumor growth	Moderate degree of variation in nuclear size and shape, hyperchromatic nucleus and presence of nucleoli	10-19 mitoses/10 HPF
3 points	Minimal or no TF (<10%)	Marked variation in nuclear size and hyperchromatic nucleus, often with one or more prominent nucleoli	> 20 mitoses/10 HPF

The grading scheme according to Peña et al. [2] is based on the assessment of three criteria: tubule formation, nuclear pleomorphism, and mitoses per 10 HPF. Each criterion is assigned points from 1 to 3 in the analysis based on the indicators visualized in Table 1. The sum of all points yields the malignancy grade of the tumor [2]. A grade of 1 (3-5 points) means the malignancy grade is "low" [2], 2 (6-7 points) stands for "intermediate" [2] and a grade of 3 (8-9 points) points out a "high" [2] grade.

2.4 Challenges in Canine Mammary Carcinoma Grading

Challenges for Pathologists The grading of CMTs can be challenging. Although the grading scheme helps to standardize the diagnosis, there is still an inter-observer variability between pathologists when assigning the malignancy grade [4]. One reason for discrepancies in the grading of tumors is that the mammary carcinoma is a very heterogeneous tumor type [2]. This can make it difficult to assign an accurate grade to the tumor. Furthermore, the grading of tubule formation and nuclear pleomorphism are subjective tasks, leading to considerable differences in the interpretation among different experts.

Challenges for AI A machine-learning algorithm could help standardize the grading process. In order to achieve good results on such complex and heterogeneous images, the

[1] HPF, high-power field

algorithm needs large amounts of high-quality data. To be used in a clinical diagnosis, it should also be interpretable.

The annotation process for a supervised learning approach is labor-intensive, which is why pathologists can only provide a small dataset. Additionally, due to the rarity of some mammary neoplasms, only a limited amount of sample images can be provided per type. This needs to be taken in consideration, because an uneven distribution in the dataset, can lead to overfitting or biased models.

Besides data shortage, the algorithm needs to function effectively for a wide range of patients and generalize over variations in staining and image artifacts that can arise in the "sample preparation [...] and during the imaging process" [15]. Finally, challenges can occur due to the different scanners that exist on the market, using different settings such as "different compression types and sizes, illumination, objectives, and resolution" [15]. This can also cause problems when developing the algorithm and needs to be kept in mind during the image preprocessing [15].

Challenges in Segmentation of Tubule Formation Tekin et al. [16] point out the challenges in segmentation of tubule formations. Tubule formations are unevenly distributed within tumoral regions and can have various shapes. Within the creation process of H&E images, it is not always possible to ensure exact shapes, which can result in reduced size, unclear borders, and blurring of the lumen. Furthermore, the lumen can be surrounded by more than one layer of nuclei, and thus an algorithm would have problems identifying the correct borders. Lastly, machine learning systems could misidentify tubules as "[other] structures with a lumen, such as adipose tissue, blood vessels [...] [or] other mammary glands" [16].

Challenges in Scoring of Nuclear Pleomorphism While challenges can occur when segmenting tubule formations, the scoring of nuclear pleomorphism can be difficult to learn for a network. When assigning a nuclear pleomorphism score, an inter-observer variability between pathologists exists. To capture a wider range of perspectives in the training dataset, incorporating annotations from multiple pathologists can be beneficial in developing an algorithm that predicts more precise and reproducible scores. [17]

3 Related Work

In recent years, deep learning has emerged as a powerful tool for solving complex computer vision problems and has been increasingly used in digital pathology applications like breast tumor grading. Traditionally, breast cancer grading has been a subjective process that can be influenced by factors such as inter-observer variability and fatigue. The motivation in newer studies is to develop an algorithm that can help standardize and automate this process, and thus provide more accurate and reproducible results. Within this chapter the popular grading schemes and accomplishments in other works, working on scoring each histopathological marker, will be evaluated. Additionally, the used materials and model architectures will be summarized. Afterward, I will present how other studies use the transfer learning method to overcome limitations of a smaller dataset and introduce methods that were applied to address the challenges in segmenting tubule formations and nuclear pleomorphism.

Scoring of Histopathological Markers Previous studies have already worked on AI systems that are relevant for the grading like segmenting tumor areas [18], counting mitosis ([19, 20, 21, 22, 23]), scoring nuclear pleomorphism [18] [17] [24] and scoring [24] or detecting tubule formations [18] [16].

While some studies focused on predicting single biologic markers, Jaroensri et al. [18] and Wetstein et al. [24] built model pipelines to predict the overall tumor grade. When predicting the overall grade, Wetstein et al. [24] showed that better results were reached when scoring all three grading components (tubule formation, nuclear pleomorphism, mitotic count) separately first and sum them up to receive the overall tumor grade afterward. On the other hand, a training dataset that provided only the final tumor grade in the supervised training, lead to more mistakes in the evaluation and is also less explainable [24].

Individual kappa scores comparing pathologists and the models' output in the works of Wetstein et al. [24] and Jaroensri et al. [18] indicate, that the best results per category were achieved for mitotic count, followed by tubule formation. In contrast, both reached lower scores for nuclear pleomorphism scoring. Especially with mitotic count, further studies (i.e., [19, 20, 21, 22, 23]) could demonstrate good results.

Results and Provided Dataset Although a supervised learning approach, that contains a score for all three markers in the dataset has yield better results, it makes the dataset creation process more time-consuming.

A recent study by Jaroensri et al. [18] relied on 498 slides with different grades of nuclear pleomorphism examples and 499 slides with different grades for tubule formation within the training dataset. With their model, they reached quadratic-weighted kappa scores of 0.45 for nuclear pleomorphism and 0.70 for tubule formation.

Another approach by Wetstein et al. [24], achieved accuracy scores of 0.70 ($\pm$0.02) on both, the tubular and nuclear score. For the training 99 samples of nuclear scores and 100 samples of tubule formation scores were available [24].

In contrast to previous studies, my training dataset incorporated 195 regions of interest for nuclear pleomorphism and 60 regions of interest for tubule formation. Therefore, the contribution of this thesis was to prove whether good results can still be achieved with a smaller dataset, to potentially reduce annotation workload in the future.

Model Architecture A model architecture, that is widely used in segmentation tasks and was developed for the application in biomedical images is the U-Net, published by Ronneberger et al. [25] in 2015. It was developed to improve localization and fine boundary detection of Convolutional Neural Networks [25]. However, Tekin et al. [16] pointed out, that the traditional U-Net architecture might not be enough to detect fine boundaries of tubule structures. When testing different architectures, they noticed that a Tubule-U-Net framework based on the EfficientNetB3-U-Net achieved better measurement results, compared to a traditional U-Net [16]. Similarly, the Tumor Segmentation Model by Wilm et al. [26] is also using a modified U-Net implementation with a ResNet encoder.

Transfer learning Another common practice in medical imaging is transfer learning. Transfer learning can help to increase the models' performance, when only a limited number of images is available. The idea of transfer learning is to use a pre-trained model that has been trained on a large dataset as a base for training a new model for another task. The pre-trained model has already adapted its parameters to recognize a wide range of features from the data. Those features can be used as a starting point for the new task. Previous research has shown that the use of transfer learning, enables higher starting performance and a faster and better convergence towards a local minimum can be expected. [27]

Focusing on the similarity of datasets in the training and pretraining, one can distinguish between "homogeneous and heterogeneous transfer learning" [28]. Homogeneous transfer learning happens, if the model is fine-tuned with a dataset from the same domain as the initial training dataset. Popular breast tumor grading approaches, e.g., Wetstein et al. [24] have also achieved good results on heterogeneous, cross-domain transfer datasets, e.g., with the ImageNet dataset [29]. The ImageNet dataset compromises a wide range of images and studies have shown that models, that learned from it, can be applied on datasets from other domains as well ([30], [31]).

Addressing Tubule Formation Segmentation challenges Besides using a more complex model architecture, Tekin et al. [16] addressed the difficulties of segmenting tubule formations, using a mirror technique. The primary objective of this approach was to preserve more intact tubule structures that would otherwise be cut off during the patch creation [16]. In their study, Tekin et al. [16], work with detailed annotations, for each tubule formation.

Addressing Nuclear Pleomorphism Scoring challenges Recent papers have treated inter-observer variability in the scoring of pleomorphism differently. Some have translated the scoring from labels 1-3 into a binary classification between "low/intermediate and high" [24]. This approach reduces complexity and generates more informative outputs [24]. In contrast, Mercan et al. [17] chose to transform the discrete scores into a continuous figure to map trends in the decision process of the annotators. This provided more information for the model. By utilizing this technique, the output scores of the network by Mercan et al. [17] achieved higher agreement rates between different experts during testing.

4 Materials and Methods

My algorithm was designed to classify and segment the three grades of tubule formation and nuclear pleomorphism in H&E-stained samples of Canine Mammary Carcinomas (CMCs). I was provided with annotated regions of interest of digital histopathological images, showing different types of CMCs. The main dataset utilized was the CMC dataset, referred to as **Mamma-Ca** dataset. Due to the time-consuming nature of annotating whole slide images, only a limited number of annotation masks were available.

Limited data can cause the model to either overfit or underfit, which will result in a low accuracy on the test data. Therefore, the training dataset was enhanced with a second open-source dataset, provided by Wilm et al. [26]. This additional dataset, consisting of annotated WSIs of squamous cell carcinoma, will be referred to as **SCC dataset**. The dataset was chosen due to the similar structure of squamous cell carcinomas and mammary carcinomas.

In the following, materials and methods will be presented in detail.

Table 2. Dataset characteristics for training, evaluation and testing.

SCC dataset			
Data	Train	Valid	Test
No. of H&E WSIs	35	5	10
Slide-level tumor class: tumor/no tumor	35/35	5/5	10/10
(Annotated area in percent)	(0.32/ 0.66)	(0.36/0.63)	(0.30/0.69)
Mamma-Ca dataset			
Data	Train	Valid	Test
No. of H&E ROIs	176	61	63
Slide-level tumor class: tumor/no tumor	118/162	51/58	43/57
(Annotated area in percent)	(0.72/0.28)	(0.75/0.24)	(0.75/0.25)
Slide-level TF score: other/tubular	170/60	61/26	63/23
(Annotated area in percent)	(0.90/0.10)	(0.93/0.07)	(0.92/0.08)
Slide-level NP score: 1/2/3	69/95/31	25/38/13	20/30/15
(Annotated area in percent)	(0.06/0.22/0.06)	(0.09/0.21/0.06)	(0.03/0.15/0.10)

Materials

4.1 U-Net Implementation

For the segmentation tasks, the Kaggle implementation of a U-Net with ResNet Blocks [2] was used. The network architecture is visualized in Figure 1.

