Unsupervised grading of prostate cancer from the feature space of a convolutional neural network

AALBORG
UNIVERSITY

STUDENT REPORT

Group 22gr10401
Biomedical Engineering and Informatics
Aalborg University
June 1st

AALBORG UNIVERSITY

STUDENT REPORT

School of Medicine and Health
Biomedical Engineering and Informatics
Fredrik Bajers Vej 7
9220 Aalborg Øst
http://www.hst.aau.dk/

Title:

Unsupervised grading of prostate cancer from the feature space of a convolutional neural network

Project:

Master Thesis

Master of Science

Project period:

February 2022 - June 2022

Project group:

22gr10401

Authors:

Henrik Paaske Lind

Supervisor:

Alex Skovsbo Jørgensen

Number of total pages: 82
Appendiks: A - D
Date of submission: 01.06.2022

Abstract:

Background and aim: Determining treatment for prostate cancer involve tissue assessment by grading cancer tissue according to its aggressiveness. Morphological structures of cancer tissue are highly heterogeneous making grading prone to inter observer variability and thus wrongful treatment. To mitigate inter observer variability objective measures are needed. The heterogeneous morphological structures of H&E stained WSIs provide an opportunity for using CNNs to learn a cancer appearance feature space, that naturally separates cancerous tissue into gleason score. Thus, the aim of this study was to investigate if the gleason score can be determined from the feature space of a supervised CNN without using grade labels. **Method:** Patches extracted from H&E stained WSIs were used in a multi output CNN consisting of reconstruction and multi class classification for learning corresponding tissue features. Two model configurations were trained and tested to compare tissue feature separation when grade labels are used. The model's performance was validated using unseen test images with an mean squared error metric for the reconstruction and confusion matrix, precision, recall and F1-score for the multi class classification. For quantifying a gleason score the of the model was extracted, where a principle component analysis was conducted using the features corresponding to >90% variance. The point-to-point-score algorithm was developed to quantify a gleason score by calculating the label distribution as a function of accumulated euclidean distance of K-means cluster centroids from benign to gleason score tissue features. **Results:** The mean feature value difference from benign and gleason score 3, 4 and 5 using the test images were (0.005, 0.059), (-0.041, 0.147), (-0.272, 0.187) and (-0.202, 0.153), (-0.290, 0.153), (-0.438, 0.170) with and without grade labels respectively. Using 10 and 25 k-means clusters, the majority of benign, gleason score 3 and 4, and 5 were present between 0 to 0.6, 0.6 to 0.9 0.9 to 1.2 (1.4 -25 clusters). **Conclusion:** Superior tissue feature separation can be obtained without using grade labels, where the PPS-algorithm can imply a gleason score based on accumulated euclidean distance from benign tissue features.

Preface

This master thesis is made by Henrik Paaske Lind, from February 1st to June 1st 2022, constituting the final project of Master of Science in Biomedical Engineering and Informatics at Aalborg University.

A special thanks to Alex Skovsbo Jørgensen for supervising throughout the Masters degree and for providing feedback throughout the project period.

Reading instructions

This study contains a problem analysis in chapter 2, elaborating the problems associated with grading prostate cancer. Based on the problem analysis, current literature is investigated to specify the aim. The methodical approaches for investigating the aim are elaborated in chapter 4. This chapter is split in two parts, whereas the first is composed of data selection and pre-processing, deep learning architectural choices and explanations, validation metrics of the deep learning network. The second part constitutes feature space analysis components and a developed algorithm for assessing prostate cancer. The results from the implemented methods are shown in chapter 5 and the master thesis is concluded through chapter 6 and 7, reviewing the methods and results.

The literature in this study is referenced using the Harvard Referencing System. As such, the concerning literature is cited by last name followed by the year of publication. The references can be used actively or passively where the concerning reference would be placed before or after a period respectively. All references are found within the bibliography whereas the proceeding pages consist of a structured literature search, theoretical explanations of deep learning components and a portfolio reflecting upon writing a master thesis.

Henrik Paaske Lind

Contents

1 Introduction

Cancer is a disease that acquires the ability to multiply cells in an uncontrolled manner. Often cancer forms tumors, that obstruct the function of the concerning organ(s) in the body. If the cancer is left untreated it can spread and eventually cause death. [Miller, 2016] In 2020 approximately 19.3 million new cancer cases and 10 million cancer deaths were recorded, which makes cancer the first or second leading cause of death before age 70 within 112 of 183 countries. The number of new cancer cases is expected to increase to 28.4 million in 2040. Prostate cancer ranks third in overall new cancer cases, accounting for 7.3%, within which 3.8% of cancer deaths were recorded. [Sung et al., 2021]

Prostate cancer affects the functions of the male reproductive organ, resulting in problems with erection, urination issues, and general pain from cancer spread. Prostate cancer is diagnosed by employing several tests for excluding any other possible disease-sharing symptoms. [Miller, 2016]

For determining the tumor tissue grade and stage, a pathologist examines the cancerous tissue by analyzing slices of tissue from a biopsy. [Ozkan et al., 2016] However, the tumor tissue can be heterogeneous and complex, which makes the diagnosis prone to interobserver variability and a wrongful evaluation could mean an unnecessary invasive treatment. [van Santvoort et al., 2020]. As such, an accurate and objective evaluation is important for correct treatment.

Current deep learning methods show great accurate performance for classifying prostate cancer. [Campanella et al., 2019; Ström et al., 2020; Gummeson et al., 2017] However, the majority use annotated data by pathologists, which are subjective by nature. Using unsupervised methods for grading prostate cancer, could mitigate the current challenges within prostate cancer classification.

Prostate Cancer 2

Throughout this chapter the disease progression of prostate cancer is elaborate. This leads to the course of diagnosis that is currently conducted, whereas challenges associated with prostate cancer assessment namely grading are explained.

2.1 Progression of Prostate Cancer

The prostate is a part of the male reproductive organ and is located below the bladder. The prostate surrounds the urethra, which allows the passage of urine and seminal fluid. [Rajal B. Shah, 2012] There are two main cell types in the prostate, epithelial and stromal. Stromal tissue provides structural support to surrounding cells and is composed of smooth muscle cells, fibroblast, and myofibroblast, figure 2.1. The epithelial cells are composed of secretory cells, which secrete fluid, basal cells surrounding the secretory cells, and rare neuroendocrine cells. [Oxley, 2014]

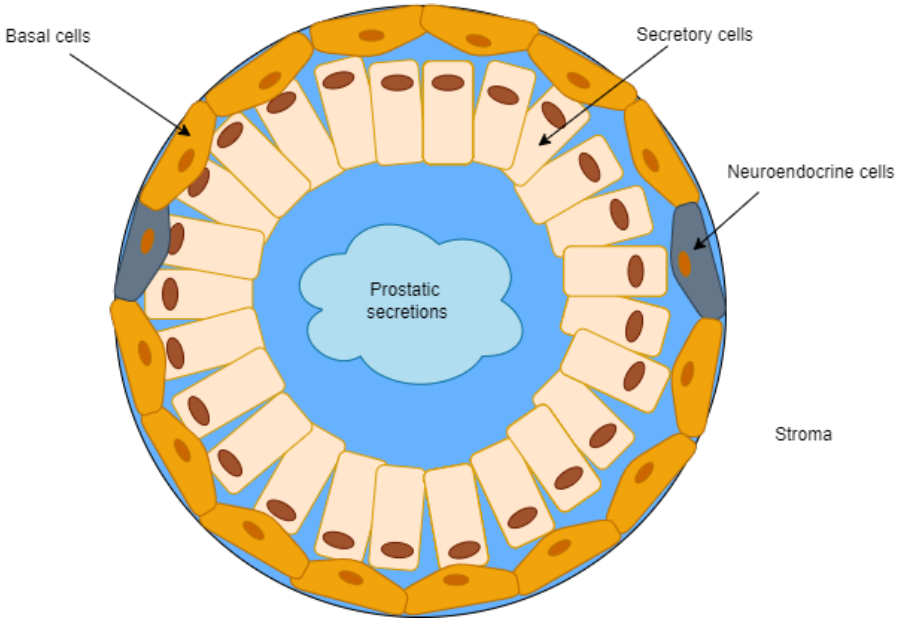


Figure 2.1. Representation of the cell types in the prostate

Within some of these cells, prostate cancer (PCa) forms, and approximately 95% of malignant PCa form from the epithelial cells, which are called adenocarcinoma (AD). [Cho et al., 2012]. In general, the progression of cancer starts within the DNA of a normal cell, which carries a particular instruction that describes what the function of that cell is. If a copy of the DNA becomes damaged, the function of the replicated cell is altered

and it behaves and appears differently (dysplasia), becoming more undifferentiated as the tumor evolves. Fortunately, few abnormal cells are controlled by the immune system and are harmless. However, if the abnormal cells start to grow uncontrollably, eventually tumors can form, which can be cancer and can be dangerous. Both benign and malignant tumors can be life-threatening, however, malignant tumors can infest surrounding tissue (carcinoma in situ) and become invasive (invasive carcinoma), figure 2.2. If the cancer is invasive, it can enter the bloodstream, which can cause cancer cells to invade other organs where tumors can form which are generally more dangerous than benign tumors. [Martini et al., 2003; InformedHealth.org, 2019]

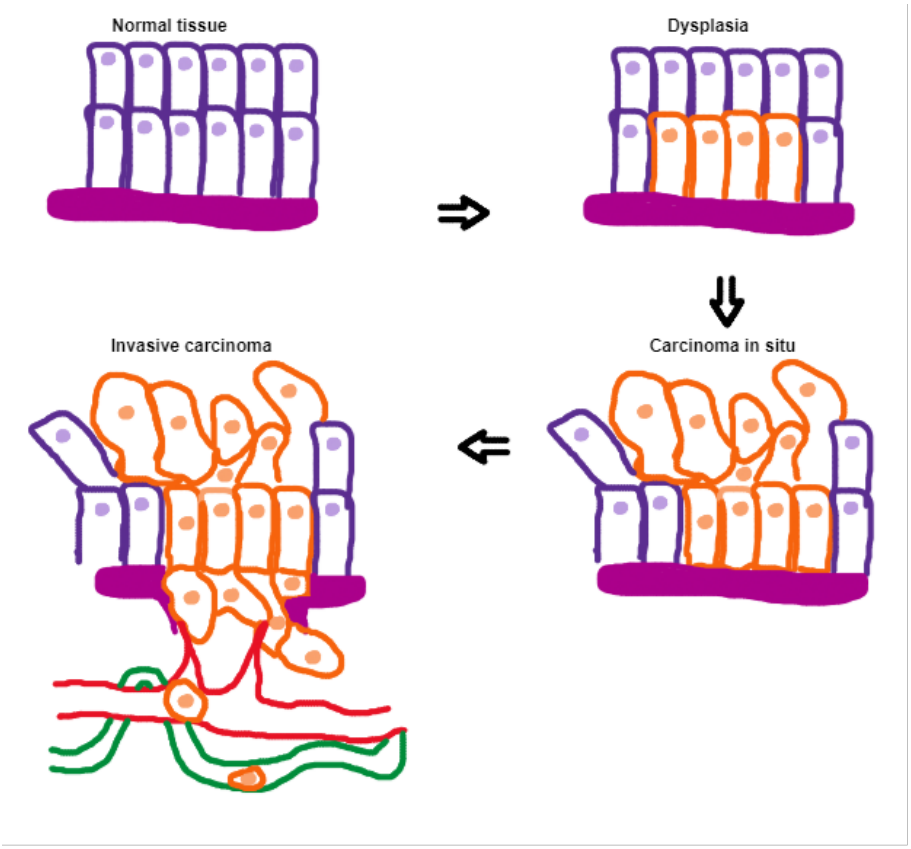


Figure 2.2. Drawing of the progression of cancer from normal tissue to invasive tumor. Inspired by Martini et al. [2003]

Initially, PCa develops without symptoms during the early stages. As it advances, patients may experience difficulty urinating or blood in urine or semen, difficulty in getting an erection, weakness or numbness in their legs, feet, bladders, or bowels due to tumor pressure on the spinal cord, or pain in their hips, spines, or ribs due to cancer spreading to bones. However, this does not guarantee PCa as the causing disease, since other health issues produce similar symptoms. In addition, noncancerous growths can form in the prostate, which makes an early diagnosis difficult. [Miller, 2016] According to WHO [2022] early diagnosis reduces mortality since the cancer is more likely to respond to treatment, which is a reason for employing several tests before reaching a final diagnosis.

2.2 Prostate Cancer Diagnosis and Treatment

When PCa is suspected, a rectal examination is conducted, palpating the prostate and nearby tissue to detect any unusual lumps or masses of cells. The examination may also involve a trans-rectal ultrasound for imaging the prostate and surrounding tissue. [Miller, 2016]. If any unusual anatomy is found, several tests are warranted to uncover the cause, which are physical examination, medical history, imaging, PSA level, staging, and grading. These are essential prognostic factors contributing to the final diagnosis and treatment plan and are shown in figure 2.3. [Gospodarowicz et al., 2017]

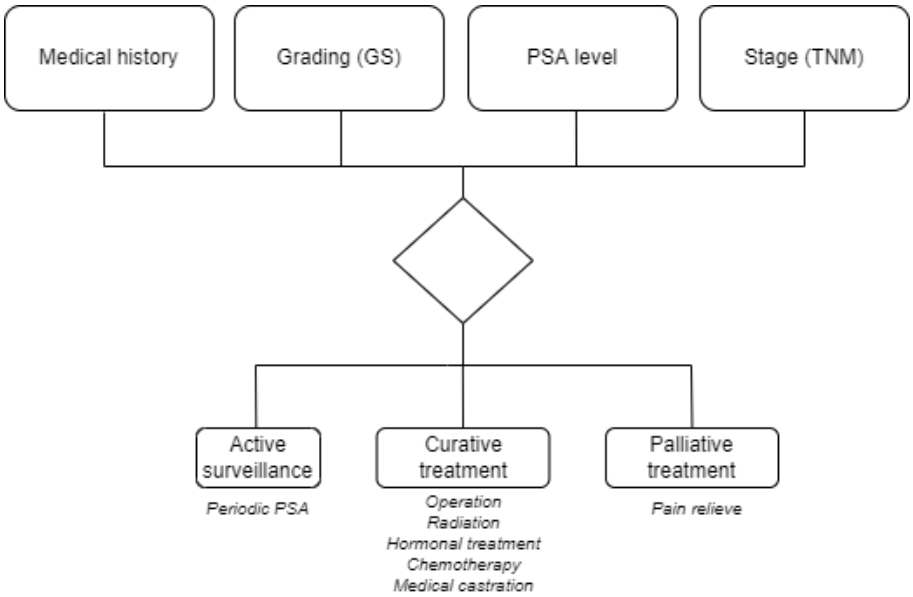


Figure 2.3. Flow diagram explaining the essential prognostic factors for deciding treatment options

After the rectal examination, the blood is analyzed for prostate-specific antigen (PSA), which is often increased in men with PCa. However, increased PSA does not guarantee a diagnosis for PCa, since several benign conditions, such as the increased size of the prostate, age, and infection also increase PSA levels in the blood. [Miller, 2016; Michael Borre, 2019] Conversely, patients with low levels of PSA, have also been diagnosed with PCa. [Miller, 2016] Since PSA levels vary in different individuals, the thresholds for concern vary depending on the patient's condition, medical history, and physician experience. [Miller, 2016; Michael Borre, 2019; Borre et al., 2019] If the rectal examination and PSA test suggest PCa, a biopsy, and subsequent histological analysis is currently conducted.

The principles of the histological analysis are to evaluate the behavior of cancer and subsequently classify it using the international standard for assessing a malignant tumor called the Tumor Node Metastasis (TNM) system. The principle of the system is to classify cancer according to an anatomical extent, where a particular classification gives essential information about expected patient survival and treatment planning. Specifically, the primary tumor (T), spread to lymph nodes (N), and metastasis (M) are given a score ($T1-2_{a-c}, T3_{a-b}, T4$, $N0-1$, $M0, M1_{a-c}$), which tells what stage the cancer is in, higher being bigger and generally more dangerous tumors (Stage 1-4). [Gospodarowicz et al., 2017]

In addition, the aggressiveness of the tumor, called tissue grading, is determined by visually examining the tissue from the biopsy and estimating a Gleason score (GS). Before a GS can be determined the tissue is prepared by fixation, sectioning/slicing, and staining, permitting visualization of the tissue structures. The commonly used stain for light microscopy is the hematoxylin and eosin stain, which provide contrast between cells and stromal tissue. Initially, the tissue is stained with Hematoxylin, coloring the cell nuclei dark blue. The contrast color is provided by eosin, coloring the stromal tissue in a pink/red color, allowing a clearer visualization of the structural features of the tissue. He et al. [2012]; Gospodarowicz et al. [2017]; Suvarna et al. [2018]

As such, the tissue structures can be examined, which determines what GS and subsequent Gleason grade the tissue is given. [Gospodarowicz et al., 2017] The GS specifies if a tumor cell looks similar to healthy epithelial cells (Grade group 1) or very abnormal (Grade 4-5), whereas cells that are in between are given a Grade of 2-3. Since PCa can be heterogeneous, the two most aggressive patterns of the biopsy are given scores, resulting in a GS from 2 to 10. [Ozkan et al., 2016]. The sum of the two GS determines what grade group the cancer is in (Grade group 1-5), higher being more aggressive cancers and higher risk cancers. However, this has an obvious limitation since there is no objective tumor grade assessment, which could improve the diagnostic accuracy.

Depending on the results from the initial tests, there are three main treatment options provided to the patient, which are active surveillance, curative and palliative treatment. Active surveillance is a treatment option, whenever cancer is not aggressive, and a periodic PSA test can determine if further intervention is necessary. The curative treatment is divided into five different treatments depending on the extent of the cancer. If the cancer is limited to the prostate, both radiation, and operational treatment is an option. However if cancer has spread to lymph nodes, other organs, or bones, hormonal treatment, chemotherapy, and medical castration are options. Lastly, if the curative treatments have no effect, the only option is to treat the pain. [Michael Borre, 2019; Borre et al., 2019].

Given that the treatment options rely on the cancer assessment, it is important to obtain objective results, since a wrongful grading could mean operation instead of active surveillance, where some of the side effects are incontinence and impotence due to nerve damage from the operation [Michael Borre, 2019; Thomsen et al., 2015],

According to Ozkan et al. [2016], the Gleason grading is an independent variable for determining a suitable PCa treatment and ideally, GS should be independent of the observer and thereby identical among different uro-pathologists. [Thomsen et al., 2015] However, with heterogeneous cancer tissue and complex morphological structures, the subjective nature of Gleason grading could lead to interobserver variability. [Ozkan et al., 2016]

2.2.1 GS Interobserver variability

Several studies have investigated interobserver variability in PCa in current research, and the results are mentioned in table 2.1, while the studies are elaborated upon throughout this section.

In a study by Ozkan et al. [2016], it was acknowledged that grading systems are subjective

methods, and accurate diagnosis is an important problem concerning intraobserver and interobserver variability. For testing the interobserver variability within GS, two pathologists determined the GS from the same tissue. A concordance with respect to Gleason sum was 58%, with a general moderate agreement of all evaluated slides $\kappa = 0.43 - 0.68$, exemplifying the interobserver variability.

Thomsen et al. [2015] showed that the interobserver agreement on GS was 63.4% with a weighted κ of 0.670, from uro-pathologists examining the same tissue. In addition, the recommendation on whether to remain in active surveillance or proceed to curative intended treatment differed by up to 10.1%. However, the differences in GS assessment were minor, where $< 5\%$ were more than one GS apart, but the consequences mean that a particular patient would be recommended curative treatment instead of active surveillance.

Owing to the difficulty of grading, Egevad et al. [2011] tested the interobserver variability when tissue patterns (Gleason sum S 6 and 7) are similar. In a specific case, three pathologists voted a Gleason sum of 9, and two voted a Gleason sum of 5 from the same tissue. From the perspective of the GS system, a Gleason sum of 5 means that the patient is at low risk, whereas a Gleason sum of 9 means that the patient is at high risk, and the treatment would likely differ and could affect the patients survival and quality of life. [van Santvoort et al., 2020]

Article	Method	Result
Ozkan et al.	Gleason sum estimation	kappa = 0.43 - 0.68
Thomsen et al.	Gleason sum estimation	Interobserver agreement = 63.4%
Egevad et al	Gleason sum estimation	Interobserver agreement = 93% - 67%

Table 2.1. Interobserver variability in three studies

Currently, the GS system is used for determining a suitable PCa treatment but is subjective which could cause wrongful treatment. Using an objective method would mitigate the interobserver variability, which could increase the quality of life and patient survival. Fortunately, scanner technologies allow for digitizing tissue slides as a whole slide image (WSI), which can be used for training a classification model. However, a general cancer diagnosis of a WSI yields weak labels for all pixels contained in the WSI, and it is only certain that one part of the WSI contains cancer and often requires many examples, which is technically challenging and exhausting to annotate for pathologists [Campanella et al., 2019]

As such, it is interesting to research whether convolutional neural networks (CNN) trained without grading labels can differentiate tissue, mitigating the interobserver variability and the subsequent consequences in the current grading system. For this reason, methods for grading, cancer assessment, or tissue differentiation using CNN, were researched through a systematic literature search.

Related work 3

In the current research, the majority of machine learning methods for classifying cancer are supervised, meaning that the data provided to the model is labeled by clinicians. However, some supervised methods use unsupervised elements for assisting the classification. As such this chapter is divided into two parts, which are supervised classification of cancer, and unsupervised elements for classification of cancer.

3.1 Supervised classification of cancer

The recent reviews of Tătaru et al. [2021]; Linkon et al. [2021]; Tran et al. [2021], covering machine learning in PCa and general deep learning in cancer research, explain the same general methodical approaches for classifying cancer, which is supervised machine learning.

Tătaru et al. [2021] review artificial intelligence in the pathology of prostate cancer. While the methods were all supervised and prone to interobserver variability, a study by Campanella et al. [2020] partly overcame this by obtaining 12.132 in-house biopsies and 12.727 biopsies from around the world, weakly annotated by different pathologists for predicting cancerous and non-cancerous WSIs. When using weakly annotated labels, it is only certain that one part of that WSI contains cancer. Owing to that fact Campanella et al. [2020] used multiple instance learning, dividing WSIs into patches of size 224x224 as input to a CNN, providing the weak label to all the patches from the concerning WSI. For finding the cancerous regions, the patches closer to a probability of 1, were selected for classification of the whole WSI. The AUC was 0.989 showing good performance. However, this approach is only applicable, when enough data is available, as the sheer volume of correctly annotated data, sorts out any discrepancies in the annotation.

Tran et al. [2021] summarize machine learning through different applications within cancer research, which are diagnosis, prognosis, and treatments, whereas diagnosis is the topic of concern in this study. Of the five deep learning methods assessing cancer on histological WSIs, Ström et al. [2020], Nir et al. [2019] and Ryu et al. [2019] were of PCa assessment. All had benign vs. cancer accuracy $> 90\%$ using CNNs, however, labels were either biased by one pathologist, or intrinsically included interobserver variability, as the method for labeling was by majority voting of the concerning label, demonstrating that there is a disagreement in grading. In addition the sensitivity for benign, Grade group 1, 2, 3, 4, 5 were 0.94% 0,33%, 0,38%, 0.59% 0.33% and 0.97% respectively, showing the difficulty in predicting intermediate grades, even with labels. [Ryu et al., 2019]

The review by Linkon et al. [2021] provides an insight into the evolution of deep learning within the PCa research and composes the state of the art deep learning innovations for

classifying PCa. One of the best performances in grading grades 3, 4, and 5 were by Gummeson et al. [2017], achieving an average 92.7% accuracy. Through pre-processing, they down-sampled WSIs at 40x magnification to 6x magnification, covering larger spatial areas, which is beneficial for predicting Gleason grade, as it retains more information about the WSI. As such, the input was 106x106 for a CNN, down-sampling into a flattened representation, from where a fully connected network handled the final classification of either benign, Gleason grade 3, 4, or 5. This shows that discriminative features can be learned by using simple deep learning architectures.

Unfortunately, no deep learning approaches mentioned in Tătaru et al. [2021]; Linkon et al. [2021]; Tran et al. [2021] used unsupervised classification. An overview of the included supervised methods and their limitations are presented in table 3.1

Article	Method	Performance	Limitations
Campanella et al.	Cancer vs non-cancer	AUC = 0.989	Weak cancer labels yields little rich information
Ström et al.	GS3 vs GS4 vs GS5	AUC = 0.997	Grade labels were used for training
Nir et al.	Cancer vs. non-cancer	Accuracy = 97.8%	Probable good performance due to few prediction labels
	Low grade (GS3) vs High grade (GS4, GS5)	Accuracy = 92.5%	
Ryu et al.	Cancer vs non-cancer	Sensitivity = 0.94% Specificity = 0.99%	Grade labels were used for training
	Benign vs grade group 1,2,3,4,5	Sensitivity = 0.94%, 0.33%, 0.38%, 0.59% 0.33%, 0.97%	
Gummeson et al.	Benign vs grades 3, 4, 5	Average accuracy = 97%	Grade labels were used for training

Table 3.1. Table overview of supervised methods, results, and limitations for the included articles in this study

The common trait for the included supervised articles is that they are bound by the labels used for training, giving little information about the cancer evolvement e.g in cancer vs. non-cancer and high-grade vs low-grade cancer cases. In principle, cancer should be viewed as a regression problem, where the tissue differentiation is a smooth transition from early dysplasia to invasive cancer, where many potential sub-categories of cancer exist, which can not be learned by supervised classification. Therefore unsupervised approached must be employed and approaches using unsupervised elements are explained in the following section.

3.2 Unsupervised elements for classification of cancer

As a part of their methodical pipeline, Lu and Daigle [2020] performed an unsupervised sub-grouping of liver cancer for characterizing survival differences. For feature extraction they used pre-trained CNN models (VGG 16, Inception V3, and ResNet 50), inputting the same image and computing the median values from the three models as the output. For reducing the number of dimensions and visualizing the components, they used PCA and t-Distributed Stochastic Neighbor Embedding (t-SNE), reducing the number of components to 10, and visualizing the components in 2 dimensions. Through PCA, it is possible to utilize features that describe the largest variance and eliminate redundancy, permitting a

t-SNE algorithm to represent features in a lower dimension whilst maintaining the feature relation in two dimensions. [Lu and Daigle, 2020] Using all cancer samples, Lu and Daigle [2020] discovered subgroups by performing k-means unsupervised clustering and Silhouette coefficient and Davies-Boulding index. The Silhouette coefficient ensures a good cluster separation by calculating the mean intra-cluster distance and the mean nearest cluster distance, while the Davies-Boulding index ensures good partitioning by calculating the cluster similarity. Two subgroups were determined to be optimal and the second subgroup consistently had a better survival prognosis than the first. This demonstrates that it is possible to find additional feature information from cancerous WSIs that correlate with class predictions.

While this example is applied to liver cancer, the same principle could be applied to PCa, finding possible subgroups of cancers, potentially exhibiting different Gleason grades. For example, cancer patches placed closer to healthy epithelial tissue, could indicate lower GS, while cancer patches farther away could indicate higher GS.

Bulten and Litjens [2018] proposed an unsupervised approach for finding relevant morphological features for detecting PCa, without relying on labeled data, aiding pathologists in finding tumorous regions. They used a clustering adversarial autoencoder, which clusters tissue as part of the training process, without the need for post-processing. It consists of an encoder, embedding two flattened latent vectors, trained to follow a Gaussian and categorical distribution and regularized by a discriminator.

The general idea is, that the categorical distribution should encode high-level information, assigning the input data to one of 50 clusters, while assigning style information to the Gaussian distribution in the style vector of size 20, aiding the reconstruction of the output data. Finally, a decoder reconstructs the data, sampling from the latent vectors. The input and output data were 128x128 H&E and IHC stained WSI patches at 5x magnification, obtained from patients that underwent radical prostatectomy. Using only cancer and non-cancer labels Bulten and Litjens [2018] found that features of PCa are separated from stroma but are somewhat entangled with benign tissue, by using a t-SNE. For validating the cancer/non-cancer performance, Bulten and Litjens [2018] used 1000 patches of stroma and epithelium and 2000 patches of tumor tissue. For examining the relative performance for detecting PCa, image translations from H&E to H&E and H&E to IHC, with 200 labeled patches of each class and all labeled patches were conducted. The resulting cancer/non-cancer accuracies for H&E to H&E were 59% and 63% and for H&E to IHC were 68% and 73%, using 200 labels and all labels, respectively. [Bulten and Litjens, 2018] This approach shows that by learning a cross-mapping such as H&E to IHC, there is some added benefit in terms of classifying cancer while inherently learning more distinct features. However, cross-mapping data is difficult to obtain, since the tissue slides are at risk of staining artifacts and tissue damage during preparation. In addition, this approach is partly supervised since annotations were used for classifying cancer. Most importantly, this approach demonstrates how unsupervised learning could be used to take advantage of the ability of deep learning to learn distinct features while avoiding subjective annotations.

Bauer et al. [2021] acknowledged the difficulty in grading heterogeneous tumor tissue in PCa resulting in label noise from the interobserver variability. This raised the interest in unsupervised classification of Gleason grade in PCa for reducing the subjective influence

by pathologists on classification results. The proposed framework consisted of multiple steps. Initially, pre-processing steps weere conducted for normalizing the stain variability of the data, which were then used in a regular stacked autoencoder for embedding latent representations and reconstructing the original input. The latent representation was used for predicting Gleason grade, and the autoencoder was then trained using the reconstruction loss and the classification loss. Bauer et al. [2021] used grading labels as a method for a possible improvement of the latent representations. However, this obviously defied the label noise that the aim of the work tried to overcome. For post-processing a principal component analysis (PCA) was used on the latent representation, reducing the dimensions into a set of principal components. Subsequently, a one-class Support Vector Machine classifier (SVM) was used to classify malignant patches, filtering benign and non-relevant patches. Finally, another SVM was used on the remaining cancerous patches, to create pseudo-labels referring to Gleason sum 6, 8, and 10. The general idea is that selecting non-mixed scores results in the least overlap in feature space (FS) since the primary grade covers most parts of the data and thus easier to classify. The classifications were then transferred and used as pseudo-labels for a CNN. In doing so, the accuracy in classifying stroma, benign tissue, GS3, GS4, or GS5 increased from between 64-70% to 72-78%, demonstrating that the features learned by using pseudo-labels, can be more consistent than the original label from an observer annotating such complex structures. However, the accuracy of the primary and secondary Gleason grade of the SVM providing the pseudo-labels was between 18-48% and 20-43%, respectively, questioning the validity of the final classifications of the CNN. This approach shows that meaningful clusters can be found by using an autoencoder and that mixed GS are difficult to classify. [Bauer et al., 2021]

Throughout the structured literature search, studies exploring cancer grading without using the corresponding labels were not found. However, Lu and Daigle [2020]; Bauer et al. [2021] prove that additional feature information exists for potential sub-grouping of cancer, by using some degree of supervised labels. For visualizing high dimensional features t-SNE and PCA were used, which permits FS analysis and determining the number of sub-groups for unsupervised clustering. An overview of the included unsupervised methods are presented in table 3.2

Article	Unsupervised element	Beneficial outcome
Lu et al.	CNNs for feature extraction without labeled information	Subgroups can be found using CNNs
Bauer et al.	Stacked autoencoder	Benefits of using labeled information with reconstruction
Bulten et al.	Clustering adversarial autoencoder	Separates features by learning morphological structures in H&E images

Table 3.2. Table overview of unsupervised methods, unsupervised elements, and beneficial outcomes from those elements for the included articles in this study

For overcoming interobserver variability, it would interesting and novel to model a cancer appearance FS, quantifying the natural tissue variation from benign to malignant tissue, using it as a natural cancer grading method.

Therefore, this study investigates whether a supervised CNN, trained using WSIs labeled in cancer and non-cancerous patches, organizes features in a way that makes it possible to find and quantify GS sub-groups. This interest formulated the research question:

Aim of study: Can the GS be determined from the FS of a supervised CNN without using the GS labels?

Method 4

This chapter is arranged in the following order. Initially, the general strategy is explained, followed by a description of the data used in this study. Next the framework for quantifying a gleason score and the contributing parts are elaborated, which is split in supervised training and FS analysis components.

For researching the aim of this study, the adopted approach was two fold. First, an autoencoder architecture in combination with a supervised classification network, denoted CNN_{dual}, was used, showed in figure 4.1. The general idea was that the autoencoder should learn features corresponding to the morphological structures of the tissue, while the classification part learns tissue specific features from benign stroma and cancer. Secondly, stroma, benign and cancer patches were used to investigate the relative position of features in FS. As an example, the distance between benign tissue and cancerous tissue, could indicate higher or lower gleason pattern, respectively. The features were reduced to a lower dimensional space to visualize the feature relation and perform clustering. For quantifying the relative position of features, a point-to-point-score (PPS) was developed. The idea was, that the GS could be determined by quantifying the tissue differentiation of benign and stroma tissue, to the GS tissue. Therefore, GS labels were used but not for training the CNN_{dual}, overcoming any interobserver variability existing within the labels. Finally, a validation was performed, using the output of the PPS, to indicate what GS the input was. As an exploratory step, grade labels were used in CNN_{dual} to compare the FS with the FS using only cancer labels. This could indicate whether the additional supervised label information is needed or not. For example, if the two FSs are similar in terms of feature relation, there could be no need for the additional grade label information.

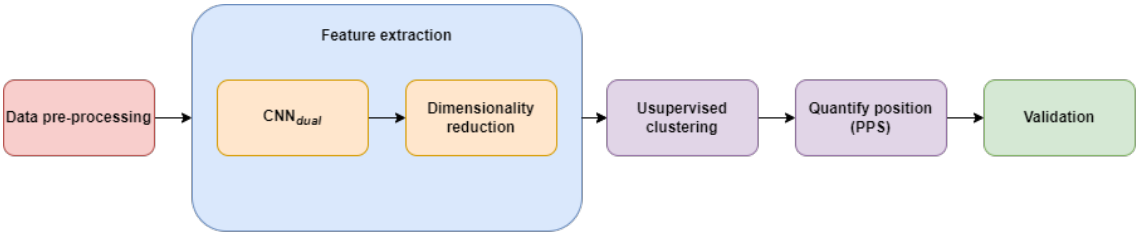


Figure 4.1. General structure of the method in this study

4.1 Data

In preparation of this study, data was downloaded from Radboudumc and Institutet [2020]. Specifically, the downloaded dataset was the prostate cancer grade assessment (PANDA) challenge, consisting of 10616 H&E stained WSIs, with corresponding 10516 annotated

WSIs. The data is provided to kaggle by the Karolinska Institute and Radboud University Medical Center. All providers scanned the WSIs at a resolution of 20x magnification and the digitized slices were subsequently annotated by different pathologists. The annotations were done on every pixel of the WSIs, where one pixel intensity represented a class. Karolinska Institute labeled values as background, stroma and cancerous tissue, while Radboud University labeled into six classes, background, stroma, healthy, GS3, 4 and 5. For the purpose of this study, the data from Radboud University was used, as it provided an opportunity for using highly detailed labeled images with gleason scores for analyzing FS. Moreover, in initial data preparation, it was revealed that Karolinska Institute labeled some background as cancer, and therefore disregarded for use in this study. As such, 5144 WSIs were left for subsequent patch extraction. [Radboudumc and Institutet, 2020] Figure 4.2 shows a flow diagram depicting the general patch extraction solution. Only stroma, benign and cancerous tissue (GS 3, 4 and 5) were considered by the patch sorting algorithm, discarding any background patches.

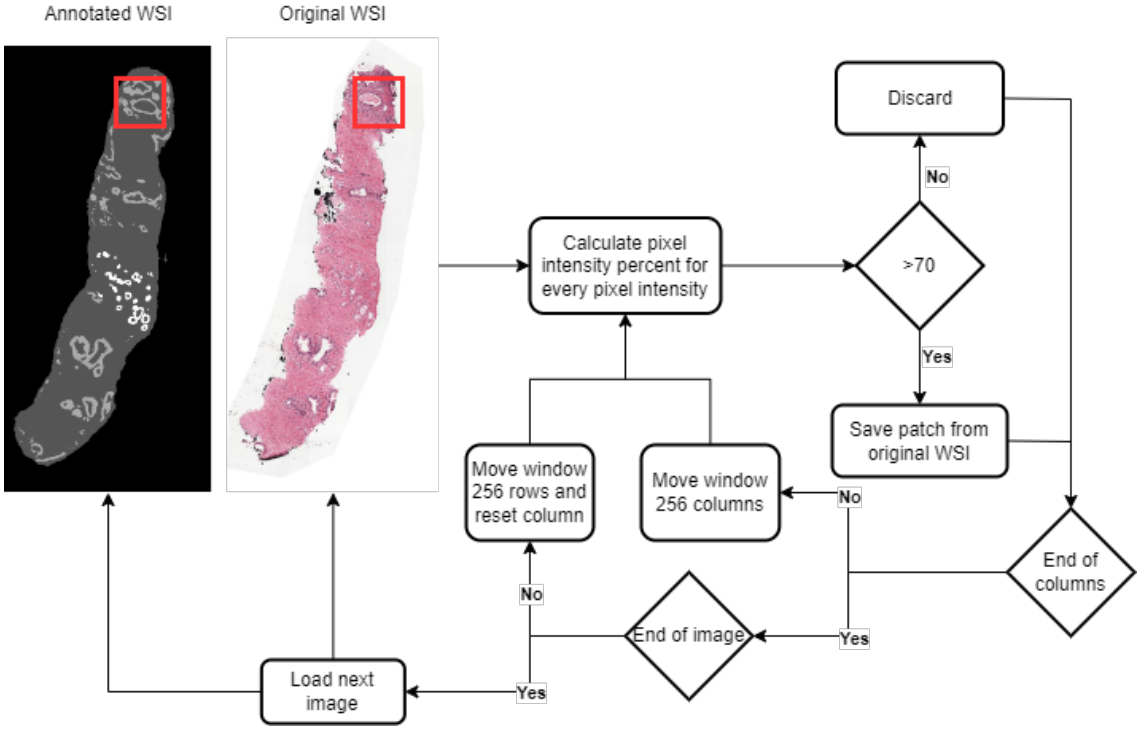


Figure 4.2. Flow diagram showing the steps of the patch sorting algorithm.

Patches were extracted using a sliding window on both the annotated and H&E stained WSIs. The annotated WSIs were used for categorizing the H&E patches, by setting a threshold $> 70\%$ for a single pixel intensity in the concerning patch. As an example, if the annotated patch had $>70\%$ of annotated cancer, benign or stroma tissue, the corresponding H&E patch was saved. The threshold was a trade-off between having the majority of one tissue class in one patch and sufficient amount of patches. In a similar manner the window size of 256x256 was chosen for obtaining enough patches, but also for capturing sufficient visual field containing the structure of the prostate glands. Five exemplary patches are shown in figure 4.3, depicting the annotated and original patches, all containing $> 70\%$ of a certain tissue type.

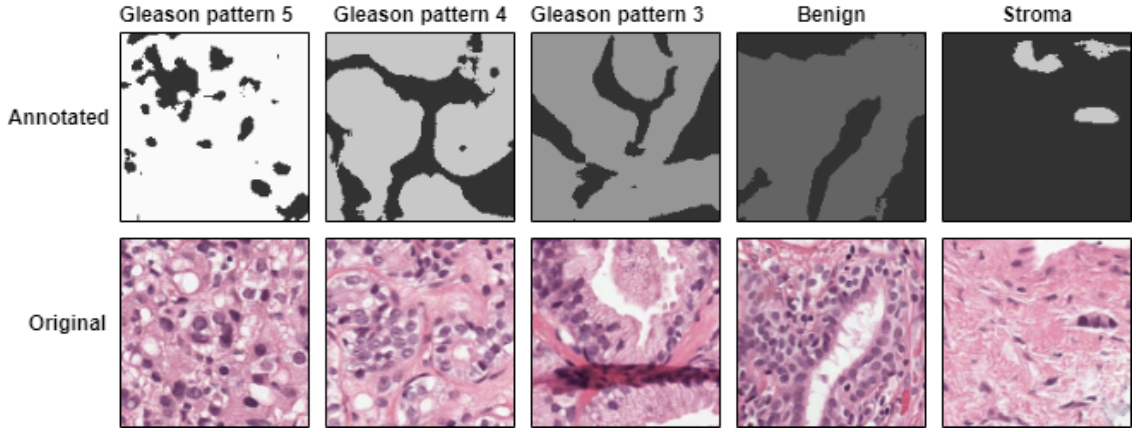


Figure 4.3. Array of patches depicting all types of tissue used for this study

When performing the patch extraction algorithm, more cancerous patches than benign patches were extracted, creating a large class imbalance. In order to balance the dataset, the total amount of benign patches extracted determined the amount of cancerous patches for training, which can be viewed in table 4.1 in addition to the amount of patches for validation, testing and the remaining patches.

However, some class imbalance was necessary to represent the variance of cancer patches (GS3, GS4 and GS5) within the data set. Therefore, the amount of cancer patches for training are three times as large as stroma and benign patches, with an equal division between GS3, GS4, and GS5 patches.

		Training	Validation	Testing	Total	Remaining test patches
Stroma patches		3646	450	450	4546	1346
Benign patches		3646	450	450	4546	0
Cancer patches	GS3	3646	450	450	4546	5172
	GS4	3646	450	450	4546	38533
	GS5	3646	450	450	4546	2059
Total		18230	2250	2250	22730	47110

Table 4.1. Table of the amount of patches split in training, validation, testing and remaining

4.2 Framework for Quantifying Gleason Score

For quantifying the gleason score, the framework in figure 4.4 was developed.