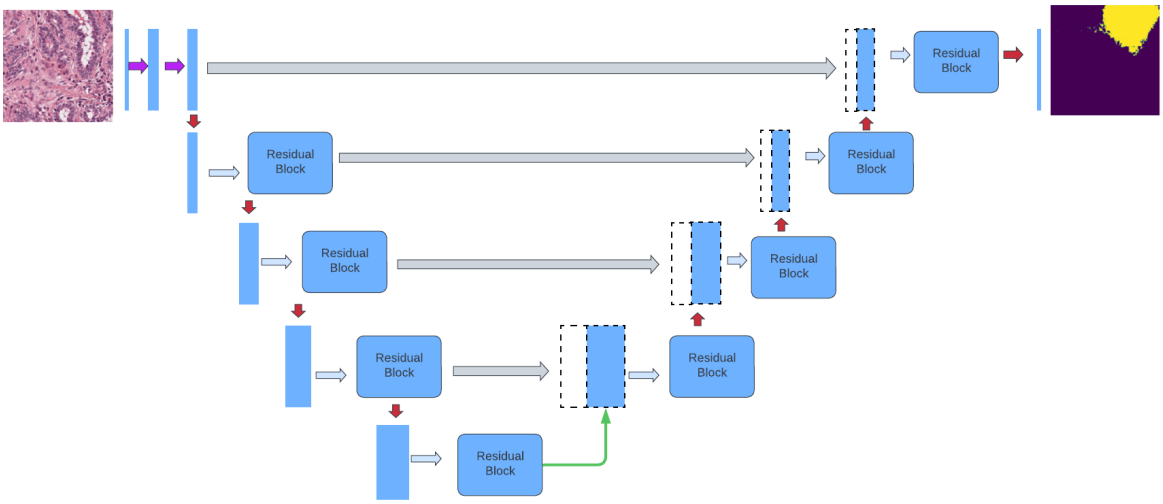

Fig. 1. U-Net with ResNet Blocks architecture adapted from [32].

4.2 Canine Cutaneous Tumor Segmentation Model

The publicly available Canine Cutaneous Tumor Segmentation Model, published by Wilm et al. [26], served as the foundation for initializing the weights of my Tumor Segmentation Baseline Model. It was initially built for the segmentation and classification of Canine Cutaneous Tumors (CCTs) and should help to further increase the robustness of my model.

Dataset To train, test and validate the Canine Cutaneous Tumor Segmentation Model, Wilm et al. [26] employed a dataset of 350 whole-slide images, containing 50 samples for "seven cutaneous tumor subtypes: melanoma, mast cell tumor (MCT), squamous cell carcinoma (SCC), peripheral nerve sheath tumor (PNST), plasmacytoma, trichoblastoma, and histiocytoma" [26]. The images were 16-fold down-sampled to a resolution of $4\frac{\mu m}{px}$

[2] https://www.kaggle.com/code/ateplyuk/pytorch-starter-u-net-resnet

(2.5X), in order to provide "more context" [26]. To determine the "background" class, Otsu's adaptive thresholding [33] was used by Wilm et al. [26].

Model architecture For the segmentation, the "fastai U-Net implementation with a ResNet18 [encoder][...] pre-trained on ImageNet "[26] was applied. The down-sampled input patches had a size "of 512×512 pixels at a resolution of $4\frac{\mu m}{px}$ (2.5X), which corresponds to a tissue size of 2048 × 2048 µm^2" [26].

Evaluation The model reached "a class-averaged Jaccard coefficient of 0.7047" [26] and a "frequency-weighted coefficient of 0.9001" [26]. When comparing the class scores, "[background and tumor] scored high Jaccard coefficients of 0.9757 and 0.9044, respectively" [26], whereas non-tumor cells were often misclassified as malignant, with a Jaccard coefficient of 0.3023 [26].

4.3 SCC dataset

The SCC dataset, which was part of the study by Wilm et al. [26] was utilized, to retrain the Canine Cutaneous Tumor Segmentation Model to perform the segmentation of the "tumor", "no tumor" and "background" classes.

4.3.1 Sample selection

The SCC dataset, which contained 50 cutaneous tissue samples of squamous cell carcinoma, was provided by "the biopsy archive of the Institute for Veterinary Pathology of the Freie Universität Berlin" [26]. It is available online and can be downloaded following the instructions in the ReadMe file of the corresponding GitHub repository[3].The slides "were routinely fixed in formalin, embedded in paraffin, and tissue sections were stained with H&E" [26]. The samples were scanned using two similar scanners at a resolution of around "0.25 $\frac{\mu m}{px}$ (40X objective lens)" [26].

4.3.2 Annotation workflow

The open-source software SlideRunner was used by one pathologist and medical students (8th semester) to annotate the images. The polygon annotations were saved as an SQLite

[3]https://github.com/DeepPathology/CanineCutaneousTumors/blob/main/README.md

database and transferred into an MS COCO (Microsoft Common Objects in Context) format using the provided Python file[4] in the repository. [26]

4.3.3 Records and classes

I divided the SCC dataset so that there were 35 WSIs in the training, 5 in the validation, and 10 slides in the test set (Table 2). All slides contained the classes "tumor" and "no tumor". Upon examining the class distribution percentages in Table 2, it is noticeable that the "no tumor" class occupied a greater area within the dataset.

4.4 Mamma-Ca Dataset

4.4.1 Sample selection

The MammaCa Dataset is my main dataset and was created for the tubule formation and nuclear pleomorphism grading algorithms. It consisted of 294 slides from 226 dogs and included breast tumor types such as: ductal tubular adenocarcinoma, ductal tubulopapillary adenocarcinoma, ductal solid adenocarcinoma, comedocarcinoma, lymphangiosis carcinomatosa, and complex adenocarcinoma. Similar to the study of Wilm et al. [26], the biopsy archive of the Institute for Veterinary Pathology of the Freie Universität Berlin prepared the H&E-stained images along with corresponding annotations within the regions of interest. The images were scanned with the Aperio Scanscope CS2 at a resolution of 0.253 $\frac{\mu m}{px}$ (40X objective lens) [26].

4.4.2 Annotation workflow

The annotations were performed by a single pathologist (C.P.) using the SlideRunner software and stored in an SQLite database.

4.4.3 Records and classes for Tumor Segmentation

The Mamma-Ca dataset was divided into three parts: the training set consisted of 170 slides, the validation set included 61 slides, and the test set comprised 63 slides (Table 2). The dataset was used to fine-tune the pre-trained SCC Baseline Model. Therefore, the annotations were grouped into the two categories "tumor" and "no tumor", while any irrelevant labels were assigned to the "other" category. The "other" category was not

[4]https://github.com/DeepPathology/CanineCutaneousTumors/tree/main/annotation_conversion

considered in the training and validation of the Tumor Segmentation Model. Compared to the SCC dataset, Table 2 indicates that the "tumor" class now occupied a considerably larger area than the "no tumor" class.

4.4.4 Records and classes for Tubule Formation Segmentation

For the segmentation of tubule formations, the super-categories "tubule" and "other" were created. The "tubule" class combined the different tubule formations inside the tumor regions: single-layer tubules (tubules that have only one cell layer), multi-layer tubules (tubules with at least 2 cell layers), and the tubulo-papillary [2]. Because, the model should distinguish between tubule formation regions and those that are not within that class, the "other" class was now relevant for the training. The tubule formations covered a smaller area than the "tumor" or "no tumor" classes from the pre-trained model. Only 109 samples contained tubule formations, as visualized in Table 2.

4.4.5 Records and Classes for Nuclear Pleomorphism Segmentation

For the segmentation of nuclear pleomorphism, I created the super-categories: "pleo 1", "pleo 2", "pleo 3" and "other". "Pleo 1" indicates regions containing "uniform or regular small nucleus and occasional nucleoli" [3]. Within regions of "pleo 2" there is a "moderate degree of variation in nuclear size and shape, hyperchromatic nucleus, and presence of nucleoli" [3]. The class with the most deformations is "pleo 3" [3]. The grades of nuclear pleomorphism were observed in the following tissue structures: single layer tubules, multi-layer tubules, tubulo-papillary and solid. Different from the Tumor Segmentation model, the "other" category was relevant in the training.

Nuclear pleomorphism, was not annotated in all slides. The "pleo 1" class was present in 114 slides, the "pleo 2" class in 163 and 59 slides contained "pleo 3" (Table 2). In contrast to that distribution, the areas that are covered by the different grades of nuclear pleomorphism were distributed differently. Although "pleo 1" appears relatively often throughout the dataset, its area made up only 6 percent of all annotated regions. Areas of "pleo 3", also made up around 6 percent. On the contrary, 22 percent of the annotated areas were labelled as "pleo 2" in the training dataset.

4.5 Computational Setup

During the training and code development for the Canine Cutaneous Tumors project, I was fortunate to receive support from the High Performance Compute System from AImotion Bavaria in Ingolstadt. The resources they provided included access to a single NVIDIA A100 GPU with 80 GB of VRAM, which was shared among the students. The GPU server was running the Linux Ubuntu operating system. The code was written in Python version 3.8.10 and utilized PyTorch version 1.12.1. The code can be found on GitHub[5].

Methods

4.6 Data Collection and Dataframe Manipulation

The process of collecting and manipulating data for this project involved two distinct datasets: the SCC dataset, obtained from the study by Wilm et al. [26], and the Mamma-Ca dataset, prepared by the Institute for Veterinary Pathology of the Freie Universität Berlin. Both datasets were annotated, but since the Mamma-CA dataset only had annotations in regions of interest (ROIs), the images were cropped accordingly. For each dataset, the polygon annotations were stored in an SQLite database, which was then converted to an MS COCO (.json) format in order to prepare the database for the dataloader. In that file, all annotation classes were saved under "categories" with their corresponding "id" and "name" from the original database. Additionally, the label "supercategories" was added to group the classes for the different training steps. Furthermore, I added an "area" label to the file to be able to analyze the size of the annotated areas. Finally, a "bbox" column was added to the new dataframe, which contained the minimum and maximum x, y coordinates for each polygon area. The bounding box is used in the sampling process to randomly select a position for the next training patch that contains a specific class.

4.7 Techniques to overcome data challenges

In the Chapter 2.4 challenges that occur when using histopathological images to develop a multi-class segmentation algorithm were mentioned. Furthermore, the Chapter 4 examined that a limited amount of images was provided and classes were unevenly distributed, which can lead to problems in the generalization of the machine learning model. In order to

[5]https://github.com/IngolstadtMedicalImaging/BiomarkersInCanineMammaryTumors

overcome these challenges and increase the probability of training an accurate model, the following techniques were applied:

4.7.1 Transfer Learning Pipeline

In my setup, I consider heterogeneous transfer learning because the datasets contain samples from the same domain (histopathological images).