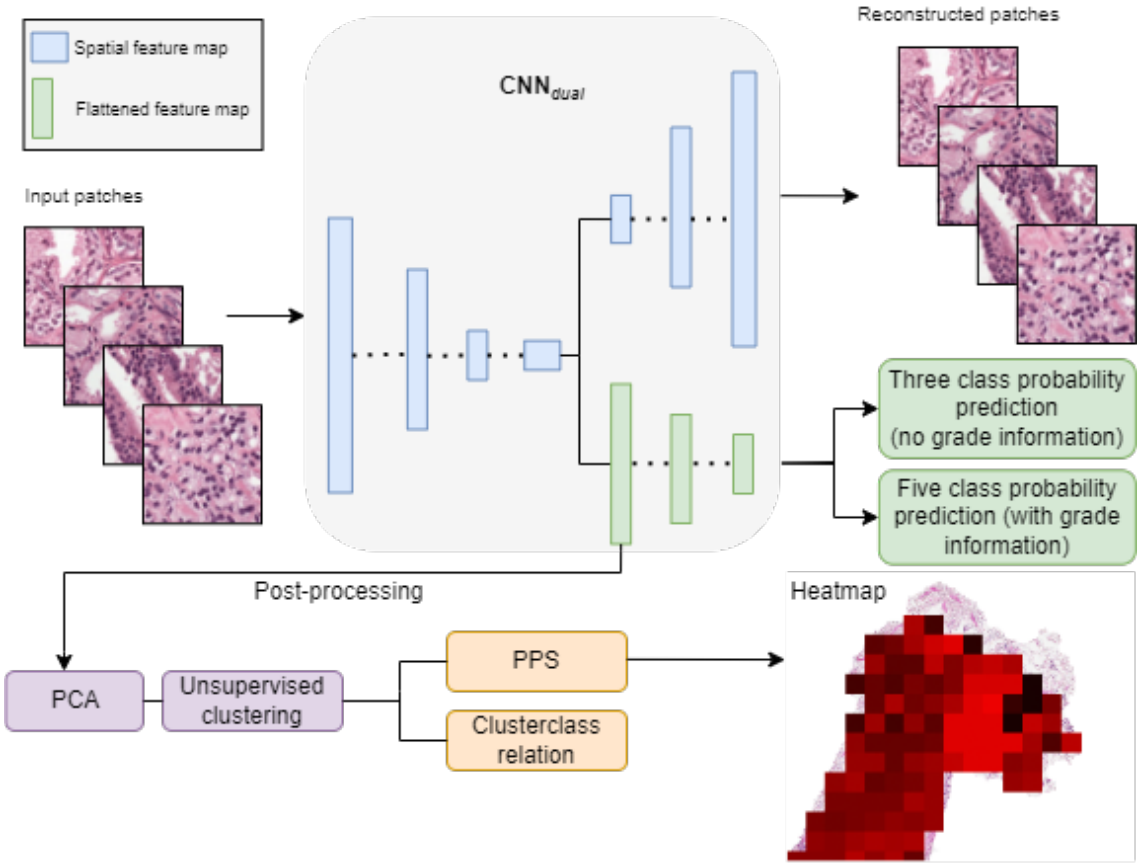


Figure 4.4. The general framework for developing a quantification of gleason score. The orange rectangles represents the constituents of the algorithm for calculating the gleason score.

There are two main steps within the framework of this study, these are CNN_{dual} and post-processing. CNN_{dual}, is a multi output CNN trained to differentiate between benign, stroma and cancerous tissue, while revealing potential sub-groups within the cancerous tissue, which was shown possible by, [Lu and Daigle, 2020]. In addition CNN_{dual} is trained with grading labels i.e GS3, GS4 and GS5, for later comparison of potential sub-grouping. Note that these are two separate training sessions. After the training the model is saved and the feature space is extracted from the first flattened feature map, as this represents the entire compressed patch information in a single feature vector, from where the model should be able to classify into the three and five classes. Subsequently, a dimensionality reduction was performed with a PCA, which permits visualizing features in two or three dimensions, while retaining most variance of the data, within which feature space analysis can be conducted. In addition, fewer dimension uses less computational resources when performing clustering. The clustering was done using K-means on the features extracted from the PCA for revealing naturally occurring sub-groups in feature space i.e GS3, GS4 and GS5 features. For quantifying the gleason score, the PPS used the euclidian distance measure between the cluster centroids of the unsupervised clustering method. Recalling that the more aggressive the cancerous tissue is the more differentiated the tissue looks, and

thus the features should also become differentiated, placing farther away from features of benign tissue. The distance of cancer patches should then indicate higher or lower gleason score respectively, which was used in a heatmap, representing location and aggressiveness of the cancer on top of the original image. Additional information in the form of cluster class relation was also provided, where the label distribution within the clusters were showed.

4.2.1 Supervised deep learning CNN_{dual}

The architecture of this study was a supervised CNN, which learned by taking input patches of size 256x256x3 with a corresponding label. The architectural choices were inspired by previous experience with CNNs, and current state of the art architectures used by the included studies, such as VGG16 by Lu and Daigle [2020]. The common trait from these state of the art architectures was to downsample an input image into a fully connected classification, using maxpool layers for downsampling and convolutional layers for learning the features. [Chollet et al., 2018]. The theoretical explanations of the components constituting CNN_{dual} are found in appendix B

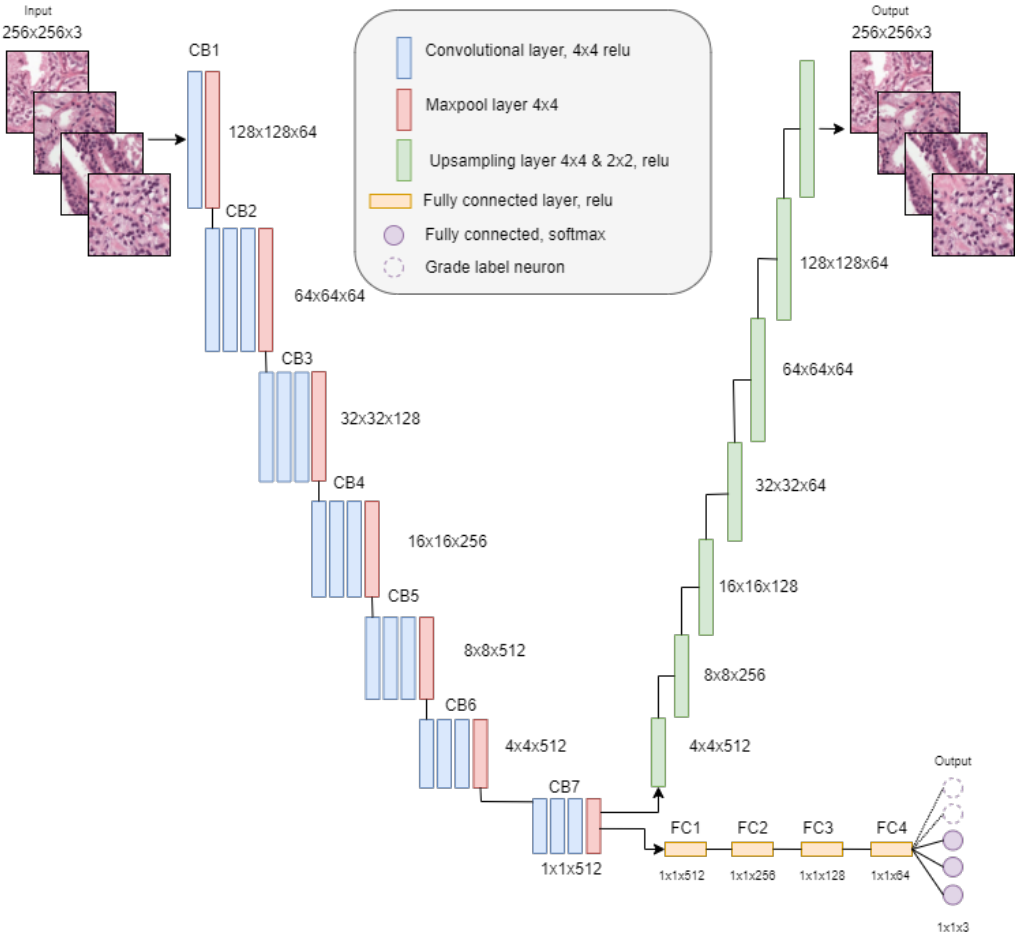


Figure 4.5. Architecture of CNN_{dual} comprised of an encoder part downsampling the input into a 1x1x512 feature space, from where the information is split to both upsample and reconstruct the image, and classify the image into three classes (five using grade labels)

The main objective for CNN_{dual} was to learn distinct information about the the tissue classes, such that the corresponding classes could be separated for subsequent feature analysis. For this reason seven convolutional blocks were used as feature extractors, increasingly downsampling the image dimensions, while increasing the amount of feature maps (encoder). Initially, one convolutional layer with 64 filters was used with a maxpooling layer, stemming from the VGG16 architecture, for extracting many low-level features such as edges, corners and shapes. To enable a convolution with the entire input and keep the dimensions of height and width, a stride of 1 was defined. This means that the filter is moved across the featuremap or input image by the stride. The depth of the output featuremaps increased with the amount of filters used for the concerning convolution. The subsequent layers used three convolutional blocks following a maxpooling layer. This allowed the model to sequentially combine features into more complex features for differentiating between cancer, benign and stroma. [Chollet et al., 2018] An additional dropout component was used in the CB5 and CB6 with dropout rate 0.1, as means of preventing overfitting. [Reinholdt, 2019] The maxpooling layers handled the downsampling, while retaining the best match from the convolution from the featuremaps. The concluding two outputs were comprised of both fully connected layers and upsampling layers. The fully connected layers were used for classification in addition to FS analysis during post-processing for assessment of feature relation between patches. The fully connected layers provided the ability for the model to use the values of the entire compressed information in FS, and thus a flow of every feature space value between the input and output classification. This created flexibility in the classification which was desired. [Chollet et al., 2018] At the end of every convolution and concluding fully connected output, an activation function was applied. For the convolution this was the ReLU activation function. This was introduced as it provided non-linearity to the model, which can aid in convergence. The softmax activation function was used in the fully connected output, to compute a class probability prediction. [Nwankpa et al., 2018] From the FS the information was upsampled through a sequence of upsampling layers (decoder), concluding the architecture by reconstructing the input images. The encoder-decoder architecture is commonly used for feature learning, because it is forced to prioritizes the most useful information, when decoding into a reconstruction. As such, it was useful for learning the morphological structures of the tissue. [Goodfellow et al., 2016]

4.3 Training CNN_{dual}

For CNN_{dual} to gain better reconstruction and classification performance, some improvements to the networks weights and bias' were done during training. The output of a layer is determined by the value of the weights and bias', and thus the resulting prediction depended on the output of all preceding layers outputs. Therefore, the training of CNN_{dual} consisted of finding the set of weights and bias' such that the output was as close to the true label as possible. In order to quantify the difference between the models prediction and the true label, a loss function was defined. To avoid any symmetric outputs from the convolutional and fully connected layers, the weights were initialized from a random normal distribution. [Goodfellow et al., 2016] All weights and bias' were updated in small increments using the learning rate based on the loss, using random batches of data with a stochastic gradient descent optimization algorithm. The learning rate can prevent the network of finding a minimum of the loss function, if the learning rate is to large. Conversely, if it is to small it could get stuck at a local minimum, and is thus not the best set of parameters for the optimization problem. This can be solved by an adaptive optimizer. The entirety of the training is summarized in the following steps:

- Step 1: Retrieve a predefined batch of data
- Step 2: Process that batch through the network to get a prediction
- Step 3: Calculate the loss between the prediction and true label
- Step 4: Calculate the gradient of the loss with respect to the weights and bias' of the model using backpropagation
- Step 5: Update the weights and bias' proportionally to the learning rate in the direction of the gradient
- Step 6: Repeat Step 1-5 until the model is converged

[Chollet et al., 2018]

Since CNN_{dual} had two different outputs comprised of a reconstruction and a classification, two corresponding loss terms were specified, which is elaborated in the following sections.

4.3.1 Loss functions

The Mean squared error loss (MSE) is commonly used in autoencoders, where the objective is to reconstruct a given input image, which is principally the same objective for the output of CNN_{dual}. Specifically, the MSE is used in regression problems, which fits with the ReLU activation function in the output layer of CNN_{dual}, that approximates all individual pixel values in the reconstructed image. [Goodfellow et al., 2016] The MSE is defined as the summation of the N squared differences between the true pixel intensity y and the predicted pixel intensity $\widehat{y}$, shown in equation 4.1

$$MSE = \sum_{i}^{N} |y_i - \widehat{y}_i|^2 \tag{4.1}$$

[Chollet et al., 2018]

Categorical crossentropy

The multiclass classificaiton problem in CNN_{dual} means that the output can only take a fixed number of possible values, which was why the categorical crossentropy loss function was used. The output was a measure of distance between the predicted and true probability distribution. The true distribution was presented to the loss function as a one-hot encoded vector, where only one entry was 1, corresponding to what class the concerning input belongs to. This fits well with the softmax activation function, forcing the last layer of CNN_{dual} to output a probability distribution. Chollet et al. [2018] As such the categorical crossentropy loss function can be formulated as in equation 4.2, where CE is the sum of N predicted probability distributions j and $\widehat{j}$ is the true probability distribution.

$$CE = -\sum_{i=1}^{N}(j_i \cdot log(\widehat{j}_i)) \tag{4.2}$$

[Koech, 2022]

Optimizer

The Adaptive Moment Estimation (Adam) optimizer computes individual gradients for every parameter, which is useful in models with many parameters, as it can be hard to converge. The adam optimizer computes a first moment and a second moment. The first moment is a exponential moving average term where the new weight update is controlled by the average exponential of the previous weight, effectively reducing the learning rate, when reaching the minimum of the loss function. The second moment comes from the optimization algorithm called Root Mean Squared Propagation (RMSprop). The principles of RMSprop is to keep a moving average of the squared loss with respect to the weights and bias' in the network. [Kingma and Ba, 2015; Bushaev, 2022]

The mathematical approach of the first and second momentum is described in appendix B. The collective effect allows for a higher learning rates, which was why it was used in this study. [Kingma and Ba, 2015]

4.3.2 Training hardware and hyperparameters

The performance of CNN_{dual} was monitored, by using a validation dataset for reducing overfitting. Specifically, when the validation loss did not decrease for a predetermined amount of epochs the training was stopped and the model was saved at that epoch.

During initial development phases of the model, the RTX 3060 ti GPU was used, which provided an opportunity for exploring many different machine learning architectures, before reaching the final and best one. The final model was trained on Aalborg University's computing cluster holding 32 Tesla V100-SXM3 GPUs. The model was set to train for 300 epochs, with 20 epochs early stopping, since that indicated overfitting. Moreover, 20 epochs was chosen due to preliminary results showing a decreasing validation loss after 10 epochs of overfitting. Finally, the batch size was 64.

4.3.3 Performance evaluation of classification and image reconstruction

Evaluation of classification performance

A confusion matrix summarizes the performance of a classification. For a multiclass classification problem, the confusion matrix can be summarized as in figure 4.6. The diagonal line depicts the true positives (TP), while the off-diagonal depicts the mis-classified samples. The false negative (FN) of e.g class A, can be calculated by adding E_{ab} and E_{ac}. The false positive for a predicted class is simply calculated by adding the error of a row. [Tharwat, 2018; Mohajon, 2020]

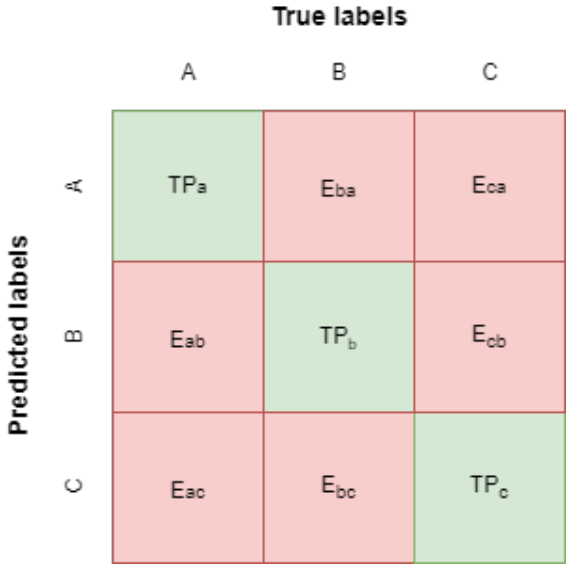


Figure 4.6. Multiclass confusion matrix for a three class classification problem. TP_{a-c} depicts the true positives for the concerning class, and $E_{(a-c)-(a-c)}$ depicts the mis-classified samples

It was important to quantify the classification, as the performance reflected the models ability to separate the classes in feature space. Since there was a class imbalance, the accuracy favors predictions of the class containing the most images, why the precision, recall and F1-score was used. [Korstanje, 2021; Tharwat, 2018] It was important to quantify the models predictive performance, as it was expected that it was more difficult for the model to separate benign features from GS3 features, than benign features from GS5 features. As such the precision, recall and $F_1 score$ was calculated as in equation 4.3, 4.4 and 4.5

$$Precision = TP/(TP + FP) \tag{4.3}$$

The precision metric explains the proportion of patches predicted as the evaluated class (TP) in relation all patches in that class.

$$Recall = TP/(TP + FN) \tag{4.4}$$

The Recall metric explains the proportion of patches predicted as a class, belonging to the evaluated class.

$$F_1 score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{4.5}$$

The $F_1 score$ computes the average of precision and recall and represents the harmonic mean between precision and recall.

[Tharwat, 2018]

Evaluation of recontruction performance

The Mean Squared Erorr (MSE) is a simple quantitative measure for comparing two images similarities, which was done on every pixel intensity of the predicted H&E patches and the true H&E patches. [Wang and Bovik, 2009] The equation for MSE is principally the same as described by the MSE loss function, and the reader is referred to equation 4.1 for clarification. The only difference was that $\frac{1}{N}$ was multiplied and thus the average MSE was calculated for the validation.

4.4 Differentiation of cancer sub-types with feature space analysis

Since the feature space of the model was to be used for analysis, the second step of the framework was done post training of CNN_{dual}. This entailed using the test data (450 benign, stroma, GS3, GS4 and GS5) and the reason for this was two fold. First, the test data was used for assessing the classification and reconstruction performance of CNN_{dual}. Second, the PPS output should depict a generalized performance, and thus the need for using the test data. The methods used for differentiating cancer sub-types were PCA, K-means clustering and the developed algorithm called PPS, depicted in figure 4.7.

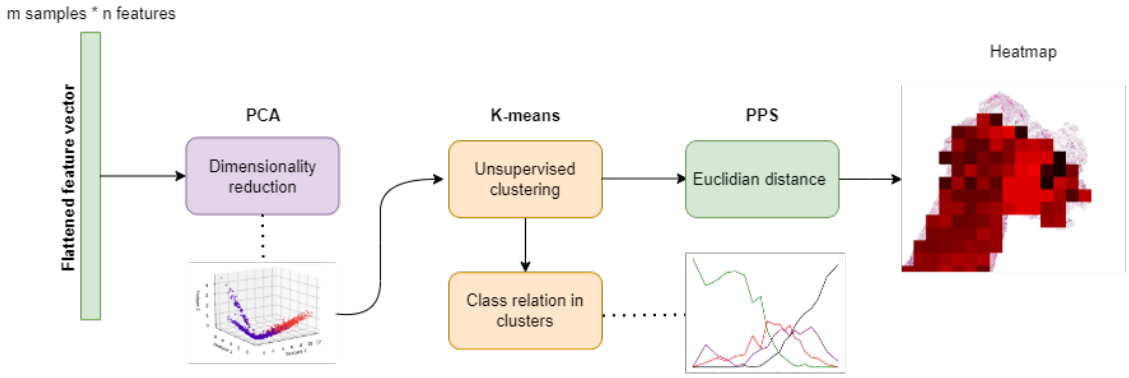


Figure 4.7. Flow diagram depicting the post-processing steps.

Initially, the flattened feature space was reduced by using a PCA. PCA is often employed as a means of finding how many features are sufficient to explain the variance of a dataset, thus

reducing the dimensions of the data set to effeciate eg. a machine learning performance. [Semmlow and Griffel, 2004]. For this study, the reason for PCA was two fold. First, the PCA was used for visualizing the relative placement of benign, stroma, GS3, GS4 and GS5 samples in feature space, which was used for revealing possible sub-groups. Note, when using the PCA the values are moved to the center around origin, which alters the position of the individual features and thus is not the natural placement of that feature in feature space. A theoretical explanation is found in appendix B. For this reason, the PCA was used to find original features, for revealing naturally occurring sub-groups, showing any tissue differentiation. Second, the visualization was used for devising a method for quantifying tissue differentiation. A threshold of $> 90\%$ explained variance was set, for retaining most information within the dataset, which then was the threshold for reducing the dimensions.

For quantifying sub-groups within the reduced representation of the dataset, the K-means clustering method was used. This was used as it is a method for finding similar points in a dataset, see appendix B for further explanation, which was appropriate, when searching for possible sub-groups in a dataset. [Bharadwaj et al., 2021] As an exploratory step, 10, 25 and 50 k-clusters were tested. This step was based on the intuition that the aggressiveness of cancer is continues and not discrete, meaning that the transition from one GS to the next in feature space, could be smooth rather than sharp. Thus the more k-clusters used, the greater the resolution of the transition between GSs.

4.4.1 Accumulated euclidean distance measure algorithm (PPS)

For quantifying the tissue differentiation the euclidean distance measure was used between the k-means cluster centroids on the reduced flattened feature space. The centroids were used as it represented the average feature of a cluster and thus the average placement in feature space. The euclidean distance was used as it calculates the straight line distance from one point to another point in an k-dimensional space. The euclidean distance in k dimensions can be formulated as:

$$dist(a,b) = \left(\sum_{i=1}^{d} |a_i - b_i|^k \right)^{1/k} \tag{4.6}$$

[Bharadwaj et al., 2021]

The intuition behind the PPS algorithm was that the distance from benign tissue would increase as the tissue became more differentiated and thus possessed more distinct features. For this reason, the extent of the following condition was tested, where $dist(B, GS5)$ is the accumulated euclidean distance from benign to GS5 tissue etc.

$$dist(B, GS5) > dist(B, GS4) > dist(B, GS3) \tag{4.7}$$

Since the accumulated distance calculated by the PPS algorithm was conducted on the test data, the resulting distances to the cluster centroids functioned as a "distance-look-up-table" in where the accumulated distance and corresponding cluster centroid was saved.

This was useful when using the remaining GS3, GS4 and GS5 data from the sorting algorithm in k-means cluster prediction, since the cluster prediction had an associated accumulated distance.

In addition, the label distribution of every class confined in every cluster was calculated and saved together with the accumulated distance for the concerning cluster. This was done as means of providing additional information about the distribution of benign, stroma, GS3, GS4 and GS5 in comparison with a certain image. Moreover, it was a useful addition to the accumulated distance, because it could indicate if the majority of the concerning image was one GS or for example a mixture of scores.

The principles of the accumulated distance algorithm is elaborated in the following code snippet and the corresponding figure 4.8.

```
current_cluster = benign cluster centroid  farthest  away from cancer tissue
accumulated_distance = 0
distance_dictionary = {}
label_distribution_dictionary = {}
for  i  in  range of( clusters  - 1)
    for  j  in  range of( clusters )
        distance  = euclidean_distance(current_cluster - clusters[j])
        if  0 < distance < previous distance
            if  clusters [j]  in  distance_dictionary
                continue:
            else
                current_cluster  = clusters [j]
    accumulated_distance += distance
    distance_dictionary[current_cluster]  = accumulated_distance
    B,GS3,GS4,GS5 = cluster_class_distribution(current_cluster)
    label_distribution_dictionary[current_cluster]  = B,GS3,GS4,GS5
```

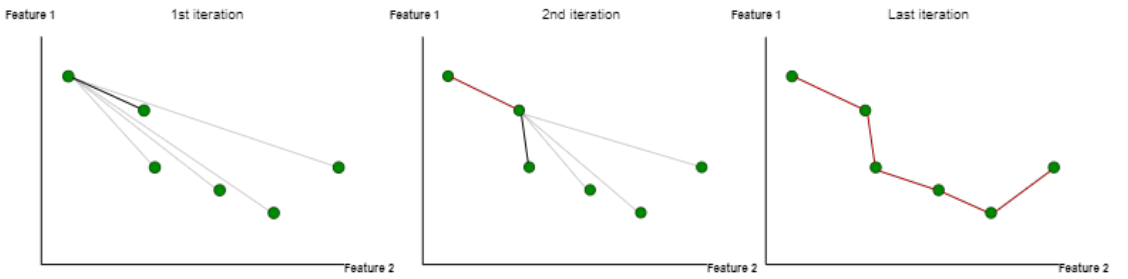


Figure 4.8. Figurative example of how the accumulated distance was calculated. The green dots are centroids of k-clusters, while the red lines are the euclidean distances between two centroids.

The algorithm starts at the centroid of benign tissue farthest away from cancer tissue. From there the euclidean distance to every other centroid is calculated, finding the shortest distance. The accumulated distance and corresponding centroid is saved in a dictionary, which is used in the following iteration for checking if the concerning centroid already had been saved. Effectively, this prevents the algorithm from looping infinitely. In addition, the label distribution was calculated in a function, that finds the total number of data points and the concerning labels in a cluster, and calculates the relative label distribution. The corresponding cluster and label distribution was saved in a dictionary. The PPS is concluded when the initiating for-loop is iterated the amount of clusters - 1 times.

Grade distance Heat-map

The final component in quantifying the gleason score is generating a heat-map using the distances derived by the PPS. Practically, the remaining input images containing either GS3, GS4 or GS5 of a WSI were saved and organized such that they could be reassembled with the corresponding WSI again. Next, every input image, was processed through the entire methodical pipeline, which entailed the following steps.

- Step 1: Predict on patch using CNN_{dual}
- Step 2: Extract feature space and perform PCA
- Step 3: Perform k-means prediction on reduced feature set
- Step 4: Attribute a distance with the cluster prediction for the concerning patch

Finally, the red channel of the patches were replaced with the corresponding distances obtained from step 4, where the minimal and maximal distance were 0 and 1 respectively. To be able to see the color contrast in the heat-maps, as in figure 4.9, the product between the distance and 255 was calculated, resulting in a spectrum of 8bit red colors, brighter and darker red being equivalent to more or less aggressive cancers respectively.

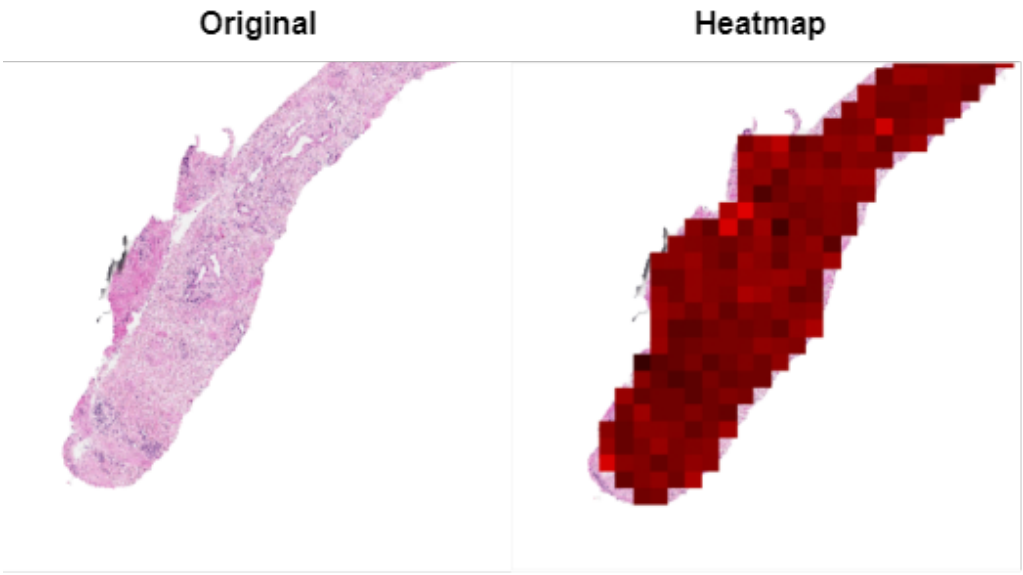


Figure 4.9. An original WSI with the corresponding heatmap

The patches were then placed in accordance with the original placement, for visualizing the area and aggressiveness of cancer tissue.

Results 5

The results from the implemented methods are visualized in this chapter. First the results from the classification and reconstruction performance of CNN_{dual} is shown. Next the features from the PCA are visualized and elaborated upon. 10, 25 and 50 k-means clusters are shown, followed by the resulting label distribution and boxplots of the accumulated distance from the PPS algorithm. Finally, Three heatmaps of selected WSIs containing GS3, GS4 and GS5 patches are shown.

5.1 Results from training CNN_{dual}

5.1.1 Training and validation losses

The losses from CNN_{dual} with and without grade labels (denoted as $CNN_{dualWGL}$ and $CNN_{dualNGL}$ respectively) during training are shown i figure 5.1. The training was stopped around epoch 70 for both the training of $CNN_{dualWGL}$ and $CNN_{dualNGL}$, by the implemented early stopping algorithm. Therefore the model starts overfitting to the training data at approximately epoch 50, where a model i saved.

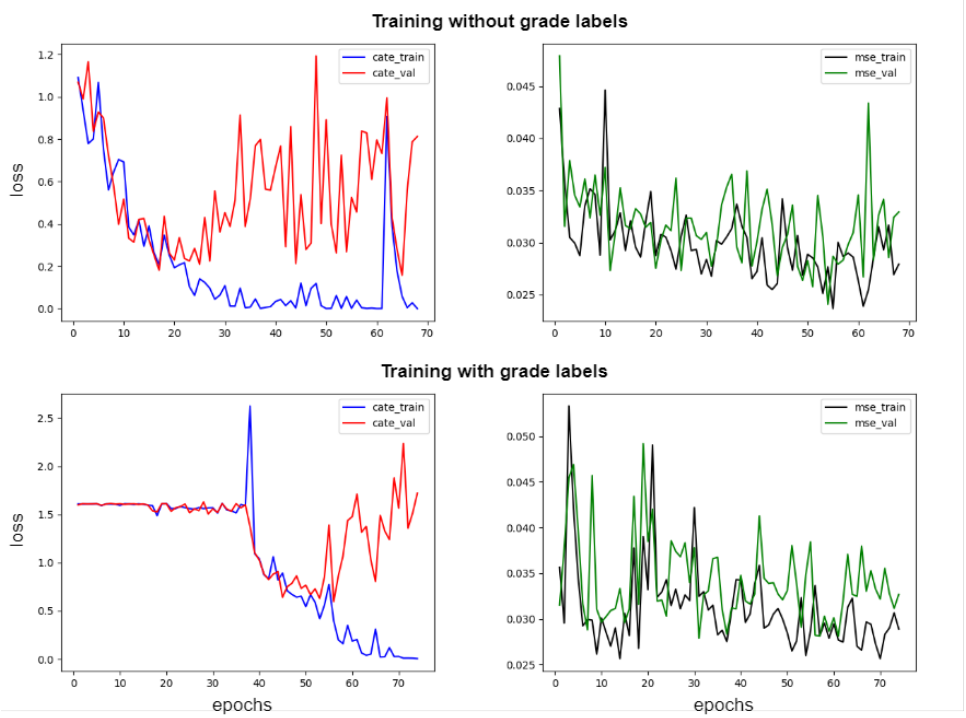


Figure 5.1. Loss curves from training CNN_{dual}. The leftmost figures depicts losses from the classification, while the right most depicts the losses from the reconstruction.)

From figure 5.1 it can be seen that it is the categorical training loss, that continuously decreases and thus overfits, while the mse validation and training loss continuously decreases over time. A greater overfitting is observed in the categorical loss with grade labels, where the validation loss is approximately 1.6, while the validation loss is 0.8 in $CNN_{dualNGL}$. All mse losses are similar starting and ending at a loss of approximately 0.05 and 0.03 respectively, while fluctuating throughout the training. The categorical training loss from training $CNN_{dualNGL}$ and $CNN_{dualWGL}$ converge.

5.1.2 Classification performance

The confusion matrix in figure 5.2, is used to summarize the performance of CNN_{dual} and shows the ratio of true labels and predicted labels from using the test images.

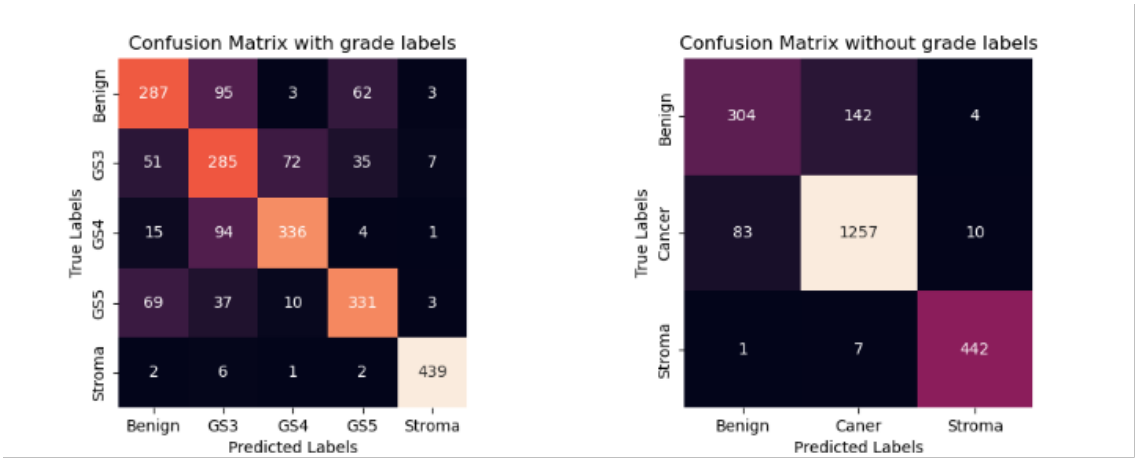


Figure 5.2. Confusion-matrix showing the model predictions of CNN_{dual} using the test images and corresponding true labels with grade labels (left) and without grade labels (right).

Figure 5.2 reveals that the classification performance of stroma is similar between performances. It becomes more difficult for the model using grade labels to classify gleason scores as opposed to classifying cancer. It is notably difficult for the model to classify benign tissue in both cases, where roughly 30% of benign tissue are classified as cancer. To analyze the false positives and false negatives, the Precision, Recall and F1-score were used.

Precision of CNN_{dual}			
		with grade labels	without grade labels
Benign		0.63	0.67
Cancer	GS3	0.63	0.93
	GS4	0.74	
	GS5	0.73	
Stroma		0.97	0.98

Table 5.1. Table showing the precision of CNN_{dual} with and without grading labels.

Recall of CNN_{dual}			
		with grade labels	without grade labels
Benign		0.67	0.78
Cancer	GS3	0.55	0.89
	GS4	0.79	
	GS5	0.76	
Stroma		0.97	0.96

Table 5.2. Table showing the recall of CNN_{dual} with and without grading labels.

$F_1 score$ of CNN_{dual}			
		with grade labels	without grade labels
Benign		0.65	0.72
Cancer	GS3	0.59	0.91
	GS4	0.76	
	GS5	0.74	
Stroma		0.97	0.97

Table 5.3. Table showing the accuracy of classification performance with and without grade labels respectively.

The F1-score of table 5.3 show that stroma was 97% and 98% for $CNN_{dualNGL}$ and $CNN_{dualWGL}$ respectively. This indicates that precision and recall were high. In addition, the model performance between cancer and GS3, GS4 and GS5 are notably different. The F1-score of cancer was 0.91 were superior to the F1-scores of GS3, GS4 and GS5 being 0.59, 0,76 and 0,74 respectively. The lowest of which was the GS3 F1-score, which is in fact also the lowest in precision and recall. A precision of 93% is observed for $CNN_{dualNGL}$, which means that 7% of the annotated cancer patches were classified as either, stroma or benign tissue. Generally the precision, recall and F1-score reveal that, $CNN_{dualNGL}$ were superior to $CNN_{dualWGL}$.

Examples of classifications from $CNN_{dualNGL}$ and $CNN_{dualWGL}$

Since there were misclassification from $CNN_{dualNGL}$ and $CNN_{dualWGL}$ using the test images, some correctly classified and misclassified samples are shown in figure 5.4 and 5.3

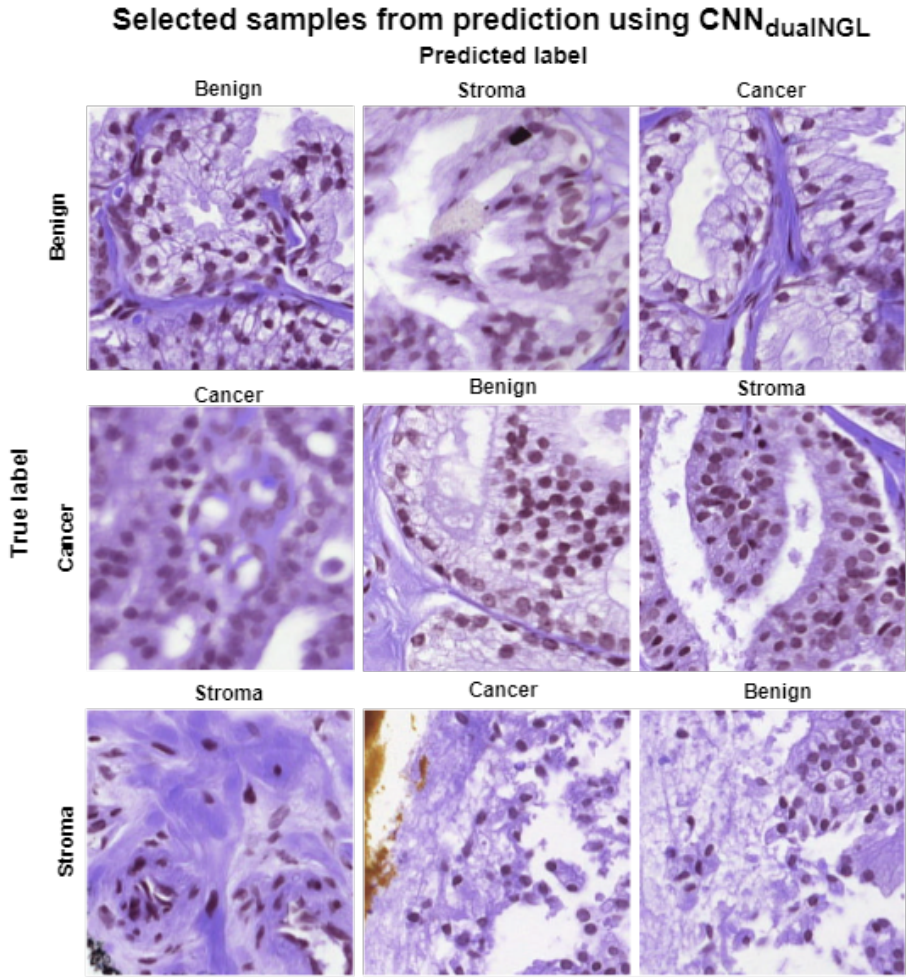


Figure 5.3. Grid of sclassifications from $CNN_{dualNGL}$ using the test images

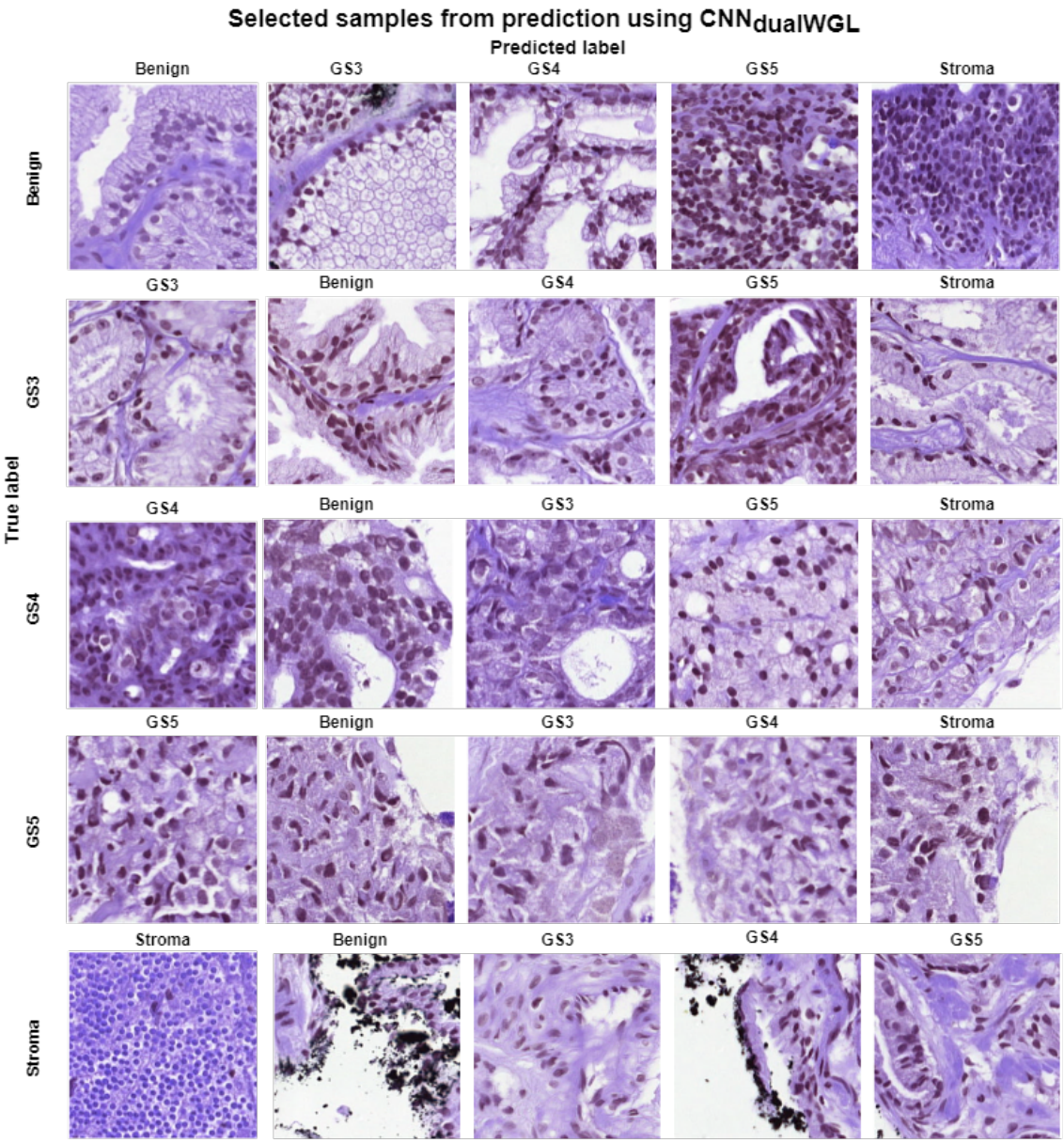


Figure 5.4. Grid of classifications from $CNN_{dualWGL}$ using the test images

5.1.3 Reconstruction performance

The reconstruction performance of $CNN_{dualNGL}$ and $CNN_{dualWGL}$ are visualized in 5.5 to analyze the predictive performance of the models between classes.