There is currently no state-of-the art strategy about which layers to freeze and which layers to train in the transfer learning [34]. However, existing research has observed that features in the starting layers seem to be more generic than those in the later layers, which are then more task-specific [35]. Therefore, two steps were used for the fine-tuning. In the first step, the Pre-Training, the model should learn to segment the tumor regions. In the second step, the Head Training, all layers of the tumor segmentation model were frozen. Only the last convolutional layers (the head) of the network were trained with the task-specific datasets. The final step resulted in the Nuclear Pleomorphism Segmentation Model and the Tubule Formation Segmentation Model.

In the following, the Pre-Training and Head Training will be described in more detail. Within the segmented classes and the layers that were frozen throughout the training will be mentioned.

Baseline Model The overall goal of the Pre-Training was to train a U-Net with ResNet Blocks to segment the "tumor", "no tumor", and "background" classes. To benefit from the advantages of transfer learning, the U-Net model was initialized with the pretrained parameters from the Canine Cutaneous Tumor Segmentation Model by Wilm et al. [26]. This model was chosen as a starting point, because task and datasets were similar to my approach. The model by Wilm et al. [26] was also trained on H&E-stained histopathological images and had to segment tissue classes within cutaneous tumor subtypes [26]. However, there were some differences, specifically in terms of image resolution. The cutaneous tumor images had a resolution of $4\frac{\mu m}{px}$, whereas my mammary tumor images had a lower resolution of $0.5\frac{\mu m}{px}$ in order to learn the detailed information on a nuclear level.

Pre-Training with the SCC dataset Due to the limitation that the Mamma-Ca dataset was quite small, I decided to fine-tune the Canine Cutaneous Tumor Segmentation Model by Wilm et al. [26] with a dataset of squamous cell carcinoma first. The

SCC dataset, was already part of the network in [26], but was now resized to a lower magnification level ($0.5\frac{\mu m}{px}$). The SCC dataset was chosen due to its biological similarities to mammary carcinomas.

Transfer Learning with the Mamma-Ca dataset In the next Pre-Training step, all parameters until (including) the third residual block were frozen. The following layers were fine-tuned with the main Mamma-Ca dataset.

Final Head Training with the Mamma-Ca dataset After the Pre-Training stage, the encoder and decoder layers were frozen to proceed with the Head Training. The goal of the Head Training was to train two separate models, one for the segmentation of tubule formations and the other one for the segmentation of nuclear pleomorphism. In this step, only the weights and biases of the last convolutional layers were fine-tuned.

4.7.2 Image Augmentation and Preprocessing

The segmentation models were trained using supervised learning. Therefore, each image had a corresponding annotation mask.

The image preprocessing and augmentation approaches were an adaptation of the pipeline in the study by Wilm et al. [26], because my dataset structure and segmentation task were similar.

Preprocessing In the preprocessing step, areas without a relevant annotation for a specific training step, were labelled as "other". Additionally, the "background" class was added to the target mask.

The "background" class was necessary for the algorithm to be able to distinguish between background and tissue. It was extracted using a fixed "white" value threshold of 210. This way, pixels that had a gray value of 210 or above were considered "background" pixels.

Next, the patches for the training were prepared. Since the whole images were too big to be processed in a CNN, patches with a size of 512×512 pixels at a resolution of $0.5\frac{\mu m}{px}$ were extracted. The chosen resolution corresponded to a down-sampling rate of 2. At this magnitude, the nuclei are presented in great detail. Information about the nuclei was important for the further segmentation steps of tubule formation and nuclear pleomorphism, which is why this level had been selected.

Sampling Only patches with relevant context were extracted from the slides, to avoid patches that solely contain the "background" or "other" class label.

Classes in the training and evaluation datasets should be distributed evenly. In order to achieve this, a label was chosen randomly. Each label had the same probability of being depicted from the files that contained the label in their annotation mask. The file was selected randomly. Within that file, coordinates for the patch were chosen in an area that contained a polygon annotation of that type. The training and evaluation patch preparation happened similarly.

Augmentation Finally, patch augmentation was applied to the training dataset using common image augmentation techniques for histopathological images. This should help to avoid the overfitting on the training patches. Spatial augmentations, like horizontal flip, vertical flip, and 90-degree rotations, were used to enhance robustness in terms of shaping with a probability of $p = 0.75$. To enhance robustness in terms of deformation, transformations such as: optical distortion, grid distortion and affine transformations were applied with a probability of 0.75 for all datasets, except for nuclear pleomorphism, as those scores measure the deformation. To increase color and staining diversity, the HedLighterColorAugmenter by Otálora et al. [36] for histopathology images ($p = 0.3$) was used. Finally, all patches were normalized with the same RGB statistic that had been applied in [26].

Addressing Tubule Formation Segmentation challenges Tekin et al. [16] were able to improve their results in the segmentation of tubule formations, using a mirror technique. I decided against this approach within this work. Reasons are, that a mirror technique leads to unnatural shapes of the tubule areas, on the one hand. On the other hand, complete structures can still not be guaranteed every time. Within my dataset, tubules close to each other were sometimes combined into one annotation area, which would make a mirror technique more inaccurate. Instead, my sampling process should ensure that entire tubule formations were located in the middle of a patch, so that there were less incomplete forms.

Addressing Nuclear Pleomorphism Scoring challenges Recent studies that focus on the scoring of nuclear pleomorphism, adjusted the scores in the training process. Mercan et al. [17] transformed the discrete labels into a continuous figure, while Wetstein et al.

[24] decided to translate the grading schema into a binary classification. For my study, the dataset provided labels only from one pathologist, which is why no continuous label could be calculated. Furthermore, in order to follow the grading scheme by Peña et al. [2], the traditional score was needed.

4.8 Evaluation Metrics

Metrics for Training The multiclass Cross-Entropy Loss is computed for the output of the model to compare a prediction against the ground truth. The overall goal of the network is to reduce the loss. The mean Cross-Entropy Loss function, $meanCE$, is defined as follows:

$$meanCE = -\frac{1}{n} \cdot \sum_{i=1}^{n} [y_i * ln(S(\hat{y}_i))]$$

Within that formula n is the total number of inputs and y is the desired output. $\hat{y}$ is the prediction of the model and S stands for the Softmax probability that is calculated from $\hat{y}$.

The final loss function includes Cross-Entropy Loss on the one hand and Jaccard Loss on the other. The Jaccard index [37] calculates the similarity between classes. An index of 0 indicates no overlap between the classes, whereas an index of 1 means the output mask is identical to the expected output. The following formula is used to measure the Jaccard similarity between the output mask A and the expected mask B. The Jaccard similarity is also known as intersection over union (IoU):

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

The final costs of this network are computed as the sum of $meanCE$ and the inverted Jaccard similarities J between the predicted class J and C_{pred} ground truth class C_{gt} for all classes c.

$$C(f) = \alpha \cdot meanCE + \beta \cdot \sum_{c} (1 - J(C_{pred}, C_{gt}))$$

In order to put more emphasis on the rare samples in my dataset, I also used the Focal Loss instead of $meanCE$. Focal Loss is an adaptation of Cross-Entropy Loss and was used for the segmentation of nuclear pleomorphism. The parameter γ can be tuned in order

to put more focus on poorly classified classes. The probability that is predicted by the model is p_i. [38]

$$FocalLoss = -\sum_{i=1}^{n}(1-p_i)^{\gamma}\log(p_i)$$

Metrics for Testing To describe the performance of the multiclass classification model, I included confusion matrices in the evaluation. The rows of the matrix correspond to the actual classes, and the columns correspond to the predicted classes. Accuracy is a measure of how many of the samples were correctly classified overall. It is calculated as the sum of the diagonal entries of the confusion matrix (the number of true positives for each class) divided by the total number of samples.

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

Another measure that indicates how many of the predicted samples actually belong to a particular class is the precision. It is calculated as the number of true positives (samples correctly predicted to be in a class) divided by the total number of predicted positives (samples predicted to be in the class, whether correctly or incorrectly).

$$Precision = \frac{TP}{TP+FP}$$

Furthermore, Recall is another measurement that can be computed from a confusion matrix. It measures the ability of the classification model to identify all pixels of a target class in a dataset. Specifically, recall represents the proportion of true positive predictions (i.e., instances of the target class that were correctly identified by the model) out of all instances of the target class in the dataset, both true positives and false negatives (i.e., instances of the target class that were incorrectly classified as a different class by the model).

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

4.9 Model Training

In the following sections, the hyperparameter settings of the different segmentation models will be presented.

4.9.1 SCC Baseline Model

Dataset The SCC Baseline Model was trained using the SCC dataset. Because it consisted of more annotated regions than the initial Mamma-Ca dataset, it was used to pretrain the Tumor Segmentation Model. To instantiate the weights, the Canine Cutaneous Segmentation Model developed by Wilm et al. [26] was used.

The output layer had four dimensions: "tumor", "no tumor", "background" and a fourth "other" class. Here, "other" included all non-annotated areas and was ignored within the training and evaluation.

For the training, a total of 50 patches of 512×512 pixels were extracted from each of the 35 whole-slide images, which were downsampled to a resolution of $0.5\frac{\mu m}{px}$. The downsampling was performed, because the further training steps required a high resolution of $0.5\frac{\mu m}{px}$, to properly display the detailed nuclear and tubule formations. The training dataset consisted of 1750 patches, and updates were performed on mini-batches of seven patches. Due to the relatively small batch size, I decided to freeze the Batch Normalization Layers for testing purposes first, to determine if the results would be more accurate. However, when comparing the results, it was observed that the results improved when including the Batch Normalization Layers. Therefore, in the final training, the Batch Normalization Layers were not frozen. Furthermore, all patches were pre-processed and sampled using the methods described in Chapter 4.7.2. All classes were sampled with the same probability. As illustrated in Table 3 the training dataset then consisted of 0.06, 0.43, 0.2 and 0.31 percent of the classes "other", "background", "tumor" and "no tumor".

Table 3. Class-wise distribution of pixels in the training samples for the SCC Baseline Model.