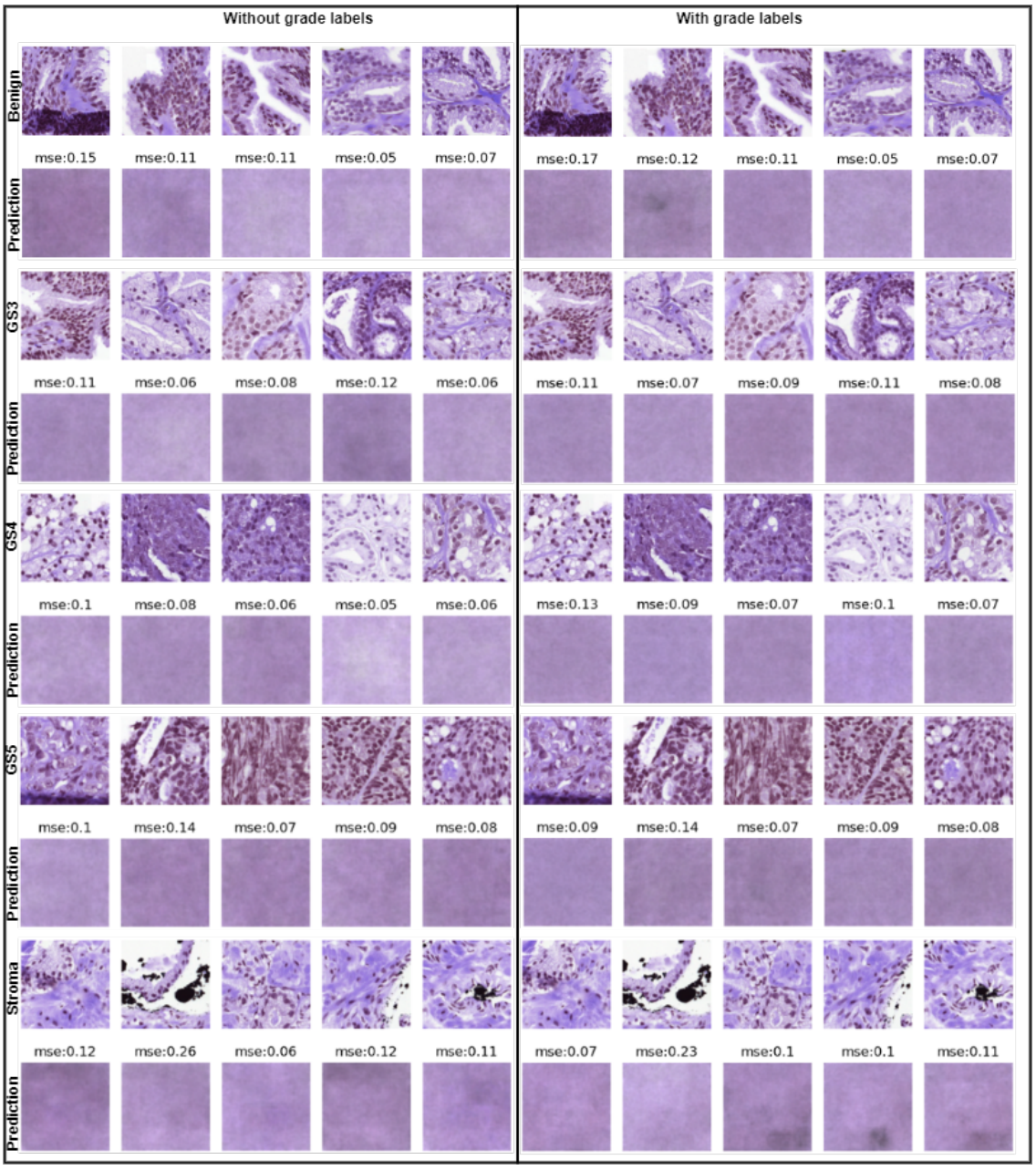


Figure 5.5. A grid of five test images for every class and the corresponding prediction using $CNN_{dualNGL}$ (left) and $CNN_{dualWGL}$ (right)

The reconstruction performance between classes of $CNN_{dualNGL}$ and $CNN_{dualWGL}$ shown in figure 5.5 was similar, producing shades of purple blurry images where poor structural information is seen within the reconstruction. It is near impossible to visually differentiate between classes from the reconstruction. The average MSE for all test patches are shown in table 5.4, and reveals that the lowest and highest average mse for CNN_{dual} with and without grade labels, were benign and stroma patches respectively.

Average MSE					
	Benign	GS3	GS4	GS5	Stroma
without grade labels	0.13	0.20	0.21	0.20	0.26
with grade labels	0.14	0.19	0.20	0.19	0.25

Table 5.4. Table describing the average mse for every class using CNN_{dual} trained with and without grade labels

5.2 Results from PCA, K-means and PPS-algortihm

During preliminary results it was found that stroma was primarily responsible for explaining the variance within the first principle component. Consequently, the benign, GS3, GS4 and GS5 tissue features of that principle component were clustered together yielding poor results. The features of the principle components explaining >90% of the variance with stroma are explained in appendix C shown in figure C.1. Therefore the stroma patches were not used in subsequent PCA, k-means clustering and in PPS algorithm, to focus on features from cancer and benign tissue.

5.2.1 PCA results

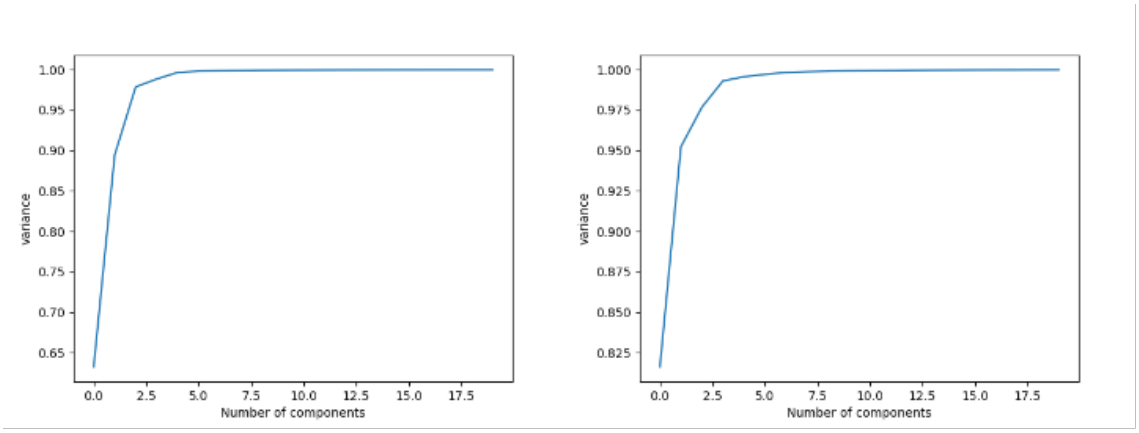


Figure 5.6. Variance as a function of principle components for $CNN_{dualWGL}$ (left figure) and $CNN_{dualNGL}$ (right figure)

PCA on feature space of CNN_{dual}							
With grade labels				Without grade labels			
PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
63.5%	**26.6%**	8.4%	0.9%	**81.6%**	**13.5%**	0.2%	0.1%

Table 5.5. Tabel of four principle components (PC) from CNN_{dual} trained with and without grade labels. The green marked PCs explain > 90% of variance.

Figure 5.6 shows that approximately five features explain 100% of the variance from the FS of $CNN_{dualWGL}$ and $CNN_{dualNGL}$. Table 5.5 reveals that the first two principle components of $CNN_{dualWGL}$ and $CNN_{dualNGL}$ explained 90.1% and 95.1% of the variance

respectively. The distribution of variance notably differs within PC1 and PC2. Of the total variance in PC1 and PC2, $CNN_{dualNGL}$ had most variance in PC1, while $CNN_{dualWGL}$ had roughly 20% less variance in PC1. The features corresponding to PC1 and PC2 are normalized as in equation 5.1 and visualized in figure 5.7

$$data_{norm} = (data - min(data))/(max(data) - min(data)) \tag{5.1}$$

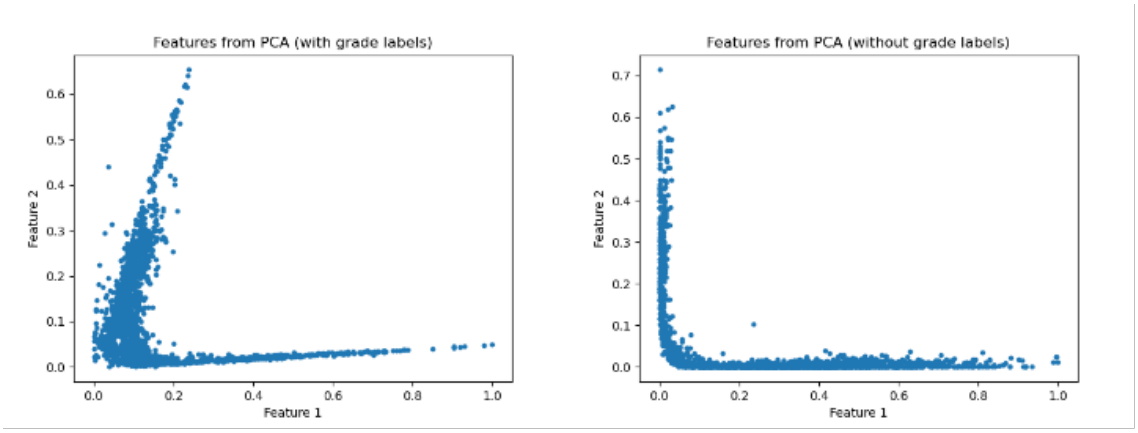


Figure 5.7. The two most explaining features using $CNN_{dualWGL}$ and $CNN_{dualNGL}$ showed in the left and right figure respectively.

From figure 5.7 it can be seen that the features are mostly in one cluster, which means that the features from the corresponding classes overlaps as shown in figure 5.8.

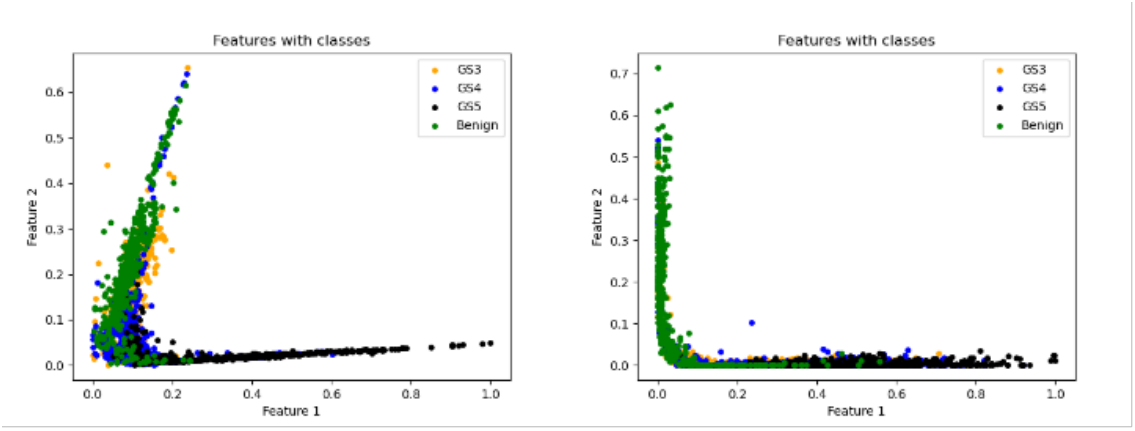


Figure 5.8. The two most explaining features using $CNN_{dualWGL}$ and $CNN_{dualNGL}$ with the corresponding classes showed in the left and right figure respectively

From figure 5.8 it can be seen that the relative placement of the feature values from the respective classes of $CNN_{dualWGL}$ and $CNN_{dualNGL}$ follow a similar pattern on the manifold, where the benign feature values are farther away from the GS5 features. The GS3, GS4 and benign features are clustered together, more so within the features derived from $CNN_{dualWGL}$ than $CNN_{dualNGL}$. The mean tissue feature value for benign, GS3, GS4 and GS5 and corresponding feature value differences are depicted in table 5.6.

The compared differences of benign and GS3, GS4 and GS5 between $CNN_{dualNGL}$ and $CNN_{dualWGL}$ using the test images reveal that the mean feature value difference is greater of $CNN_{dualNGL}$ than of $CNN_{dualWGL}$. This implies that the general tendency is that benign, GS3, GS4 and GS5 features are more separated and thus more distant when using $CNN_{dualNGL}$ than using $CNN_{dualWGL}$, which means that there is no added benefit in using labels for learning the features corresponding to the classes.

Mean feature value of Benign, GS3, GS4 and GS5		
	$CNN_{dualNGL}$	$CNN_{dualWGL}$
Benign (feature 1, feature 2)	(0.051, 0.174)	(0.096, 0.218)
GS3 (feature 1, feature 2)	(0.253, 0.021)	(0.091, 0.159)
GS4 (feature 2, feature 2)	(0.341, 0.021	(0.137, 0.071)
GS5 (feature 1, feature 2)	(0.489, 0.004)	(0.368, 0.031)
Difference between mean Benign feature values and GS3, GS4 and GS5		
Benign vs GS3	(-0.202, 0.153)	(0.005, 0.059)
Benign vs GS4	(-0.290, 0.153)	(-0.041, 0.147)
Benign vs GS5	(-0.438, 0.170)	(-0.272, 0.187)

Table 5.6. Mean feature value and mean feature value difference of benign, GS3, GS4 and GS5

5.2.2 K-means Clustering

The purpose of the K-means clustering was to find similar points and to provide a space from where the PPS algorithm could function. It can be seen from figure 5.9 that 10 and 25 clusters found similar points, but 50 clusters finds single points as clusters both for $CNN_{dualNGL}$ and $CNN_{dualWGL}$. Increasing the cluster resolution to 50 clusters indicates that it is to many for finding sub-groups for this feature value placement.

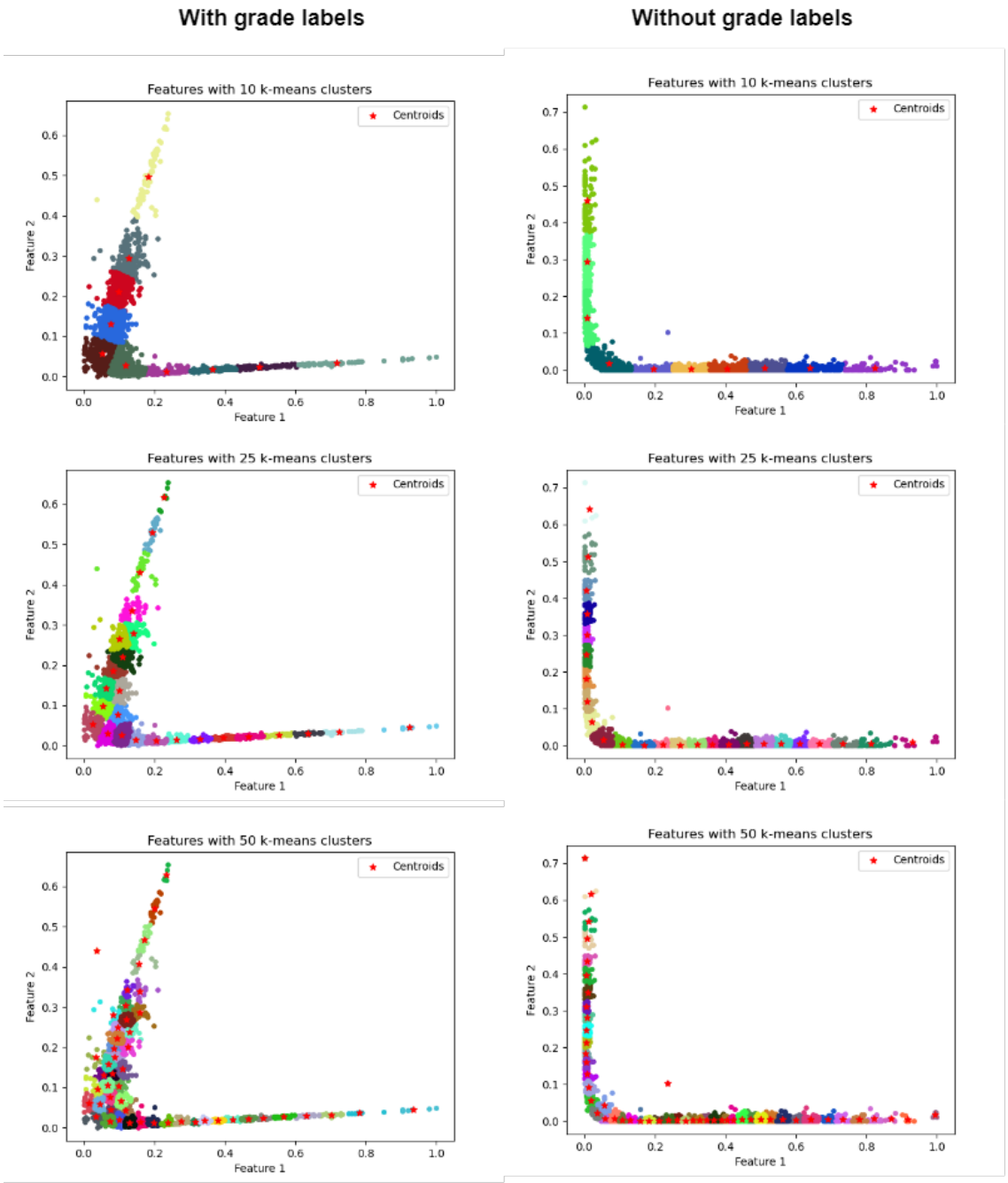


Figure 5.9. 10, 25 and 50 K-means clusters of the two most explaining features placed at the top, middle and bottom row respectively.

5.2.3 Label distribution and Accumulated euclidean distance

The label distribution was used as a tool to analyze if the concerning patch was closest to a cluster containing the majority of one class or a mixture of several classes.