SCC Baseline Model		
class	pixels	percentage
other	25.428.207	0.06
background	198.980.197	0.43
tumor	91.872.300	0.2
no tumor	142.209.152	0.31

Hyperparameter setting To achieve the goal of minimizing the cost function to a local minimum, the training process utilized stochastic gradient descent (SGD) with a learning rate of 0.001 and a momentum parameter of 0.9. The loss function employed was a combination of Cross-Entropy Loss multiplied by 1.0 and intersection over union (IoU) loss from all 3 relevant classes multiplied by 0.5. The chosen hyperparameters, along with the SGD optimizer, served to optimize the weights of the model throughout the training process. To address the issue of fluctuating loss updates on mini-batches, momentum was incorporated into the optimizer, facilitating a better convergence of the model.

4.9.2 Mamma-Ca Baseline Model

Dataset In this study, the Mamma-Ca dataset was utilized to fine-tune the pretrained SCC Baseline Model. In order to preserve the generalization performance of the pretrained SCC model, the U-Net parameters of all layers until (including) the third residual block were frozen. The same classes as in the SCC Baseline Model ("tumor", "no tumor", "background", "other") were segmented. The size of the patches, which were extracted from a total of 170 ROIs, was 512×512. Because the ROIs were much smaller than the WSIs, only 7 patches per image could be extracted. Consequently, an epoch consisted of 1232 patches, that were again grouped into batches of seven samples. Again, the sampling procedure happened as described in Chapter 4.7.2, while all classes had the same probability to be in the dataset. The training dataset then consisted of 0.16, 0.18, 0.49 and 0.16 percent of the classes "other", "background", "tumor" and "no tumor" pixel areas as visualized in Table 4.

Table 4. Class-wise distribution of pixels in the training samples for the Mamma-Ca Baseline Model.

Mamma-Ca Baseline Model		
class	pixels	percentage
other	52.266.285	0.16
background	59.528.506	0.18
tumor	159.540.506	0.49
no tumor	51.363.967	0.16

Hyperparameter setting To calculate the loss, the Cross-Entropy Loss function was utilized, along with an IoU loss from the three relevant classes multiplied by 0.3. The impact of the IoU loss should be smaller, because it can become less accurate in regions with more "other" areas. Specifically, when a prediction is performed within an "other" area that has no annotation, even if a prediction is correct, the IoU would be low, because the ground truth label is "other". Compared to the SCC dataset, the Mamma-Ca dataset contains more "other" areas. Additionally, a learning rate of 0.0001 to be multiplied with the loss gradient was defined. Furthermore, stochastic gradient descent (SGD) with a momentum rate of 0.95 was applied.

4.9.3 Tubule Formation Segmentation Model

Dataset After the Pre-Training, all the base layers of the Tumor Segmentation Model were frozen, and only the parameters within the model's head, the last two convolutional layers of the U-Net, were allowed to be updated. This approach should enable the model to first differentiate between "tumor" and "no tumor" areas and subsequently, segment the tubule formations within the tumor region in the last layers. The final two convolutional layers of the model were trained to accurately segment two distinct classes, namely, "tubular" and "other". Notably, a separate background class was not explicitly extracted, as the tubule formations contain a lumen that belongs to the tubular structure and should not be erroneously classified as background. Therefore, the "other" class included both the surrounding tissue and background areas. This time the "other" class, representing all non-tubular areas, was included into the training. Regarding the patch sampling process, seven patches of size 512×512 were extracted from each image and batches of size eight were used during the training process. To avoid redundant patches, they were exclusively sampled from areas labeled "tubule" since the "other" regions are present in every patch. Per epoch, a total of 1232 patches were utilized during the training process. In detail, the training dataset then consisted of 0.58 and 0.42 percent of the classes "other" and "tubular" as visualized in Table 5. All classes were sampled with the same probability.

Table 5. Class-wise distribution of pixels in the training samples for the Tubule Formation Segmentation Model.

Tubule Formation Segmentation Model		
class	pixels	percentage
other	134.551.743	0.58
tubular	95.872.833	0.42

Hyperparameter setting Optimizer and Loss Criterion were the same as in Chapter 4.9.2, this time enhanced with a momentum rate of 0.9. Furthermore, a learning rate of 0.001 was used and a weight decay rate of 0.001 was added to avoid overfitting.

4.9.4 Nuclear Pleomorphism Segmentation Model

Dataset For the segmentation of nuclear pleomorphism, the smallest amount of data was available. To avoid overfitting, the transfer learning approach, freezing all the base layers of the pre-trained Mamma-Ca Baseline Model and only fine-tuning the parameters of the last two convolutional layers, was adopted. As in the segmentation of tubule formations, the model should first separate "tumor" and "no tumor" areas and then segment the nuclear pleomorphism within the tumor region.

This time the dataset was separated in five classes, including the different grades of nuclear pleomorphism ("pleo 1", "pleo 2" and "pleo 3"), "background" and "other".

The patch and batch creation happened in the same way as for the Tubule Formation Segmentation Model and resulted in a total of 1232 patches per epoch. Since "pleo 2" regions made up a 4 times larger area than "pleo 1" and "pleo 3", the minority classes in the segmentation were up-sampled. Therefore, weights were applied to each class to be chosen in the sampling process. Weighting factors were defined as the following: "pleo 1" = 0.8, "pleo 2" = 0.5, and "pleo 3" = 0.7. Consequently, the training dataset consisted of 0.36, 0.14, 0.12, 0.21 and 0.17 percent of the classes "other","background", "pleo 1", "pleo 2" and "pleo 3" as visualized in Table 6.

Table 6. Class-wise distribution of pixels in the training samples for the Nuclear Pleomorphism Segmentation Model.

Nuclear Pleomorphism Segmentation Model		
class	pixels	percentage
other	83.186.285	0.36
background	31.742.415	0.14
pleo 1	27.469.555	0.12
pleo 2	48.655.139	0.21
pleo 3	39.371.182	0.17

Hyperparameter setting The Hyperparameters were the same as in the Tubule Formation Segmentation Model training. This time, the Cross-Entropy Loss was exchanged with Focal Loss to put more emphasis on classes with a low IoU score. Again, the IoU score was added with a weighting factor of 0.3.

5 Results

This chapter presents the performance of the trained models. Each algorithm was trained for 25 epochs or until convergence. Finally, the model with the lowest validation loss was stored. The chapter is organized into five sections. The first two sections will present the baseline models, that were trained to segment "tumor", "no tumor" and "background" areas, while the following two sections will provide the results for the Tubule Formation Segmentation and Nuclear Pleomorphism Segmentation Models. The evaluation of each model includes the IoU scores achieved for each class, the confusion matrix and further performance measures of the model. The chapter concludes with example segmentation masks produced by each model. Example images are presented in patch size and also merged together into the initial image size, along with an implementation of the scoring.

5.1 SCC Baseline Model

The SCC Baseline Model achieved a test IoU of 0.58 for all classes combined. Separate classes achieved IoU scores of 0.89, 0.38, and 0.46 for "background", "tumor" and "no tumor", respectively (Table 7). When examining the confusion matrix in Figure 2, it became apparent that the "background" class had the highest level of overlap between predicted and ground truth labels along with a high precision (0.91) and recall score (0.95). In contrast, the "tumor" and "no tumor" classes were mixed up in certain areas, which resulted in lower precision and recall scores. Although some pixels of the "tumor" and "no tumor" classes were misclassified, the accuracy reached a score of 73.19 percent.

5.2 Mamma-Ca Baseline Model

With the Mamma-Ca Baseline Model, an overall IoU of 0.44 and separate IoU scores of 0.52, 0.57, and 0.24 for "background", "tumor" and "no tumor" were reached on the test dataset (Table 7). Comparing the different classes, it was noticeable that the "tumor" class reached the highest IoU score and had the highest overlap with the ground truth pixels, visible in the confusion matrix (Figure 3). Nevertheless, the "tumor" label was often incorrectly assigned to the other classes as well. This can also be noticed in the high recall of 0.91, but in contrast a lower precision of 0.49 for the "tumor" class.

Similar to the SCC Baseline Model, the "background" predictions achieved a high degree of overlap with the ground truth regions (precision = 0.94). The "no tumor" class

Table 7. Class-wise evaluation measures per model on the test dataset on patch-level.

SCC Baseline Model					
Segmented class	IoU score	Overall IoU score	Precision	Recall	Overall Accuracy
background	0.89	0.58	0.91	0.95	73.19%
tumor	0.38	-	0.65	0.59	-
no tumor	0.46	-	0.63	0.66	-
Mamma-Ca Baseline Model					
Segmented class	IoU score	Overall IoU score	Precision	Recall	Overall Accuracy
background	0.52	0.44	0.94	0.61	62.81%
tumor	0.57	-	0.49	0.91	-
no tumor	0.24	-	0.74	0.37	-
Tubule Formation Segmentation Model					
Segmented class	IoU score	Overall IoU score	Precision	Recall	Overall Accuracy
other	0.56	0.48	0.60	0.79	63.5 %
tubular	0.39	-	0.70	0.48	-
Nuclear Pleomorphism Segmentation Model					
Segmented class	IoU score	Overall IoU score	Precision	Recall	Overall Accuracy
other	0.4	0.25	0.31	0.80	37.39%
background	0.62	-	0.98	0.66	-
NP score 1	0	-	0	0	-
NP score 2	0.21	-	0.24	0.42	-
NP score 3	0	-	0	0	-

was harder to identify by the model, with a recall of 0.37, instead the majority of "no tumor" labels were labeled as "tumor".

5.3 Tubule Formation Segmentation Model

On the test dataset, an overall IoU of 0.48 was reached with the Tubule Formation Segmentation Model. The two classes "other" and "tubule" achieved IoU scores of 0.56 and 0.39, respectively (Table 7). The recall of 0.48 shows, that the model had problems finding tubule formations, but reached a higher precision of 0.70 when predicting the "tubule"

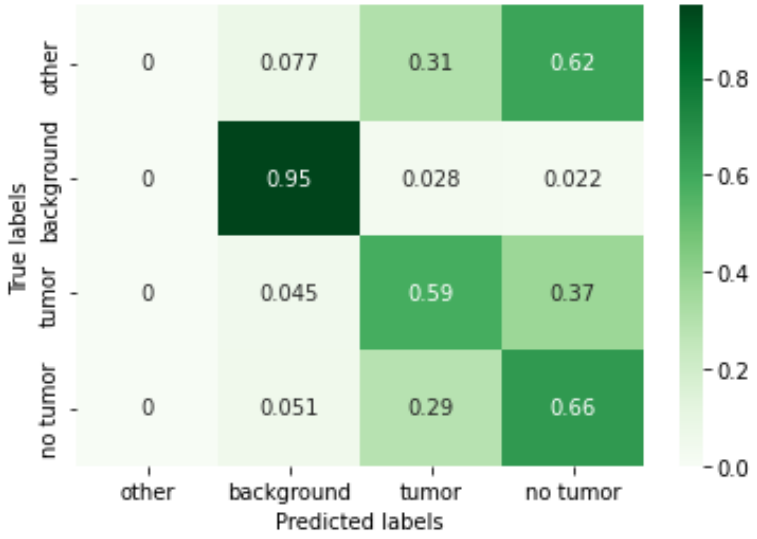

Fig. 2. Confusion matrix of the SCC Baseline Model that segments "background", "tumor" and "no tumor" class. The "other" class was ignored in the training and evaluation. The results are normalized between 0 and 1 along the true labels.