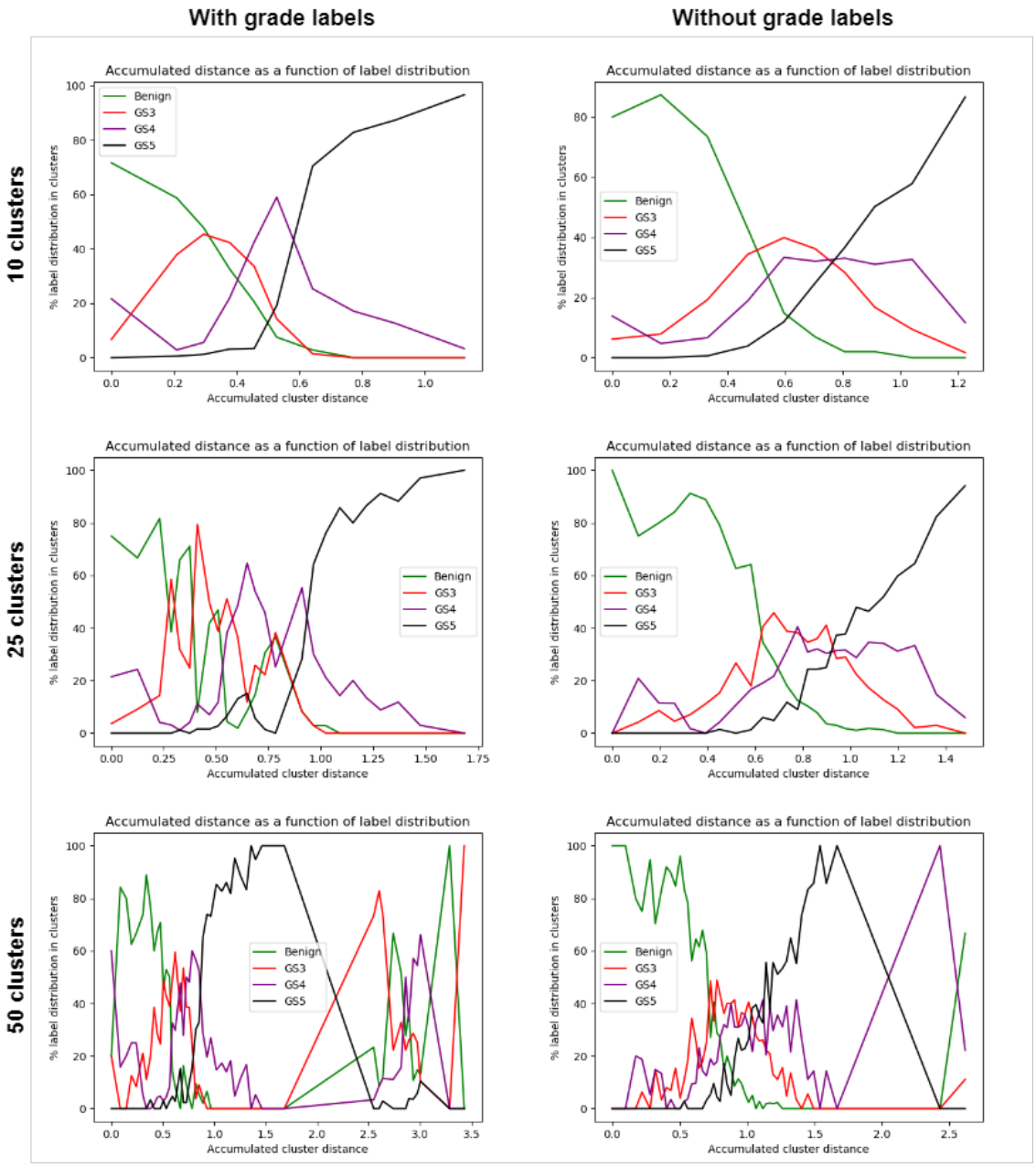


Figure 5.10. Label distribution is clusters as a function of accumulated distance to centroids of 10, 25 and 50 k-means clusters with and without grade labels using the PPS algorithm.

When using 50 clusters the label distribution within the clusters as a function of distance becomes ambiguous. It is expected that the label distribution favors more aggressive cancer, the farther the distance becomes. This is the case for the benign and GS5 tissue features when using 10 and 25 clusters without grade labels. Interestingly, at distance 0 and thus the first cluster, the label distribution containing most benign tissue is 25 clusters and 50 clusters of $CNN_{dualNGL}$, whereas all others contain approximately 20% GS4 tissue features and slightly less GS3 tissue features. The subsequent distances follow the expected label distribution as a function of distance whereas the 50 clusters label distribution both with and without grade labels does not, which is a limitation of the PPS algorithm.

The distances calculated in the boxplots in figure 5.11 are done using the remaining test

patches, ans the saved "distance-look-up-table" from the PPS-algorithm as described in chapter 4, section 4.1 table 4.1.

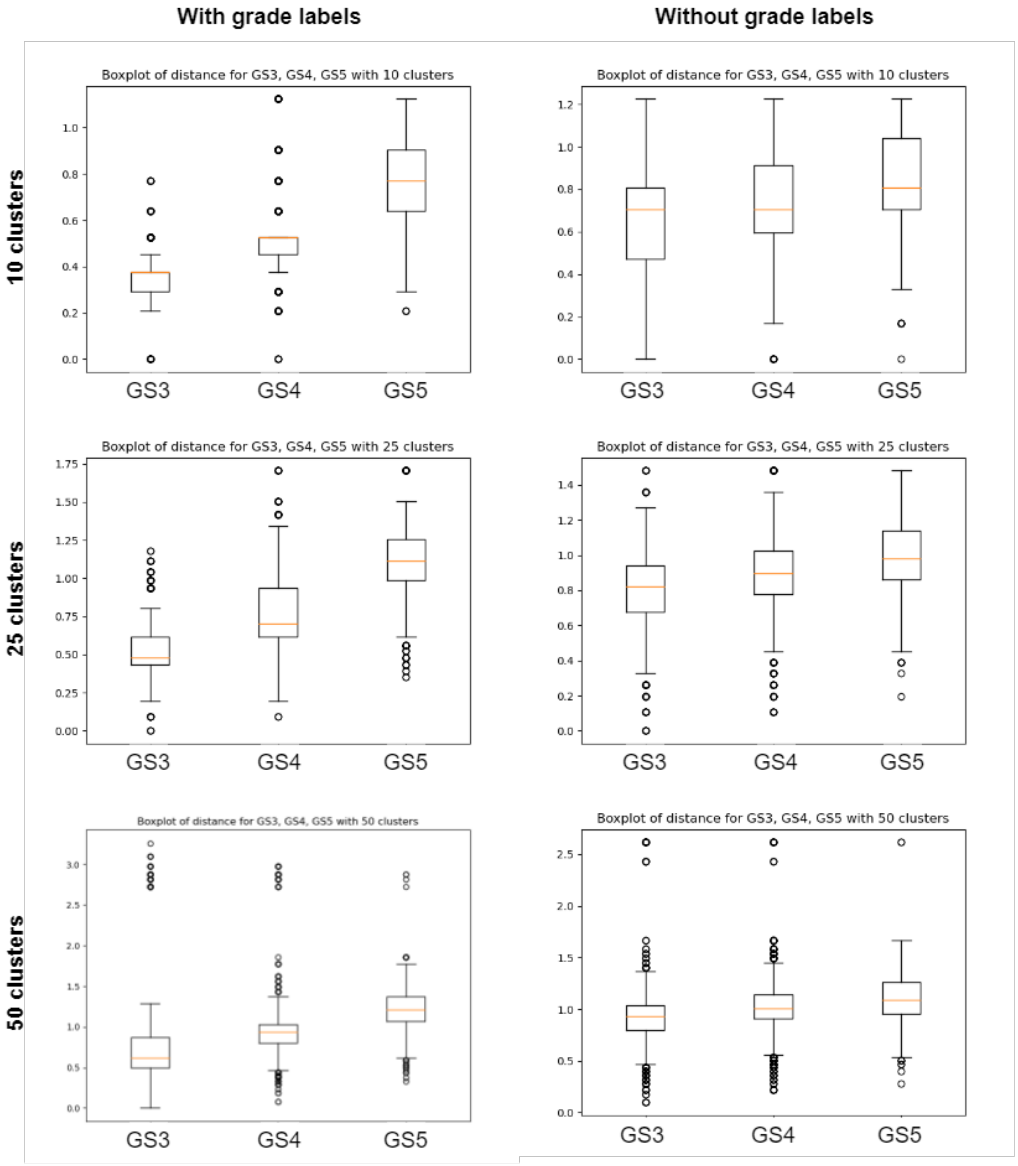


Figure 5.11. Boxplots of the distances from the remaining GS3, GS4 and GS5 test patches using 10, 25 and 50 clusters with and without labels

The boxplots of the distances from the remaining patches in figure 5.11 show interesting results. There is a general tendency for the median value to increase with the GS, except from 10 clusters without grade labels, where the median value of GS3 and GS4 are similar. The boxplots using 25 clusters reveal that without grade labels, the distance median value slightly increase across GS compared to using grade labels, where the median distances are farther apart. Most importantly, this shows that there is a distance difference between GS, when grade labels are not used. For the boxplots using 50 clusters, it is evident from the label distribution as a function of distance using 50 clusters the distances becomes ambiguous and loses meaning when reaching 1.5 and above. This means that it is difficult to interpret distances from GS3, GS4 and GS5 without using grade labels, and GS4 and

GS5 using grade labels. Therefore, using 50 clusters was disregarded in the following results.

5.2.4 Heatmap examples of $CNN_{dualNGL}$ using 10 and 25 cluster distances

Three examples are shown, where the patches for the heatmaps were annotated as GS3, GS4 and GS5 in figure 5.12, 5.13 and 5.14 respectively. The distances derived using the PPS-algorithm using 10 and 25 clusters are used to produce the heatmaps in the figures. The corresponding original and annotated WSI are showed as additional information. A boxplot of the distances derived using 10 and 25 clusters and the corresponding label distribution as a function of distance is also presented.

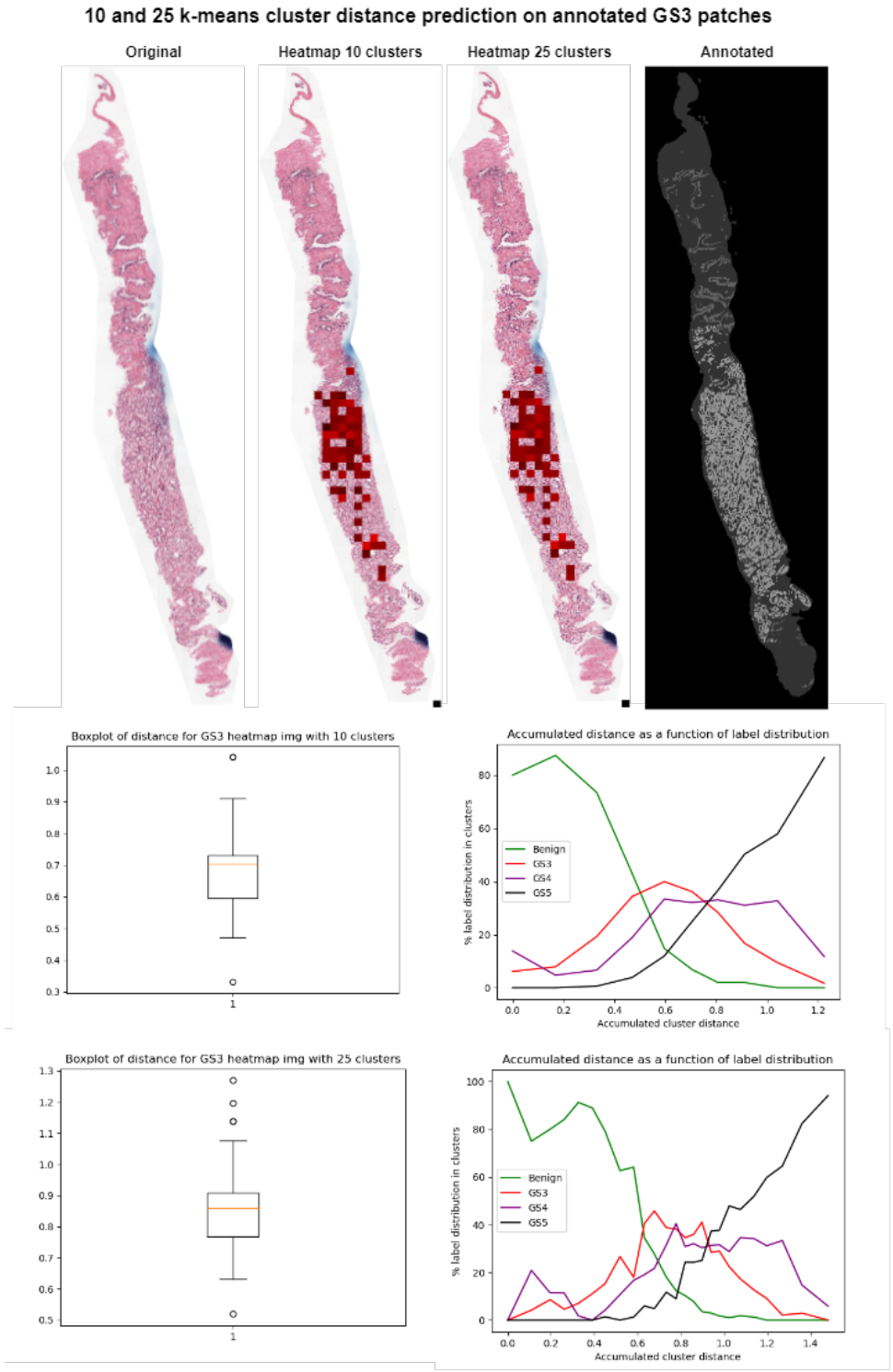


Figure 5.12. Heatmaps derived from 10 and 25 cluster distances of a WSI containing GS3 annotated patches. The original and annotated WSIs are placed adjacent to the heatmaps, with the corresponding boxplot and label distribution placed under.

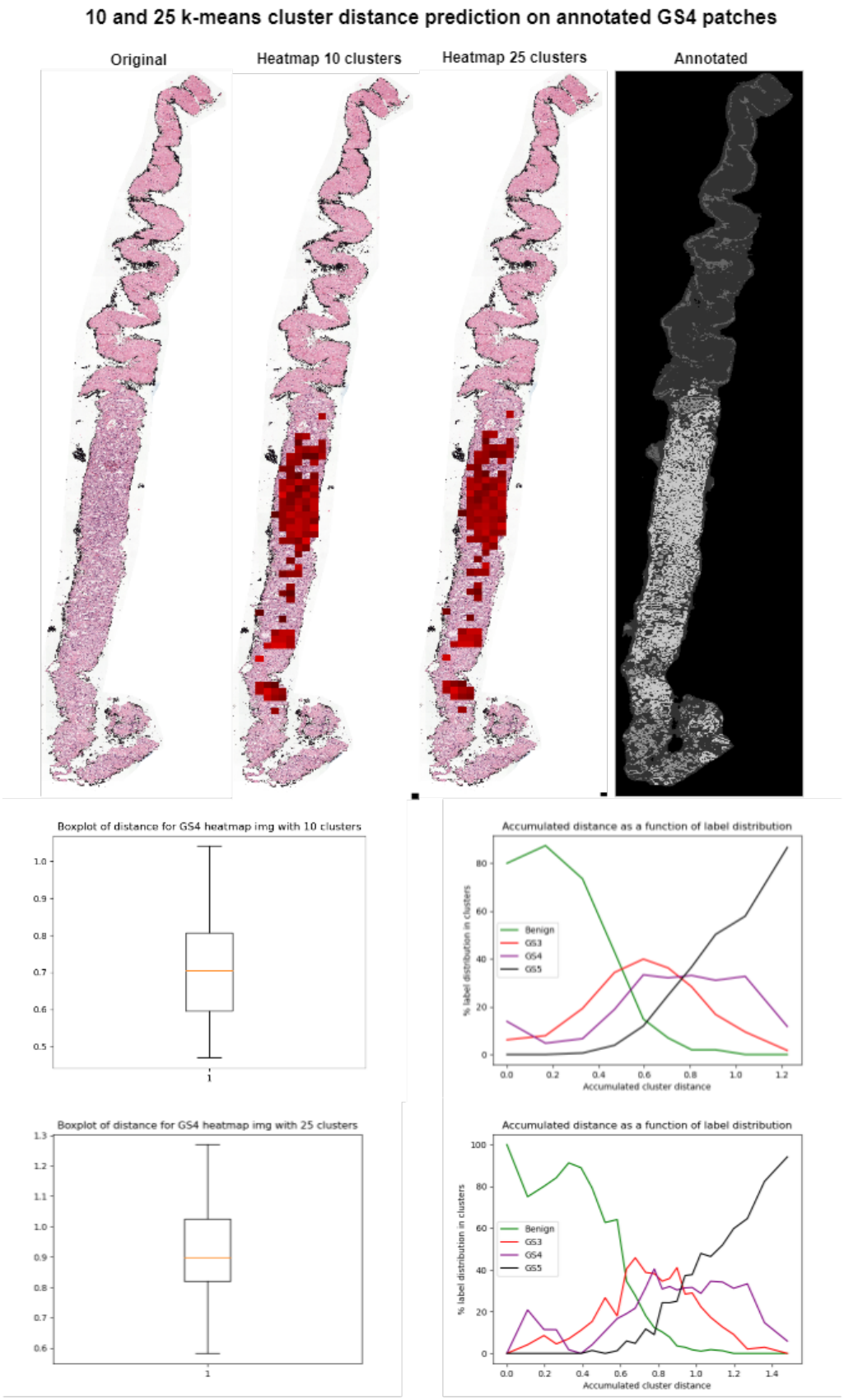


Figure 5.13. Heatmaps derived from 10 and 25 cluster distances of a WSI containing GS4 annotated patches. The original and annotated WSIs are placed adjacent to the heatmaps, with the corresponding boxplot and label distribution placed under.

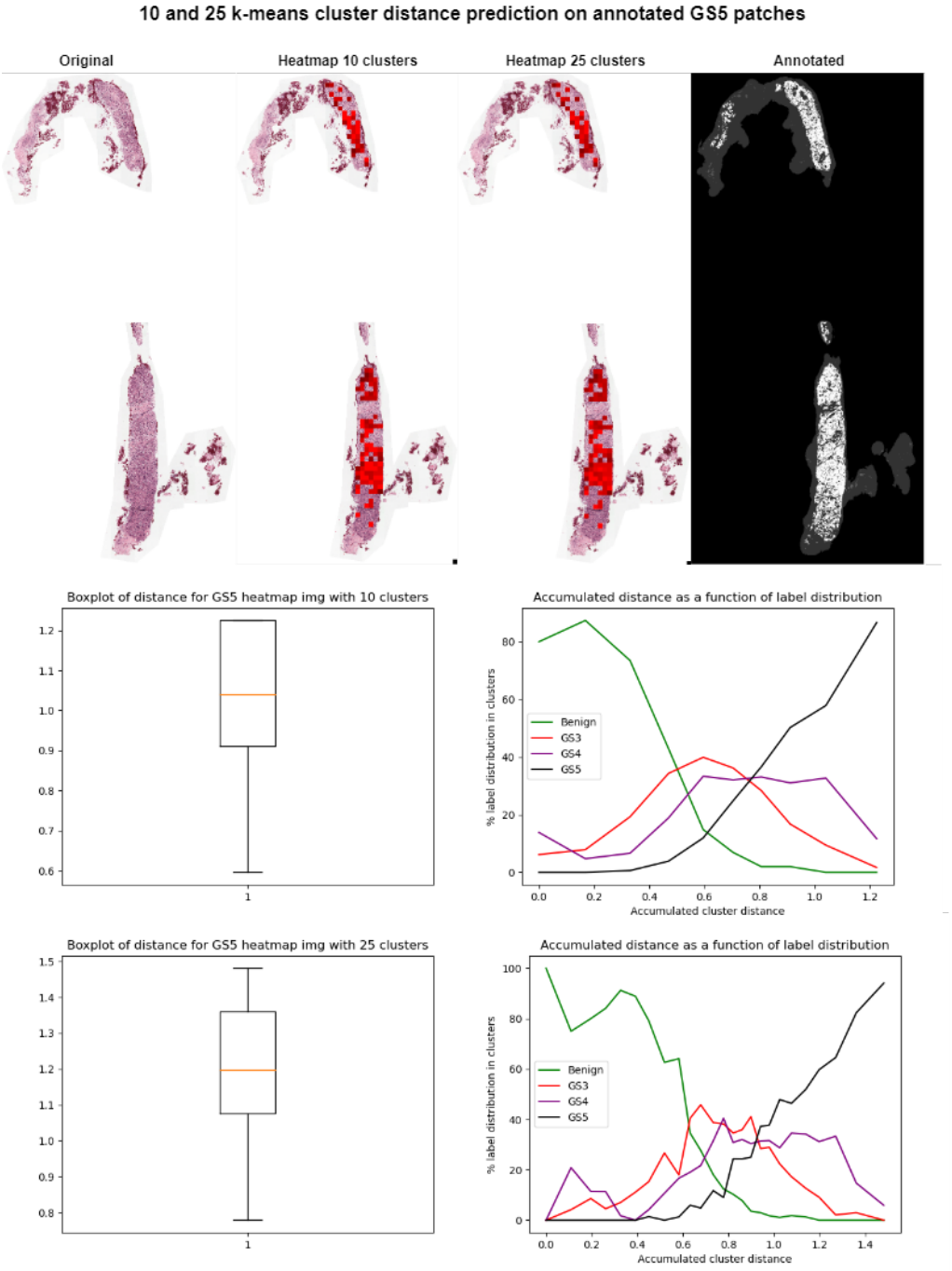


Figure 5.14. Heatmaps derived from 10 and 25 cluster distances of a WSI containing GS5 annotated patches. The original and annotated WSIs are placed adjacent to the heatmaps, with the corresponding boxplot and label distribution placed under.