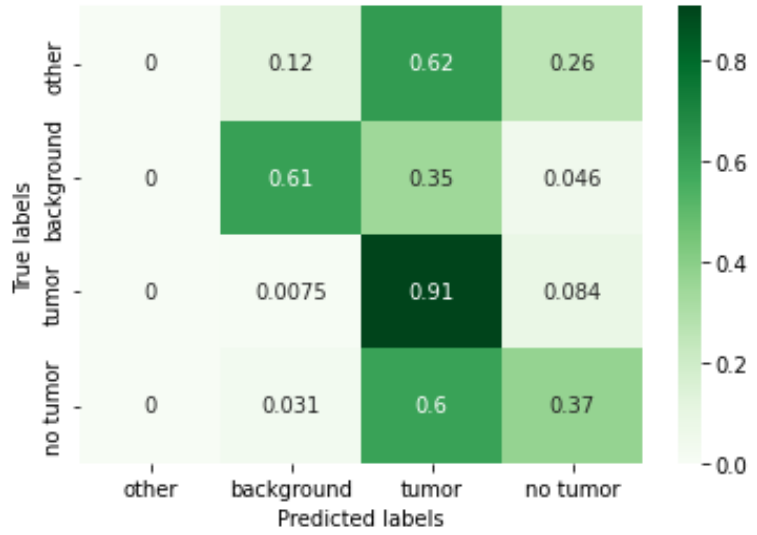

Fig. 3. Confusion matrix of the Mamma-Ca Baseline Model that segments "background", "tumor" and "no tumor" class. The "other" class is ignored in the training and evaluation. The results are normalized between 0 and 1 along the true labels.

class. Nevertheless, the precision score for the "other" class was only 0.60, because tubule formations could not be found in many cases. This can be observed in Figure 4.

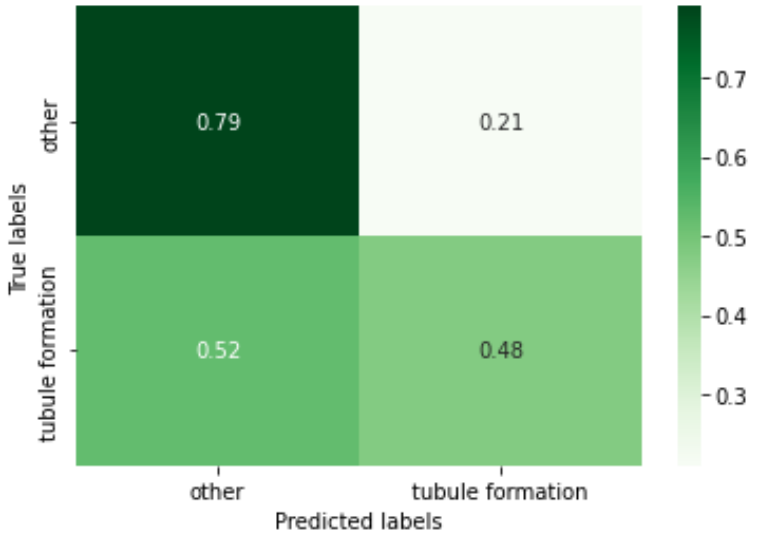

Fig. 4. Confusion matrix of the Tubule Formation Segmentation Model that segments "other" and "tubule formation" classes. The "other" class contains background and tissue areas. The results are normalized between 0 and 1 along the true labels.

5.4 Nuclear Pleomorphism Segmentation Model

The Nuclear Pleomorphism Segmentation Model reached an overall IoU score of 0.25 and the scores of 0.4, 0.62, 0.00, 0.21 and 0.00 for the classes: "other", "background", "pleo 1", "pleo 2", "pleo 3" (Table 7). Both, "pleo 1" and "pleo 3" classes showed a very low IoU of less than 0.01 in the test dataset, while in the training dataset, the IoU score for "pleo 1" remained close to zero, and for "pleo 3", there were some peaks at an IoU of 0.04, despite being close to zero as well. Contrary, the nuclear pleomorphism classes were mostly assigned to the "other" or "pleo 2" label. With a precision of 0.24 the "pleo 2" class was not often identified correctly (Figure 5). Regarding transfer learning, the pretraining process helped the model to converge quickly, but it stagnated after ten epochs. Since the "pleo 1" class predictions did not improve, the training was not continued after 25 epochs.

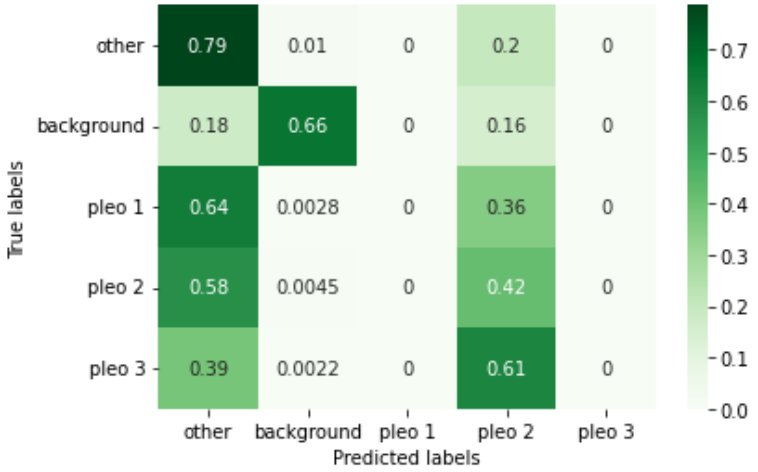

Fig. 5. Confusion matrix of the Nuclear Pleomorphism Segmentation Model that segments "other", "background", "pleo 1", "pleo 2", "pleo 3" classes. The results are normalized between 0 and 1 along the true labels.

5.5 Example Patches and Grading Results

5.5.1 Tubule Formation

In order to grade tubule formations based on the grading scheme proposed by Peña et al. [2], it is necessary to determine the percentage of tubule formation within each whole slide image. However, since the model only accepts inputs at the size of 512×512 pixels, the samples were divided into separate patches. To quantify the tubule formation areas relative to "other" pixels on whole ROIs, the prediction masks of each patch got merged together again. Due to the receptive field of convolutional neural networks, pixels in the center of the image receive more attention during the prediction process. Therefore, an overlap of 50 pixels was applied when combining the patches into the prediction mask for the whole image. Given the entire sample, the percentage of the tubule formation area could be calculated and the corresponding grade could be assigned. Examples of the prediction masks and grades are presented in the following two paragraphs:

Predictions on single patches In the evaluation of the Tubule Formation Segmentation model, it was noticeable that problems with the segmentation of tubules occurred. Tubule formations were correctly identified in less than 50 percent of the cases. The output images in Figure 6 show, that tubule formation areas with a large lumen could be iden-

tified, whereas tubule structures without or with small lumen were ignored. Furthermore, areas that contain background, but are not a tubule formation were often misclassified as "tubular".

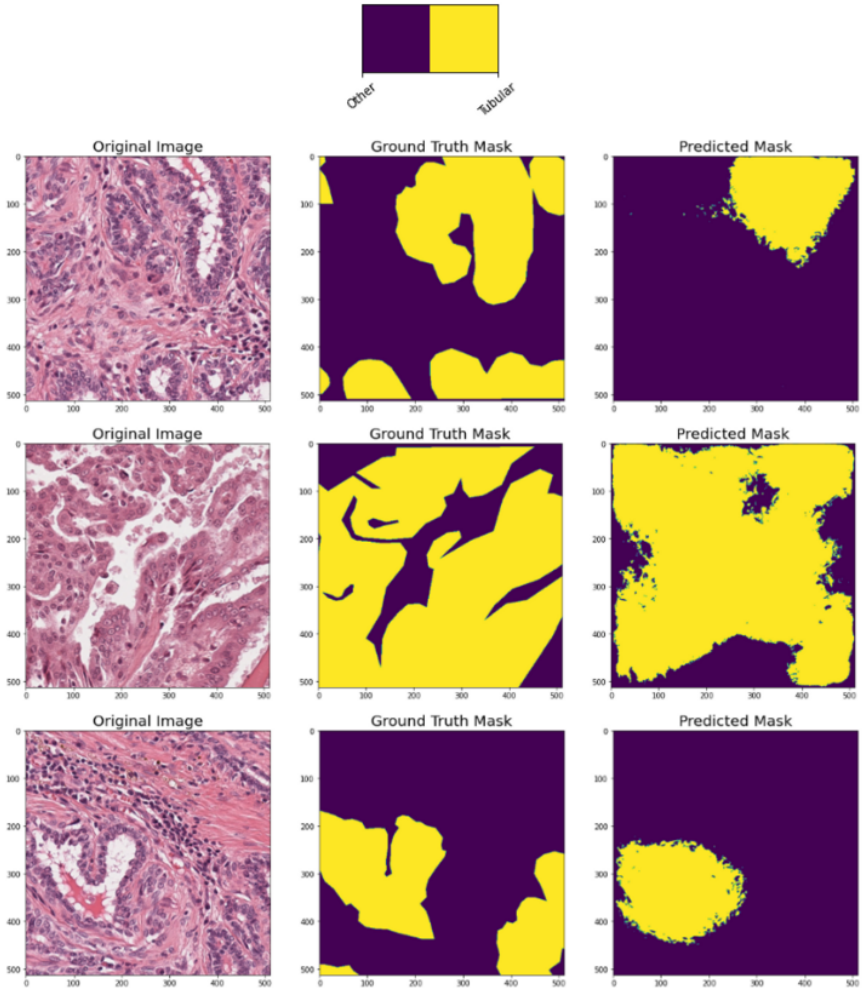

Fig. 6. Example segmentation results of the Tubule Formation Segmentation Model on the test dataset on 512×512 patches. The first image shows the original patch, followed by the expected output mask and the actual prediction mask.

Grading on whole slides Likewise, when combining the 512×512 patches into the full prediction mask, the algorithm achieved only low accuracy. In order to test the grading pipeline, I visualized the outputs, which yielded the example images in Figure 7.

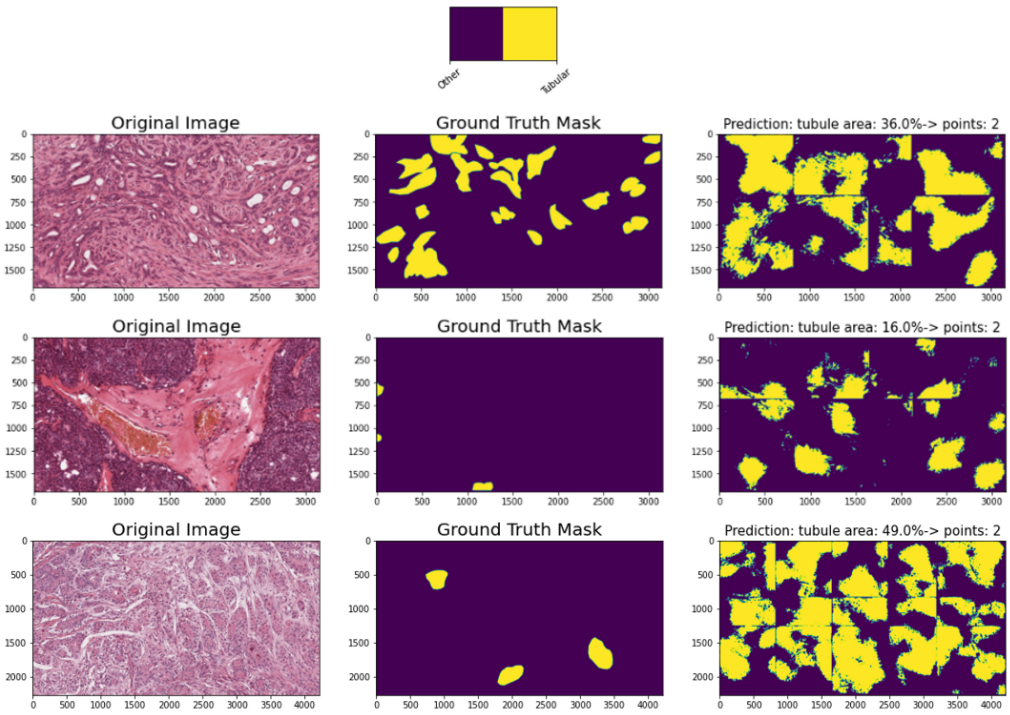

Fig. 7. Example segmentation results of the Tubule Formation Segmentation Model on the test dataset on the whole slide. The first image shows the original patch, followed by the expected output mask and the actual prediction mask. In the ground truth masks, tubules were only annotated within tumor regions. The percentage of tubule formation within the tumor region yields the grading points [2]. The Tubule Formation Segmentation Model achieved low accuracy on the test data, which is why the prediction masks contain wrongly classified pixels.

5.5.2 Nuclear Pleomorphism

In order to define the nuclear pleomorphism grades, the Nuclear Pleomorphism Segmentation Model was trained. The scores had been assigned by an expert following the grading scheme by Peña et al. [2]. Consequently, the supervised learning model should learn to segment relevant tissue areas and assign a score in the range of 1-3.

Predictions on single patches Unfortunately, the accuracy for the nuclear pleomorphism results was low and the classes could not be separated well. In most cases, the predicted mask assigned the "pleo 2" label. "pleo 1" and "pleo 3" classes, that occurred less

in the dataset, were mostly ignored by the segmentation network. Examples of the output patches compared to the ground truth annotation are presented in Figure 8.

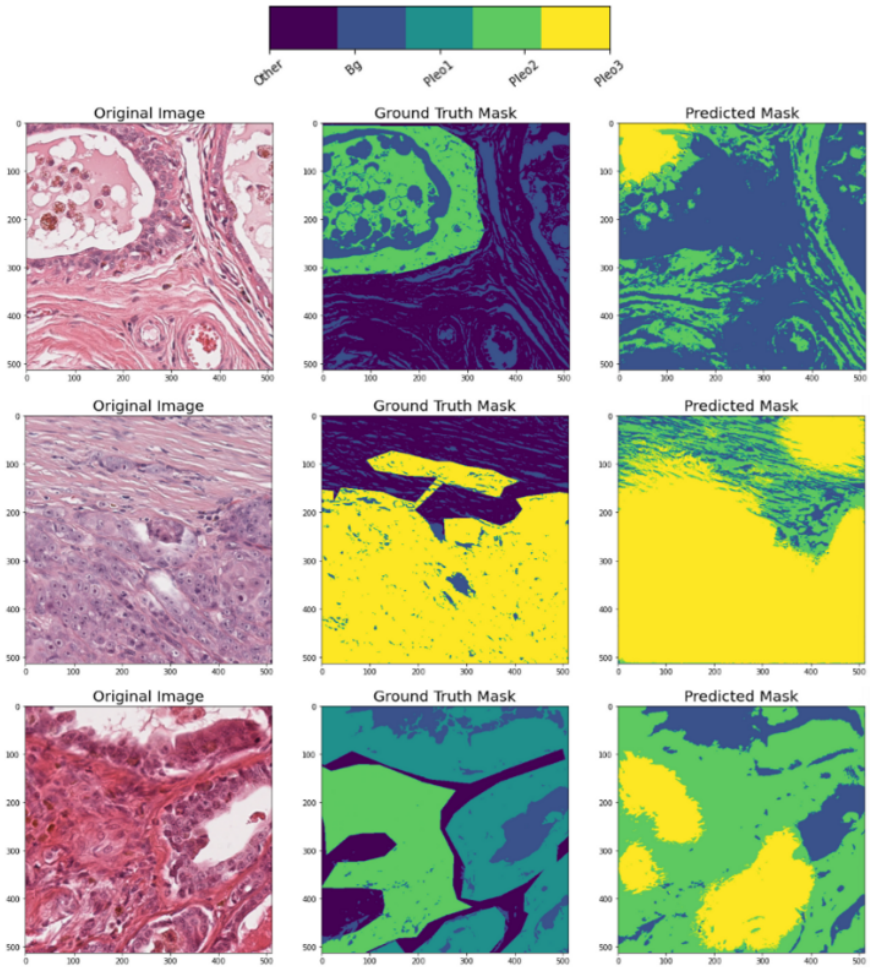

Fig. 8. Example segmentation results of the Nuclear Pleomorphism Segmentation Model on the test dataset on 512×512 patches. The first image shows the original patch, followed by the expected output mask and the actual prediction mask.

Grading on whole slides Although the predictions did not match with the ground truth masks in many cases, the grading logic was still implemented. In order to define the final grade, a list of all nuclear pleomorphism scores appearing in the image is created from the final prediction mask. Subsequently, the maximum value within that list is returned as the corresponding score above the output image in Figure 9.

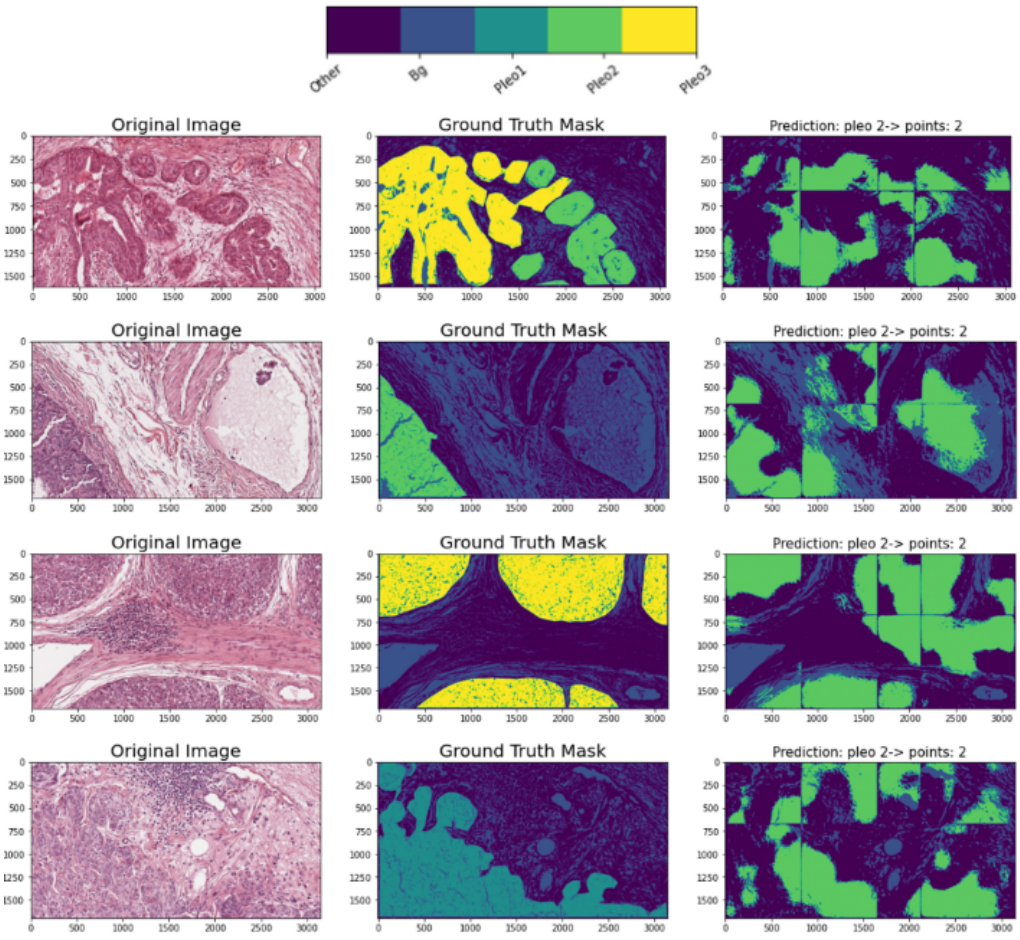

Fig. 9. Example segmentation results of the Nuclear Pleomorphism Segmentation Model on the test dataset on the whole slide. The first image shows the original patch, followed by the expected output mask and the actual prediction mask. The maximum value within the mask yields the final score [2]. The Nuclear Pleomorphism Model could not learn the pleomorphism grades 1 and 3 properly, which is why the output is always grade 2.

6 Discussion

This section reviews the methods to improve the performance of each segmentation model and the final results achieved. Moreover, the limitations and problems of this work and what could be changed in future works will be presented. For the main grading tasks, the results will be compared to previous studies.