The heatmaps are visually darker from the GS3 heatmap in figure 5.12 compared with GS4 heatmap in figure 5.13, which is darker compared to GS5 heatmap in figure 5.14. This means that the distances generally are shorter when $CNN_{dualNGL}$ is applied to GS3 image patches compared to GS4 and GS5 image patches. The median values of the boxplots from the WSI class patches using 10 clusters of GS3, GS4 and GS5 are 0.7, 0.7 and 1.05 respectively. While the median values using 25 clusters of GS3, GS4 and GS5 are 0.85, 0.9 and 1.2 respectively. Using 25 clusters is slightly better than using 10 clusters, since

it reveals a greater label distribution resolution, while maintaining an increase in distance with an increase in GS. In addition, increasing the clusters from 10 to 25 results in a label distribution of 100% for benign tissue features at a distance of 0, while using 10 clusters yields 80% benign tissue features at a distance of 0. Which indicated that 25 clusters is a better resolution than 10 clusters.

Max, median and minimum single patch distance

Selected patches contributing to the maximum, median and minimum distances of the boxplots from the heatmaps are visualized in figure 5.15.

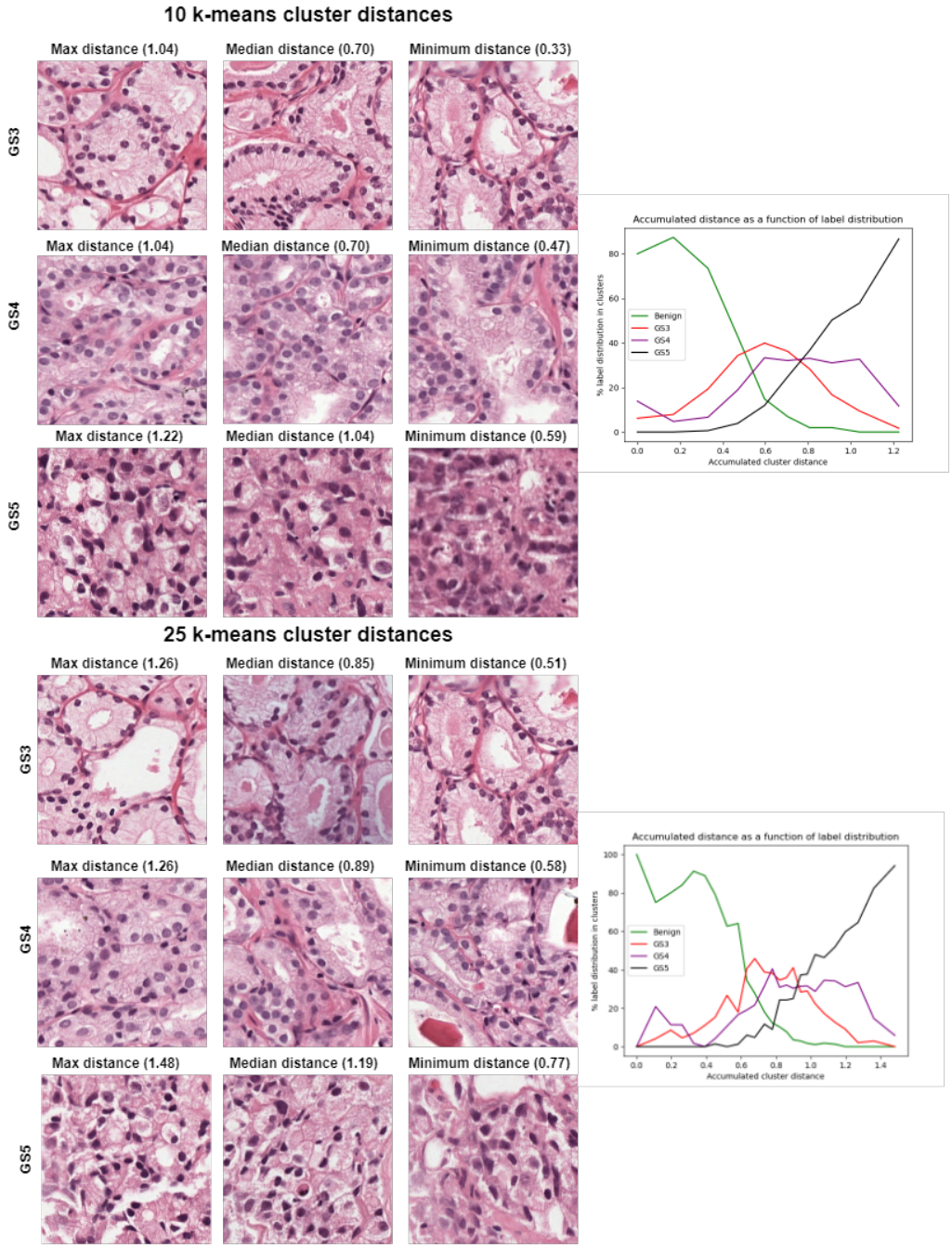


Figure 5.15. Single patches from the heatmaps with a maximum, median and minimum distance prediction for GS3, GS4 and GS5.

Combining the label distribution as a function of distance using 10 clusters, with the maximum (1.04), median (0.7) and minimum (0.33) distances of GS3 shows, that the patch producing the maximum distance yields a label distribution favoring GS5 ($\approx 50\%$) and GS4 ($\approx 35\%$) more than GS3 ($\approx 10\%$). Using the same constellation with 25 clusters, the maximum distance (1.27) produces approximately the same label distribution. The distances for the selected patches are shown in table 5.7.

		GS3	GS4	GS5
10 Clusters	Max	1.04	1.04	1.22
	Median	0.71	0.71	1.04
	Min	0.33	0.47	0.59
25 Clusters	Max	1.27	1.27	1.48
	Median	0.85	0.91	1.20
	Min	0.51	0.58	0.77

Table 5.7. Distances for the selected patches of GS3, GS4 and GS4 using 10 and 25 clusters

Discussion 6

When determining treatment for PCa an accurate GS is important for identifying the correct treatment. The GS is currently determined by pathologists visually analyzing H&E stained WSIs, which are prone to interobserver variability and thus wrongful treatment [Ozkan et al., 2016; Thomsen et al., 2015; Egevad et al., 2011]. Therefore, objective measures are needed to obtain a consistent and accurate GS, to improve patient outcomes. The variation in morphological structures within H&E stained WSIs containing benign and cancerous tissue provides an opportunity for learning a cancer appearance feature space that can be used for grading PCa and overcoming the interobserver variability.

For learning a cancer appearance feature space, WSIs containing benign, GS3, GS4, GS5, and stroma tissue were used and sorted in patches. To mitigate interobserver variability, the GS3, GS4, and GS5 patches were joined in one category and labeled as cancer. By using $CNN_{dualNGL}$ to reconstruct and classify the input patches, the assumption was that the model could learn a FS containing tissue features that naturally separate the cancer-labeled images into GS3, GS4, and GS5 features. To analyze if additional GS information is needed for separating GS features, grade labels were used in $CNN_{dualWGL}$ to compare with the FS of $CNN_{dualNGL}$. Using test images on $CNN_{dualWGL}$ and $CNN_{dualNGL}$, features from FS constituting $> 90\%$ variance were extracted using a PCA, upon which a k-means unsupervised clustering was performed. To evaluate the degree of class affiliation the PPS algorithm was used consisting of two components, which are accumulated euclidean distance to each cluster centroid and the label distribution within clusters.

Cancer appearance feature space

From the extracted and dimension reduced feature space, some feature separation of benign, GS3, GS4, and GS5 were obtained both in $CNN_{dualWGL}$ and $CNN_{dualNGL}$. These results are similar to the study by Bulten and Litjens [2018], where stroma is separated, while benign and cancerous features are entangled. Comparing $CNN_{dualWGL}$ with $CNN_{dualNGL}$, the mean feature value of GS3, GS4, and GS5 of $CNN_{dualWGL}$ were closer to the mean feature value of benign compared to $CNN_{dualNGL}$. Specifically, the mean feature value difference of $CNN_{dualWGL}$ and $CNN_{dualNGL}$ were (0.005, 0.059), (-0.041, 0.147), (-0.272, 0.187) and (-0.202, 0.153), (-0.290, 0.153), (-0.438, 0.170) respectively, which implies that $CNN_{dualNGL}$ learned a greater separation between benign, GS3, GS4, and GS5 features without using the grade labels and there could be no added benefit in using grade labels using the architecture of CNN_{dual}.

Distinction between tissue features

Features corresponding to a higher GS and thus more differentiated cancer were assumed to appear more distant from benign features on the manifold. This is well visualized within the manifold of $CNN_{dualNGL}$ using the test images, where the variance of feature 2 mostly contains benign feature values, and the variance of feature 1 mostly contains GS5 feature values, with the GS3 and GS4 feature values mixed in between. Using the PPS algorithm on the centroids from the K-means clustering revealed that the distribution of benign, GS3, GS4, and GS5 feature values, were organized such that the benign feature values were more distant from GS5 than GS4 and GS3. Specifically, the label distribution as a function of accumulated distance using $CNN_{dualNGL}$ with 10 and 25 k-means clusters, the benign, GS3 and GS4, and GS5 tissue features are mostly present at a distance between 0 to 0.6, 0.6 to 0.9 and, 0.9 to 1.2 (1.4 using 25 k-means clusters) respectively. Nevertheless, it was expected that the PPS algorithm worked well with the GS5 feature values using the test images, due to the sensitivity of 0.97 and average accuracy of 92% shown by [Ryu et al., 2019] and [Gummeson et al., 2017] on GS5 tissue respectively. Interestingly, as the GS becomes intermediate, the GS3 and GS4 label distribution as a function of accumulated distance using 10 and 25 clusters is similar, attaining almost equal label distribution regardless of the accumulated distance, which means that the GS3 and GS4 feature values are clustered together. This result fits well with the notion by Egevad et al. [2011] that GS3 and GS4 can attain borderline morphology. This means that the morphological structures of GS3 can mimic that of GS4 and GS4 can mimic that of GS3, which especially makes it difficult for a pathologist to annotate but also for a model to classify. This is reflected in this study, where $CNN_{dualWGL}$, achieved a recall score of 0.55 and 0.79 for GS3 and GS4 respectively, and in Ryu et al. [2019] with a result of 0.59 and 0.33 on GS3 and GS4 respectively. In terms of clinical application, a poor classification performance means that the tissue could be classified as a GS3 instead of a GS4. Using the current grading system, it could indicate a summarized GS of 8 instead of 6, where GS6 is categorized as grade group 1 (least aggressive) and a GS8 is categorized as grade group 4 (next to most aggressive). The subsequent course of treatment probably differs where GS6 could be active surveillance and GS8 could indicate some extension of curative treatment. However, these results indicate that the clustered GS3 and GS4 feature values could be a true depiction of the morphological distribution of GS3 and GS4 and that these are very similar, implying that GS3 and GS4 could be divided into subcategories and redefined for more accurate grading. In fact, when a pathologist determines a GS, several structural components of the tissue are considered. This could for example be the presence of necrosis, glumerularion, ciribriform, etc. and these patterns constitute a GS. Using such information in a model would be novel and interesting, possibly finding new tissue combinations that could constitute a new sub-division of GS and aid in greater differentiation between GS. Nevertheless, from the k-means prediction using the remaining test images and the PPS algorithm, the median distance using 25 clusters reveals a slight increase with GS (GS3=0.8, GS4=0.85, GS5=1.0), which implies that there is a minor difference between the feature values of GS3 and GS4.

Reconstruction of H&E patches

The reconstruction performance was poor, showing shades of purple blurry images, almost impossible to differentiate between classes. This could mean that the model hardly learned any natural tissue variation. However, the stroma and cancer classification of 0.97 and 0.91 F1-score respectively, reflected a good performance revealing that the model must have learned tissue-specific features. Nevertheless, there were three times more cancer patches and the F1-scores of cancer could be biased by the class imbalance. Interestingly, a proportionally similar FP leak for benign tissue is observed between $CNN_{dualWGL}$ and $CNN_{dualNGL}$, where 95, 3, and 62 samples are predicted as GS3, GS4, and GS5 for $CNN_{dualWGL}$ respectively, and 142 samples are predicted as cancer for $CNN_{dualNGL}$. indicating that the effect of class imbalance on $CNN_{dualNGL}$ was minor.

Patch sorting algorithm

The patch sorting algorithm only considered 256x256x3 patches, where $> 70\%$ of a single pixel intensity, corresponding to either, benign, GS3, GS4, GS5, or stroma, was present. Consequently, the heatmaps only display areas where $>70\%$ of cancer tissue is present. However, as can be seen in the annotated WSIs, it is only a portion of the true annotated cancer regions that are covered by the heatmaps. Consequently, the clinical application is limited since the heatmap does not capture all cancer within a WSI and the clinician could oversee areas containing cancer, which could lead to incorrect treatment of the patient. However, the properties of a CNN reveal that once a filter has learned a certain pattern, it can find that pattern anywhere on an image. This means that if the same pattern exists in two different patches, where one contains for example 90% GS3 and the other contains 10% GS3, that same pattern is found regardless. Assuming that the same patterns exist between all 256x256x3 cancer patches of a WSI, using all cancer patches could yield a complete heatmap, displaying the concerning area where cancer is located.

Label distribution comparison

From the intuition that the aggressiveness of cancer is continuous, 10, 25, and 50 k-means cluster configurations were tested on the test images prediction using $CNN_{dualNGL}$ with the PPS algorithm, to increase the resolution between transitions of the tissue features and possibly find sub-groups of cancer. Even though the label distribution as a function of accumulated distance increasingly fluctuates with the number of clusters showing greater resolution, a similar tendency from 0 to an accumulated distance of 1.2, 1.4, and 1.6 for the 10, 25, and 50 clusters are observed respectively. It shows that adding clusters does not add substantial information about the transition between tissue features. In addition, the PPS algorithm encounters a limitation as the distance approaches > 1.6 for 50 clusters, where approximately 70% benign, 10% GS3, and 100% GS4 tissue features are present. Analyzing the tissue features on the manifold of $CNN_{dualNGL}$ reveals that benign, GS3,and GS4 tissue features are not present between 0.8 and 1 of the normalized feature 1 axis, and thus the limitation is due to the accumulated euclidean distance calculation of the PPS algorithm. Specifically, when using 50 clusters, a single GS4 feature value is assigned a cluster centroid. That centroid is located farther away from the other centroids in terms of euclidean distance. Since the PPS algorithm must assign a distance to all centroids and finds the shortest euclidean distance from the concerning centroid to the

next that particular centroid containing a GS4 feature value is circumvented and only considered when the euclidean distance eventually becomes the shortest to that centroid. The resulting label distribution is 100% as the cluster contains one sample of GS4 tissue feature, and the accumulated distance is relatively greater as the PPS algorithm reverts to the cluster centroid that was circumvented. Consequently, the label distribution as a function of accumulated distance using 50 clusters must be disregarded and is a limitation of the PPS algorithm.

Future work

In future improvements of the reconstruction performance of CNN_{dual}, it would be interesting to implement a Generative Adversarial Network (GAN). A GAN uses a classification network called a discriminator to penalize the reconstructions of an encoder-decoder architecture, similar to the one in this study. The traditional GAN uses Gaussian noise as input [Goodfellow et al., 2020], however, the GAN can be conditioned on a specific input image, as showed by the exemplary pix2pix architecture in Isola et al. [2017]. In addition, when the structure of the objects within the patches carry important information, it would be evident to investigate a different loss function than the MSE, as it ignores the relation between pixels and disregards the structural information. The Structural Similarity loss function (SSL) measures the similarity between the reconstruction and the ground truth and has shown performance increase over GANs training difficulties and state-of-the-art semantic segmentations. [Zhao et al., 2019] When combining the SSL on the decoder output, with the discriminator penalization, the resulting reconstructions should appear greater in detail, and thus attain better morphological structure than that of this study, possibly aiding in better separation of grade tissue features. Optimal grade tissue feature separation could be useful in a clinical setting where the clinician is challenged in grading the tissue. Specifically, when the GS is determined, the two most prominent grade appearances are assessed, in which the PPS algorithm could indicate whether the concerning WSI mostly contains for example GS3 and secondly GS4 by analyzing the label distribution as a function of distance. A more accurate grade tissue feature separation would also be beneficial within the heatmap since it could guide the clinician in the specific placement of where the GS3 and GS4 are present. It could be highlighted by assigning specific colors to the concerning grade, which could mitigate the interobserver variability currently existing in PCa tissue assessment and potentially provide more accurate treatment for the patient.

In a study by Inglese et al. [2017], mass spectrometry was used on colorectal adenocarcinoma to utilize the mass/charge ratio of molecules in consecutive tissue slices to represent a spatial distribution of chemical and biological structures. The intuition was that visual inspection of tumor tissues does not reveal the complex metabolic alterations, that exist in a three-dimensional tumor environment, which contribute to the differentiation of cancer and its sub-types from healthy tissues. [Inglese et al., 2017] Specifically, the additional information in mass spectrometry could be utilized in an attempt to find sub-groups of GS in prostate cancer, to enable a greater distinction between e.g GS3 and GS4, which could aid in more accurate grading of PCa.

Conclusion 7

The aim of the study was to develop a feature space containing natural tissue variation from benign to cancer tissue to understand if a GS can be determined without using grade labels to mitigate interobserver variability. $CNN_{dualNGL}$ and $CNN_{dualWGL}$ were developed containing a reconstruction output and a multi-class classification output to learn morphological structures of the patches and to separate the learned tissue features into the corresponding classes. The models were trained and validated on 256x256 patches extracted from H&E stained WSIs, whereas the patches contained at least 70% of either benign, GS3, GS4, GS5, or stroma. $CNN_{dualNGL}$ was trained without grade labels and $CNN_{dualWGL}$ was trained with grade labels to analyze if the additional grade information is needed to obtain sufficient feature separation in FS for determining a GS. From the features of the unseen test images explaining $> 90\%$ variance constituting two features, it was found that $CNN_{dualNGL}$ learned a greater separation of benign, GS3, GS4, and GS5 tissue features than $CNN_{dualWGL}$. The mean feature value difference from benign and GS3, GS4 and GS5 using the test images of $CNN_{dualWGL}$ and $CNN_{dualNGL}$ were (0.005, 0.059), (-0.041, 0.147), (-0.272, 0.187) and (-0.202, 0.153), (-0.290, 0.153), (-0.438, 0.170) respectively. The dimension reduced FS of $CNN_{dualWGL}$ and $CNN_{dualNGL}$ using the test images was used to create a distance metric for assessing GS, whereas it was assumed that greater accumulated euclidean distance from benign tissue would contain more aggressive GS. As such the PPS algorithm was developed to calculate the accumulated distance between K-means cluster centroids from benign to GS3, GS4, and GS5 along with the label distribution contained within every cluster. Using 10 and 25 k-means clusters, the majority of benign tissue features were present between an accumulated euclidean distance of 0 to 0.6. The GS3 and GS4 tissue features follow a similar tendency, and the majority were present at an accumulated distance between 0.6 to 0.9. Finally, the majority of GS5 tissue features were present from an accumulated distance of 0.9 to 1.2 using 10 k-means clusters and 0.9 to 1.4 using 25 k-means clusters. These results showed that some natural tissue separation was obtained without using grade labels and that the majority of a GS can be implied by the PPS algorithm. However deep learning architectural improvements are needed to understand the full utility of the proposed solutions in this study.

Bibliography

Bauer et al., **2021**. Markus Bauer, Sebastian Zürner, Markus Kreuz, Dominik Otto, Georg Popp and Ulf-Dietrich Braumann. *Histological grading of the prostate carcinoma using deep learning: an unsupervised approach.* (February 2021), 31, 2021. ISSN 16057422. doi: 10.1117/12.2581043.

Bharadwaj et al., **2021**. Bharadwaj, Kolla Bhanu Prakash and G. R. Kanagachidambaresan. *Pattern Recognition and Machine Learning.* 2021. ISBN 9780387310732. doi: 10.1007/978-3-030-57077-4_11.

Borre et al., **2019**. Michael Borre, Bente Klarlund Pedersen, Naja Zenius Jespersen, Dorte Bojer, Lisa Sengeløv, Gregers G. Hermann and Gregers Hansen-Nord. *Prostatakræft*, 2019. URL `https://www.sundhed.dk/sundhedsfaglig/laegehaandbogen/mandlige-koensorganer/tilstande-og-sygdomme/prostata/prostatakraeft/`.

Bulten and Litjens, **2018**. Wouter Bulten and Geert Litjens. *Unsupervised Prostate Cancer Detection on H&E using Convolutional Adversarial Autoencoders.* 2018. ISSN 2331-8422. URL `http://arxiv.org/abs/1804.07098`.