SCC Baseline Model The SCC dataset was the first dataset which was used to train a model for the segmentation of the "tumor", "no tumor" and "background" classes. Having the weights initialized with the weights of a prior segmentation model, that already included the SCC dataset, an improvement in IoU scores was expected. Due to the fact that the initial model, which was used as a baseline, had a low IoU score for the "no tumor" class, one could have expected the model to overfit towards "tumor", which did not happen. In contrast to the expected outcomes, the overall IoU score was lower, but the score improved for non-tumorous tissue. With the help of transfer learning and pre-initialized weights, the "background" class achieved high precision and recall values. Nevertheless, the IoU score of 0.89 was lower than in the initial Canine Cutaneous Tumor Segmentation Model [26] (IoU = 0.97). An explanation for this behavior could be the different ways the "background" class was determined. In the SCC Baseline Model training, a fixed white threshold value of 210 was chosen, whereas Wilm et al. [26] extracted the "background" class with the help of Otsu's adaptive thresholding. Consequently, the white was not constant. Furthermore, I applied color augmentations, that could have also increased difficulty in finding the background class. The network also had difficulties differentiating between "tumor" and "no tumor" classes. The IoU score for non-tumorous tissue improved from 0.30 to 0.46 compared to the initial model, while the "tumor" class IoU decreased from being 0.90 to a score of 0.38.

One reason for the difficulty to distinguish between the 2 classes could be the change in resolution. The new SCC images had an 8-times higher resolution of $0.5\frac{\mu m}{px}$ than the samples of the baseline model. Because the tissue regions were available in greater detail now, the network had to update its parameters and lost the previously learned accuracy.

Furthermore, the U-Net architectures were different. The Canine Cutaneous Tumor Model was build using a fastai U-Net implementation with a ResNet18 encoder, while my approach utilized another U-Net adaptation with residual blocks. Therefore, some weights could have been irritating and needed to be changed in the back-propagation process, which could have led to lower accuracy too. It is not yet clear whether the output IoU would have been better without setting the baseline weights, and thus avoiding a shift in resolution and a different architecture.

Mamma-Ca Baseline Model After the fine-tuning step, the transfer method was used. Therefore, only the last layers of the SCC Baseline Model were fine-tuned, this time using the Mamma-Ca dataset.

Although the Mamma-Ca Baseline Model was initialized with the parameters of a very similar segmentation task, it only achieved a Jaccard score of 0.44 for all segmented classes combined. That means that the overall IoU score decreased from being 0.58 in the SCC Segmentation Model. Considering only the values of the Jaccard score, the model seems to give quite good IoU scores for "background" (0.52) and "tumor" (0.57). In contrast, "no tumor" predictions did not overlap with the expected output often (IoU = 0.24). Compared to the SCC Baseline Model, despite the"tumor" class, all IoU scores got worse. When evaluating the confusion matrix in Figure 3 it becomes clear, that the model was overfitting towards the "tumor" class. In fact, in most cases it labelled tissue areas, that did not belong to the background class, as tumorous. This can also be noticed at a recall score of 0.91, but only a precision of 0.49 for the "tumor" class, while non-malignant tissue only reached a recall of 0.37. Because the data analysis in Chapter 4.4.3 revealed that the dataset contained larger areas of tumorous tissue than non-tumorous, equal sampling was applied to avoid overfitting towards the predominant class. However, an explanation for the low accuracy towards the "no tumor" class, may be the size of areas the two classes covered in the dataset. Table 2 indicates that an around three times higher percentage of "tumor" area covered the images used for training, validation and testing. This leads to the assumption, that although a sample was extracted, because it contained the "no tumor" class, in most cases it also contained parts of malignant tissue. In more detail this can be seen in Table 4, which shows that despite an equal sampling method, the percentage of "tumor" pixels was about three times higher (49 %) than for the "no tumor" area (16%), what could be the reason for the overfitting. Due to this limitation of "no tumor" areas, better results could not be achieved with this approach. In future works, a larger and more variable dataset could prevent an overfitting towards the majority class. Furthermore, an oversampling of the minority class could also increase the performance. Another technique could be the use of cost-sensitive learning, which assigns different misclassification costs to different classes to reflect the importance of each class.

Tubule Formation Segmentation Model With the pre-trained Tumor Segmentation Model, good results when fine-tuning the last layers for the Tubule Formation Segmentation Model were expected. Within the first layers the model should distinguish between "tumor", "no tumor" and "background" areas and learn to segment tubule areas afterward. The overall Jaccard score on the test dataset was 0.48. The two classes, "other" and "tubule", reached scores of 0.56 and 0.39, respectively. The "other" class reached a recall

of 0.79, whereas "tubule formation" reached only 0.48, because tubule formation regions were often misclassified as "other". When reviewing the confusion matrix in Figure 4, the model seemed to be biased and preferred predicting the "other" class in most cases. This could be, because the model is underfitting. Another reason could be, that the tubule formations were only annotated within tumor regions. Consequently, if the model recognized a tubule formation outside a tumor area, it would be considered as false. However, the patches for training, validation and testing were sampled from only relevant areas. Hence, missing annotations outside of tumor areas should not have a high impact on the results within patch size.

Another reason could be, that the model had problems distinguishing between lumen and background, as they are both white. Looking at the example patches in Figure 6, the model seems to find tubule formations, with a large lumen and misses tubule formations, that have only a small lumen area. Furthermore, when evaluating the segmentation masks on the whole images in Figure 7, it looks like the model focused on the background area when segmenting the tubule. Areas that were misclassified as tubule formations mostly contained a white region within. As my images only contained regions of interest on a whole slide, there were only a few ROIs that contained large areas of background, so that the model could not identify background areas properly. As a consequence, in the Tubule Formation Head Training, the model could have mistaken background pixels for lumen.

Furthermore, there are many shapes of tubule formations (e.g., single-layer tubules, multi-layer tubules and tubulo-papillary) in the dataset, but only a small amount of examples per shape. Consequently, the model could not learn the different shapes of tubule formations properly and was rather focusing on the lumen, which most tubule formations contain. Compared to the previous work by Jaroensri et al. [18], that relied on 499 images for the training, my training dataset was limited to 60 images with tubule formations. Another factor that might have contributed to the accuracy of predictions is the way tubule formations were annotated within the training dataset. In contrast to previous studies (e.g., [16]), that circled each tubule on its own, tubule formations that were next to each other in my dataset, were annotated as one region. This issue and the aforementioned problem of missing examples could probably be resolved with a larger dataset. Furthermore, my pre-trained model was not accurate in distinguishing tumor and non-tumor regions, it segmented tubules in healthy cells too. This is a problem when defining the grade, because only the tubule formations within the tumor region are rele-

vant for grading. Regarding the limitation of implementation time, it was not possible to implement anymore, but in the future the tumor regions could be extracted beforehand, so that the tubule formations will only be segmented in this area.

Nuclear Pleomorphism Segmentation Model Similar to the Tubule Formation Segmentation Model, only the last layers of the Mamma-Ca Baseline Model were trained for the segmentation of nuclear pleomorphism scores. Again, within the first layers the model should distinguish between "tumor" and "no tumor" areas and learn to segment nuclear pleomorphism areas within the tumor, afterward. Regarding the segmentation of nuclear pleomorphism, compared to the Mamma-Ca Baseline Model, the "background" IoU had further improved, now reaching an IoU of 0.62 instead of 0.44. Nevertheless, the algorithm had big problems distinguishing between the "other" and nuclear pleomorphism classes.

In the confusion matrix in Figure 5, it becomes apparent that the algorithm has a relatively high recall value for the "other" class (0.80), and the "background" class (0.66). However, many false positives were observed, with classes misclassified as "other". Consequently, the precision for "other" was only 0.31. The "pleo 2" class had a precision of only 0.24. False positive values often occurred, because the model could not properly distinguish between the pleo scores and was labelling every nuclear structure as "pleo 2". The IoU for the other pleomorphism scores was lower than 0.01 for the test dataset. The low performance of the model in accurately detecting the "pleo 1" and "pleo 3" classes is most probably attributed to the limited number of samples. In Table 2 it is noticeable, that only 195 regions of interest for the different nuclear pleomorphism scores 1,2 and 3 were provided, whereas another recent study by Jaroensri et al. [18] relied on 498 slides containing nuclear pleomorphism. For the nuclear pleomorphism scores, that are difficult to identify even by experts, a model would need more data to learn from. Furthermore, there are significant differences between the classes of nuclear pleomorphism, such as single-layer tubules, multi-layer tubules, tubulo-papillary structures, and solid structures. Since the Tubule Formation Segmentation model already had problems identifying tubule areas, the Nuclear Pleomorphism Model with even more shapes and classes did not perform better. Two methods that were applied in addition to transfer learning in order to overcome the issue of limited samples, were Focal Loss and oversampling techniques applied on minority classes. However, the results did not improve in the test dataset. To improve the models' performance in the future, increasing the number of samples would be necessary. Additionally, as described in the tubule formation segmentation approach, it would also be

necessary to first extract tumor regions and then look for nuclear pleomorphism within, because the annotations can only be found in malignant areas.

For the nuclear pleomorphism it is important to note, that due to inter-observer variability it could be important to label the dataset by different pathologists, to create more stable results, which was performed in a recent study by [17].

7 Conclusion and Outlook

The goal of this thesis was to build a classification model that segments the two histopathological markers, tubule formation and nuclear pleomorphism, on histopathological images of Canine Mammary Carcinoma.

Canine Mammary Tumors are common in female dogs and can be caused by various factors. Besides other evaluation options, the histopathological examination of the tumor is a critical component in defining the tumor type, grade and stage. Although grading schemes exist, there is still an inter-observer variability, when identifying such features. With the new field of digital pathology, it is now possible to scan tissue samples and utilize them to develop classification algorithms, to aid pathologists when grading tumors.

Compared to other studies, my algorithm had a relatively small set of training data and thus the first research question was how to overcome this limitation by using transfer learning, image augmentation and image preprocessing. The second question was how precise the segmentation results would be and whether they are accurate enough to be used by a pathologist. To overcome the limitation of a smaller dataset I built a transfer learning pipeline, that was composed of 2 different datasets and a network trained within another study by Wilm et al. [26], to initialize the weights for the first training. The first training step was to segment "tumor", "no tumor" and "background" classes. Therefore, a squamous cell carcinoma dataset and my mammary carcinoma samples, were used. In the second step, only the last convolutional layers of the Tumor Segmentation Model were trained, while the other layers were frozen. The training resulted in two different models, one to segment tubule formations and the other one to segment nuclear pleomorphism scores.

In the image preprocessing I extracted patches at a resolution of $0.5 \frac{\mu m}{px}$, so that nuclei, which were important for the final training steps, were displayed in detail.

To avoid an unequal distribution of classes within my training, validation and test dataset, I built a sampling method to only create patches from relevant areas and select class areas with the same probability. Furthermore, to enhance the dataset and avoid overfitting, I applied different augmentation techniques, like spatial augmentation, deformation techniques and color staining. All patches were also normalized with an RGB statistic. Because, the analysis of the dataset showed, that nuclear pleomorphism samples mainly consisted of "pleo 2" regions, I additionally applied focal loss and oversampling of the minority classes.

Although a transfer learning pipeline was build, images were augmented using different popular techniques and samples were preprocessed, only low overall IoU scores for the segmentation of tubule formation (0.48) and nuclear pleomorphism (0.23) were reached on the test samples. The evaluation of the outputs showed, that the models were oversimplified.

In the case of tubule formations, the model recognized some tubule areas, but could not segment them in detail. Furthermore, the output masks showed that the model focussed on the lumen and thus also wrongly segmented background areas as tubule formations.

For nuclear pleomorphism scoring, the model labelled all pleomorphism areas as "pleo 2" and ignored other pleomorphism scores. An additional application of focal loss and oversampling of the minority classes did not improve the results.

Compared to other studies using similar techniques on a larger dataset, my results were less accurate. My approach showed, that the accuracy of the model highly depends on the size of the dataset. Due to the variability and complex structures of nuclear pleomorphism and tubule shapes, it is important to provide a diverse dataset, that contains enough examples for every shape. Although I applied transfer learning, image preprocessing and augmentation techniques, my model could not achieve similar or better results compared to other studies with a larger dataset.

Considering the second research question, the outputs with an accuracy of 63.5% for tubule formation segmentation and 37.39% for nuclear pleomorphism segmentation are not precise enough to be used by a pathologist and would need further samples to achieve good results and avoid a bias in the output images.

According to my findings, the conclusion that can be drawn is that my dataset, that contained only 60 ROIs of tubule formation and 195 ROIs of nuclear pleomorphism was not sufficient. Although transfer learning, image preprocessing and image augmentation techniques were applied, good results could not be achieved.

Interesting questions for future research would include investigating the extent to which additional data could enhance the algorithm's results and exploring whether more detailed annotations could improve the segmentation outcomes. Additionally, it would be valuable to explore the application of transfer learning methods using different pipelines or datasets.

References

[1] W. Wang, W. Li, D. Chu, J. Hua, X. Zhang, D. Lu, Y. Wang, and S. Zhang, "Long-term assessment of risk factors for canine tumors registered in xi'an, china," *Animal Diseases*, vol. 1, no. 1, 2021.

[2] L. Peña, P. J. de Andrés, M. Clemente, P. Cuesta, and M. D. Pérez-Alenza, "Prognostic value of histological grading in noninflammatory canine mammary carcinomas in a prospective study with two-year follow-up: relationship with clinical and histological characteristics," *Veterinary pathology*, vol. 50, no. 1, pp. 94–105, 2013.

[3] M. Goldschmidt, L. Peña, R. Rasotto, and V. Zappulli, "Classification and grading of canine mammary tumors," *Veterinary pathology*, vol. 48, no. 1, pp. 117–131, 2011.

[4] P. S. Ginter, R. Idress, T. M. D'Alfonso, S. Fineberg, S. Jaffer, A. K. Sattar, A. Chagpar, P. Wilson, and M. Harigopal, "Histologic grading of breast carcinoma: a multi-institution study of interobserver variation using virtual microscopy," *Modern pathology*, vol. 34, no. 4, pp. 701–709, 2021.

[5] S. M. Abdelmegeed and S. Mohammed, "Canine mammary tumors as a model for human disease," *Oncology letters*, vol. 15, no. 6, pp. 8195–8205, 2018.

[6] N. Sleeckx, H. de Rooster, E. J. B. Veldhuis Kroeze, C. van Ginneken, and L. van Brantegem, "Canine mammary tumours, an overview," *Reproduction in domestic animals*, vol. 46, no. 6, pp. 1112–1131, 2011.

[7] K. U. Sorenmo, D. R. Worley, and V. Zappulli, "Tumors of the mammary gland," in *Withrow and MacEwen's Small Animal Clinical Oncology*. Elsevier, 2019, vol. 6, pp. 604–625.

[8] K. S. Suvarna, C. Layton, and J. D. Bancroft, *Bancroft's theory and practice of histological techniques*, 8th ed. Elsevier, 2019.

[9] S. M. Ayyad, M. Shehata, A. Shalaby, M. Abou El-Ghar, M. Ghazal, M. El-Melegy, N. B. Abdel-Hamid, L. M. Labib, H. A. Ali, and A. El-Baz, "Role of ai and histopathological images in detecting prostate cancer: A survey," *Sensors*, vol. 21, no. 8, p. 2586, 2021.

[10] M. Slaoui and L. Fiette, "Histopathology procedures: from tissue sampling to histopathological evaluation," *Methods in molecular biology (Clifton, N.J.)*, vol. 691, pp. 69–82, 2011.

[11] S. Rashid, L. Fazli, A. Boag, R. Siemens, P. Abolmaesumi, and S. E. Salcudean, "Separation of benign and malignant glands in prostatic adenocarcinoma," *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013*, vol. 16, no. 3, pp. 461–468, 2013.

[12] J. K. C. Chan, "The wonderful colors of the hematoxylin-eosin stain in diagnostic surgical pathology," *International journal of surgical pathology*, vol. 22, no. 1, pp. 12–32, 2014.

[13] C. Wemmert, J. Weber, F. Feuerhake, and G. Forestier, "Deep learning for histopathological image analysis," in *Deep Learning for Biomedical Data Analysis*, 2021, pp. 153–169.

[14] C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up," *Histopathology*, vol. 19, no. 5, pp. 403–410, 1991.

[15] N. Dimitriou, O. Arandjelović, and P. D. Caie, "Deep learning for whole slide image analysis: An overview," *Frontiers in Medicine*, vol. 6, 2019.

[16] E. Tekin, Ç. Yazıcı, H. Kusetogullari, F. Tokat, A. Yavariabdi, L. O. Iheme, S. Çayır, E. Bozaba, G. Solmaz, B. Darbaz, G. Özsoy, S. Ayaltı, C. K. Kayhan, Ü. İnce, and B. Uzel, "Tubule-u-net: a novel dataset and deep learning-based tubule segmentation framework in whole slide images of breast cancer," *Scientific reports*, vol. 13, no. 1, p. 128, 2023.

[17] C. Mercan, M. Balkenhol, R. Salgado, M. Sherman, P. Vielh, W. Vreuls, A. Polónia, H. M. Horlings, W. Weichert, J. M. Carter, P. Bult, M. Christgen, C. Denkert, K. van de Vijver, J.-M. Bokhorst, J. van der Laak, and F. Ciompi, "Deep learning for fully-automated nuclear pleomorphism scoring in breast cancer," *NPJ breast cancer*, vol. 8, no. 1, p. 120, 2022.

[18] R. Jaroensri, E. Wulczyn, N. Hegde, T. Brown, I. Flament-Auvigne, F. Tan, Y. Cai, K. Nagpal, E. A. Rakha, D. J. Dabbs, N. Olson, J. H. Wren, E. E. Thompson,

E. Seetao, C. Robinson, M. Miao, F. Beckers, G. S. Corrado, L. H. Peng, C. H. Mermel, Y. Liu, D. F. Steiner, and P.-H. C. Chen, "Deep learning models for histologic grading of breast cancer and association with disease prognosis," *NPJ breast cancer*, vol. 8, no. 1, p. 113, 2022.

[19] M. Veta, P. J. van Diest, M. Jiwa, S. Al-Janabi, and J. P. W. Pluim, "Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method," *PloS one*, vol. 11, no. 8, 2016.

[20] C. Li, X. Wang, W. Liu, and L. J. Latecki, "Deepmitosis: Mitosis detection via deep detection, verification and segmentation networks," *Medical image analysis*, vol. 45, pp. 121–133, 2018.

[21] C. A. Bertram, M. Aubreville, C. Marzahl, A. Maier, and R. Klopfleisch, "A large-scale dataset for mitotic figure assessment on whole slide images of canine cutaneous mast cell tumor," *Scientific data*, vol. 6, no. 1, p. 274, 2019.

[22] M. Aubreville, C. A. Bertram, T. A. Donovan, C. Marzahl, A. Maier, and R. Klopfleisch, "A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research," *Scientific data*, vol. 7, no. 1, p. 417, 2020.

[23] M. Aubreville, C. A. Bertram, C. Marzahl, C. Gurtner, M. Dettwiler, A. Schmidt, F. Bartenschlager, S. Merz, M. Fragoso, O. Kershaw, R. Klopfleisch, and A. Maier, "Deep learning algorithms out-perform veterinary pathologists in detecting the mitotically most active tumor region," *Scientific reports*, vol. 10, no. 1, p. 16447, 2020.

[24] S. C. Wetstein, V. M. T. de Jong, N. Stathonikos, M. Opdam, G. M. H. E. Dackus, J. P. W. Pluim, P. J. van Diest, and M. Veta, "Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images," *Scientific reports*, vol. 12, no. 1, p. 15102, 2022.

[25] 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*, 2015, pp. 234–241.

[26] F. Wilm, M. Fragoso, C. Marzahl, J. Qiu, C. Puget, L. Diehl, C. A. Bertram, R. Klopfleisch, A. Maier, K. Breininger, and M. Aubreville, "Pan-tumor canine cutaneous cancer histology (catch) dataset," *Scientific data*, vol. 9, no. 1, p. 588, 2022.

[27] Lisa A. Torrey and Jude W. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, IGI Global, Hershey*, 2009, pp. 242–264.

[28] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, 2016.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[30] Minyoung Huh, Pulkit Agrawal, and Alexei A. Efros. (2016) What makes imagenet good for transfer learning? [Accessed: 18.04.2023]. [Online]. Available: https://arxiv.org/abs/1608.08614

[31] Z. Li and D. Hoiem, "Learning without forgetting," in *Computer vision - ECCV 2016*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer, 2016, vol. 9908, pp. 614–629.

[32] Alexander Teplyuk, "Pytorch starter (u-net resnet)," 2019, [Accessed: 13.03.2023]. [Online]. Available: https://www.kaggle.com/code/ateplyuk/pytorch-starter-u-net-resnet/notebook

[33] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[34] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[35] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer vision - ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer, 2014, vol. 8689, pp. 818–833.

[36] S. Otálora and N. Tsiknakis. (2022) stainlib. [Accessed: 18.02.2023]. [Online]. Available: https://github.com/sebastianffx/stainlib

[37] P. Jaccard, "Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines," *Bulletin de la Societe Vaudoise des Sciences Naturelles*, no. 37, pp. 241–272, 1901.

[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 2, pp. 318–327, 2020.