Bushaev, **2022**. Vitaly Bushaev. *Understanding RMSprop — faster neural network learning*, 2022. URL `https://towardsdatascience.com/understanding-rmsprop-faster-neural-network-learning-62e116fcf29a`.

Campanella et al., **2019**. Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra and Thomas J. Fuchs. *Clinical-grade computational pathology using weakly supervised deep learning on whole slide images.* Nature Medicine, 25(8), 1301–1309, 2019. ISSN 1546170X. doi: 10.1038/s41591-019-0508-1.

Campanella et al., **2020**. Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra and Thomas J Fuchs. *Deep Learning on Whole Slide Images.* 25(8), 1301–1309, 2020. doi: 10.1038/s41591-019-0508-1.Clinical-grade.

Cho et al., **2012**. Yong Mee Cho, Jae Y. Ro and Gustavo E. Ayala. *Molecular pathology of prostate cancer.* 2012. ISBN 9781451178098. doi: 10.1016/b978-0-12-800886-7.00022-4.

Chollet et al., **2018**. Francois Chollet et al. *Deep learning with Python*, volume 361. Manning New York, 2018.

Egevad et al., **2011**. Lars Egevad, Ferran Algaba, Daniel M. Berney, Liliane Boccon-Gibod, Eva Compérat, Andrew J. Evans, Rainer Grobholz, Glen Kristiansen, Cord Langner, Gina Lockwood, Antonio Lopez-Beltran, Rodolfo Montironi, Pedro

Oliveira, Matthias Schwenkglenks, Ben Vainer, Murali Varma, Vincent Verger and Philippe Camparo. *Interactive digital slides with heat maps: A novel method to improve the reproducibility of Gleason grading.* Virchows Archiv, 459(2), 175–182, 2011. ISSN 09456317. doi: 10.1007/s00428-011-1106-x.

Fei-Fei Li, **2020**. Danfei Xu Fei-Fei Li, Ranjay Krishna. *Convolutional Neural Networks (CNNs / ConvNets)*, 2020. URL `https://cs231n.github.io/convolutional-networks/`.

Goodfellow et al., **2016**. Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning.* MIT Press, 2016. `http://www.deeplearningbook.org`.

Goodfellow et al., **2020**. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio. *Generative adversarial networks.* Communications of the ACM, 63(11), 139–144, 2020. ISSN 15577317. doi: 10.1145/3422622.

Gospodarowicz et al., **2017**. James D. Brierley Gospodarowicz, Mary K. Wittekind and Christian Brierley. *Molecular pathology of prostate cancer.* John Wiley & sons, incorporated, 2017. ISBN 9781119263562.

Gummeson et al., **2017**. Anna Gummeson, Ida Arvidsson, Mattias Ohlsson, Niels C. Overgaard, Agnieszka Krzyzanowska, Anders Heyden, Anders Bjartell and Kalle Aström. *Automatic Gleason grading of H and E stained microscopic prostate images using deep convolutional neural networks.* Medical Imaging 2017: Digital Pathology, 10140(March 2017), 101400S, 2017. ISSN 16057422. doi: 10.1117/12.2253620.

He et al., **2012**. Lei He, L. Rodney Long, Sameer Antani and George R. Thoma. *Histology image analysis for carcinoma detection and grading.* Computer Methods and Programs in Biomedicine, 107(3), 538–556, 2012. ISSN 01692607. doi: 10.1016/j.cmpb.2011.12.007. URL `http://dx.doi.org/10.1016/j.cmpb.2011.12.007`.

Jette Egelund Holgaard, Thomas Ryberg, Nikolaj Stegeager, Diana Stentoft and Anja Overgaard Thomassen. Problembaseret læring og projektarbejde ved de videregående uddannelser, chapter 3, pages 60–67. Samfundslitteratur, 2014. ISBN 978-87-593-2149-2. 1. edition.

InformedHealth.org, **2019**. InformedHealth.org. *How do cancer cells grow and spread?*, 2019. URL `https://www.ncbi.nlm.nih.gov/books/NBK279410/`.

Inglese et al., **2017**. Paolo Inglese, James S. McKenzie, Anna Mroz, James Kinross, Kirill Veselkov, Elaine Holmes, Zoltan Takats, Jeremy K. Nicholson and Robert C. Glen. *Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer.* Chemical Science, 8(5), 3500–3511, 2017. ISSN 20416539. doi: 10.1039/c6sc03738k. URL `http://dx.doi.org/10.1039/C6SC03738K`.

Isola et al., **2017**. Phillip Isola, Jun Yan Zhu, Tinghui Zhou and Alexei A. Efros. *Image-to-image translation with conditional adversarial networks.* Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pages 5967–5976, 2017. doi: 10.1109/CVPR.2017.632.

Kingma and Ba, **2015**. Diederik P. Kingma and Jimmy Lei Ba. *Adam: A method for stochastic optimization.* 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, pages 1–15, 2015.

Koech, **2022**. Kiprono Elijah Koech. *Cross-Entropy Loss Function*, 2022. URL `https://towardsdatascience.com/cross-entropy-loss-function-f38c4ec8643e`.

Korstanje, **2021**. Joos Korstanje. *The F1 score*, 2021. URL `https://towardsdatascience.com/the-f1-score-bec2bbc38aa6`.

Linkon et al., **2021**. Ali Hasan Md Linkon, Md Mahir Labib, Tarik Hasan, Mozammal Hossain and Marium E. Jannat. *Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: An extensive study.* Informatics in Medicine Unlocked, 24(April), 100582, 2021. ISSN 23529148. doi: 10.1016/j.imu.2021.100582. URL `https://doi.org/10.1016/j.imu.2021.100582`.

Lu and Daigle, **2020**. Liangqun Lu and Bernie J. Daigle. *Prognostic analysis of histopathological images using pre-trained convolutional neural networks: Application to hepatocellular carcinoma.* PeerJ, 2020(3), 2020. ISSN 21678359. doi: 10.7717/peerj.8668.

Martini et al., **2003**. Frederic H. Martini, Judi L. Nath and Edwin F. Bartholomew. *Fundamentals of Anatomy & Physiology.* 2003. ISBN 9780321709332.

Michael Borre, **2019**. Henriette Lindberg Michael Borre. *Undersøgelser ved prostatakræft*, 2019. URL `https://www.cancer.dk/prostatakraeft/undersogelser-prostatakraeft/`.

Miller, **2016**. Mary E. Miller. *Cancer.* Momentum Press, 2016. 1. edition.

Mohajon, **2020**. Joydwip Mohajon. *Confusion Matrix for Your Multi-Class Machine Learning Model*, 2020. URL `https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826`.

Nir et al., **2019**. Guy Nir, Davood Karimi, S. Larry Goldenberg, Ladan Fazli, Brian F. Skinnider, Peyman Tavassoli, Dmitry Turbin, Carlos F. Villamil, Gang Wang, Darby J.S. Thompson, Peter C. Black and Septimiu E. Salcudean. *Comparison of Artificial Intelligence Techniques to Evaluate Performance of a Classifier for Automatic Grading of Prostate Cancer from Digitized Histopathologic Images.* JAMA Network Open, 2(3), 2–11, 2019. ISSN 25743805. doi: 10.1001/jamanetworkopen.2019.0442.

Nwankpa et al., **2018**. Chigozie Nwankpa, Winifred Ijomah, Anthony Gachagan and Stephen Marshall. *Activation Functions: Comparison of trends in Practice and Research for Deep Learning.* pages 1–20, 2018. ISSN 2331-8422. URL `http://arxiv.org/abs/1811.03378`.

Oxley, **2014**. Jon Oxley. *Understanding the Histopathology.* Prostate Cancer: Diagnosis and Clinical Management, pages 34–48, 2014. doi: 10.1002/9781118347379.ch3.

Ozkan et al., **2016**. Tayyar A. Ozkan, Ahmet T. Eruyar, Oguz O. Cebeci, Omur Memik, Levent Ozcan and Ibrahim Kuskonmaz. *Interobserver variability in Gleason*

histological grading of prostate cancer. Scandinavian Journal of Urology, 50(6), 420–424, 2016. ISSN 21681813. doi: 10.1080/21681805.2016.1206619.

Radboudumc and Institutet, **2020**. Radboudumc and Karolinska Institutet. *Prostate cancer grade assessment (PANDA) challenge*, 2020. URL `https://www.kaggle.com/c/prostate-cancer-grade-assessment`.

Rajal B. Shah, **2012**. Ming Zhou Rajal B. Shah. *Prostate Biopsy Interpretation: An Illustration Guide*. Springer, 2012. ISBN 9783642213687.

Reinholdt, **2019**. Jacon Reinholdt. *Dropout on convolutional layers is weird*, 2019. URL `https://towardsdatascience.com/dropout-on-convolutional-layers-is-weird-5c6ab14f19b2`.

Ryu et al., **2019**. Han Suk Ryu, Min Sun Jin, Jeong Hwan Park, Sanghun Lee, Joonyoung Cho, Sangjun Oh, Tae Yeong Kwak, Junwoo Isaacwoo, Yechan Mun, Sun Woo Kim, Soohyun Hwang, Su Jin Shin and Hyeyoon Chang. *Automated gleason scoring and tumor quantification in prostate core needle biopsy images using deep neural networks and its comparison with pathologist-based assessment*. Cancers, 11(12), 2019. ISSN 20726694. doi: 10.3390/cancers11121860.

Semmlow and Griffel, **2004**. John L. Semmlow and Benjamin Griffel. *Multivariate Analyses: Principle Component Analysis and Independent Component Analysis*. Marcel Dekker, 2004.

Srivastava et al., **2014**. Nitish Srivastava, Geoffry Hinton, Ilaya Sutskever Alex Krishevsky and Ruslan Salakhutdinov. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. Journal of Machine Learning Reasearch, 15, 1929–1958, 2014. doi: 10.1109/ICAEES.2016.7888100.

Ström et al., **2020**. Peter Ström, Kimmo Kartasalo, Henrik Olsson, Leslie Solorzano, Brett Delahunt, Daniel M. Berney, David G. Bostwick, Andrew J. Evans, David J. Grignon, Peter A. Humphrey, Kenneth A. Iczkowski, James G. Kench, Glen Kristiansen, Theodorus H. van der Kwast, Katia R.M. Leite, Jesse K. McKenney, Jon Oxley, Chin Chen Pan, Hemamali Samaratunga, John R. Srigley, Hiroyuki Takahashi, Toyonori Tsuzuki, Murali Varma, Ming Zhou, Johan Lindberg, Cecilia Lindskog, Pekka Ruusuvuori, Carolina Wählby, Henrik Grönberg, Mattias Rantalainen, Lars Egevad and Martin Eklund. *Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study*. The Lancet Oncology, 21(2), 222–232, 2020. ISSN 14745488. doi: 10.1016/S1470-2045(19)30738-7.

Sung et al., **2021**. Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal and Freddie Bray. *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries*. CA: A Cancer Journal for Clinicians, 71(3), 209–249, 2021. ISSN 0007-9235. doi: 10.3322/caac.21660.

Suvarna et al., **2018**. Kim S Suvarna, Christopher Layton and John D Bancroft. *Bancroft's theory and practice of histological techniques E-Book*. Elsevier Health Sciences, 2018.

Tharwat, **2018**. Alaa Tharwat. *Classification assessment methods.* Applied Computing and Informatics, 17(1), 168–192, 2018. ISSN 22108327. doi: 10.1016/j.aci.2018.08.003.

Thomsen et al., **2015**. Frederik B. Thomsen, Niels Marcussen, Kasper D. Berg, Ib J. Christensen, Ben Vainer, Peter Iversen and Klaus Brasso. *Repeated biopsies in patients with prostate cancer on active surveillance: Clinical implications of interobserver variation in histopathological assessment.* BJU International, 115(4), 599–605, 2015. ISSN 1464410X. doi: 10.1111/bju.12820.

Tran et al., **2021**. Khoa A. Tran, Olga Kondrashova, Andrew Bradley, Elizabeth D. Williams, John V. Pearson and Nicola Waddell. *Deep learning in cancer diagnosis, prognosis and treatment selection.* Genome Medicine, 13(1), 1–17, 2021. ISSN 1756994X. doi: 10.1186/s13073-021-00968-x.

Tătaru et al., **2021**. Octavian Sabin Tătaru, Mihai Dorin Vartolomei, Jens J. Rassweiler, Oșan Virgil, Giuseppe Lucarelli, Francesco Porpiglia, Daniele Amparore, Matteo Manfredi, Giuseppe Carrieri, Ugo Falagario, Daniela Terracciano, Ottavio de Cobelli, Gian Maria Busetto, Francesco Del Giudice and Matteo Ferro. *Artificial intelligence and machine learning in prostate cancer patient management—current trends and future perspectives.* Diagnostics, 11(2), 1–20, 2021. ISSN 20754418. doi: 10.3390/diagnostics11020354.

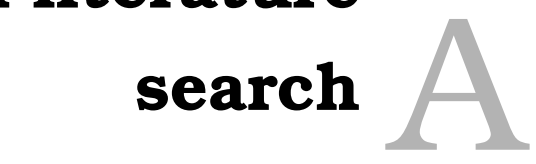
van Santvoort et al., **2020**. B. W.H. van Santvoort, G. J.L.H. van Leenders, L. A. Kiemeney, I. M. van Oort, S. E. Wieringa, H. Jansen, R. W.M. Vernooij, C. A. Hulsbergen-van de Kaa and K. K.H. Aben. *Histopathological re-evaluations of biopsies in prostate cancer: a nationwide observational study.* Scandinavian Journal of Urology, 54(6), 463–469, 2020. ISSN 21681813. doi: 10.1080/21681805.2020.1806354.

Wang and Bovik, **2009**. Zhou Wang and Alan C Bovik. *Mean Squared Error : Love It or Leave It ?* IEEE Signal Processing Magazine, 26(1), 98–117, 2009. ISSN 1053-5888. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4775883`.

WHO, **2022**. WHO. *Cancer*, 2022. URL `https://www.who.int/news-room/fact-sheets/detail/cancer`.

Zhao et al., **2019**. Shuai Zhao, Boxi Wu, Wenqing Chu, Yao Hu and Deng Cai. *Correlation maximized structural similarity loss for semantic segmentation.* arXiv, 2019. ISSN 23318422.

Structured literature search



A structured literature search was conducted to research the field of prostate cancer and the existing interobserver variability along with current deep learning methods applied for mitigating the interobserver variability. Initially a semi-structured literature search was conducted to uncover keywords relevant for a structured literature search. The keywords were organized in the four blocks as shown in figure A.1.

Cancer	Histopathology	Deep Learning	Feature space analysis
- Neoplasms - Neoplasia - Neoplasias - Tumor - Tumors - Malignancy - Malignancies	- Whole slide - Whole slides - WSI - Tissue Slide - Tissue Slides - Histology* - Histologic - Histological - Histopathology - Digital pathology - Pathology - Staining and Labeling* - Hematoxylin - Eosin - H&E	- deep learning - neural networks - convolutional neural networks - CNN - computer aided diagnosis - Autoencoder - Variational autoencoder - Unet - GAN - Classification - Decision Support	- Feature space analysis - Latent space - Embedding space - t-SNE - UMAP - neoplasm grading [Mesh]

Figure A.1

The Scopus database was used for conducting the structured literate search, covering engineering topics within biomedical science along with general health sciences. Using the keywords from figure A.1, a search string was made resulting in 77 articles as can be seen in figure A.2.

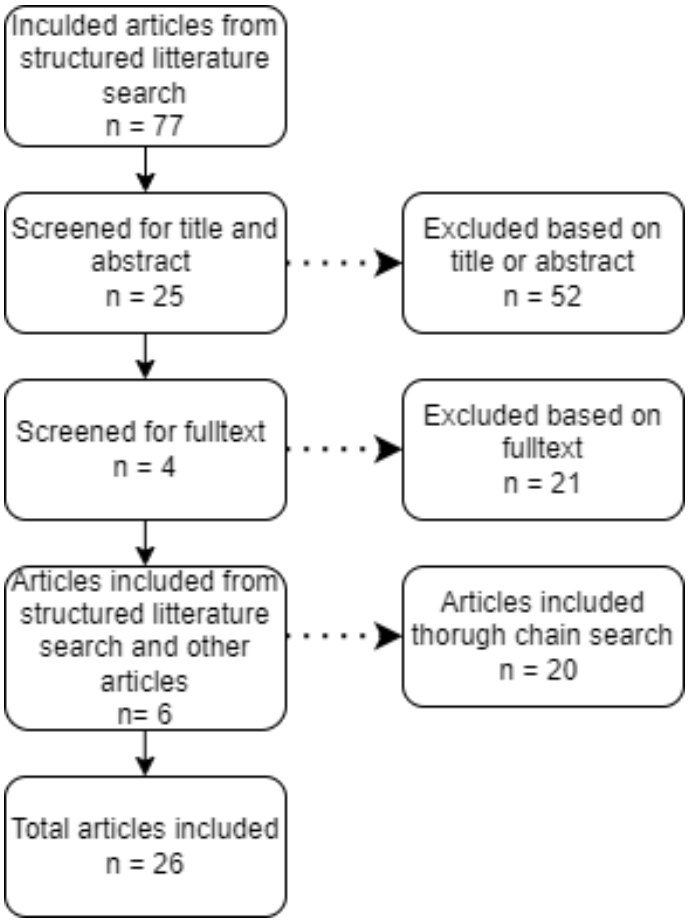


Figure A.2

For sorting the articles gained from the search string, the inclusion and exclusion criteria stated below were used. The articles were sorted title and abstract and fulltext including 6 articles. A chain search was conducted on the included articles, whereas 20 was included. As such the total amount of articles used from the structures literature search was 26.

Inclusion Criteria

- Language: English or Danish
- Year of publication: 2012 - 2022
- Data: histological data, primarily tumor tissue
- Must use machine learning for cancer assessment (grading or change in morphological structures of the tissue i.e grades of cancer)

Exclusion Criteria

- Articles using supervised machine learning for predicting grading without using methods for feature space analysis.
- Articles using supervised machine learning for predicting cancer without using feature space for tissue assessment between classifications.
- Articles having unclear method descriptions
- Articles using supervised or unsupervised methods for predicting cancer in animals or children

- Articles not about tissue differentiation in cancer
- Full text not available

Theory of the components of CNN_{dual}

B

Convolution

The convolution constitutes a dot product between a specified number of filters, with an input image or feature map. The filters are used for recognizing patters within images, and if a certain pattern is learned it can be located anywhere in the image. The filters are composed of weights, and each filter contains different weights representing a certain pattern. When performing a convolution, the dot product between the weights and the pixel values of the H&E patch or feature map values are calculated and summed with an added bias value, shown in B.1. The degree of response reveals if the filters pattern matches the concerning area, higher sum being more similar patterns. [Chollet et al., 2018; Fei-Fei Li, 2020] For this reason, the convolution is used on the H&E patches, as opposed to a fully connected network, as that would have to learn the pattern anew, if it appeared anywhere else.

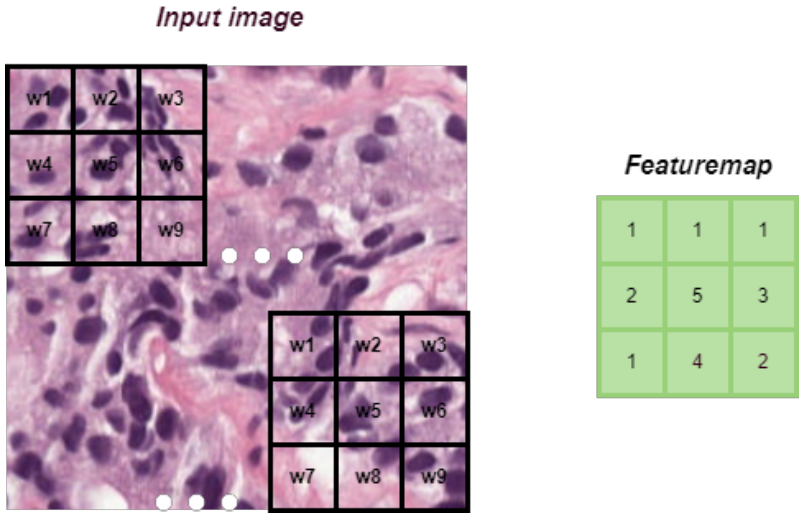


Figure B.1. Depiction of a convolution with a H&E input image with a 3x3 filter, consisting of weights denoted as w1-9, and the corresponding output featuremap from that convolution.

The transposed convolution has similar properties as the convolution, however, as it can be viewed in figure B.2, the input changes size after a transposed convolution. Therefore, the input is up-sampled inserting zeros in between the featuremap values, for the principle convolution to yield a bigger output. The transposed convolution was used for generating output images, as close to the input as possible, to learn CNN_{dual} the relevant features within the images.

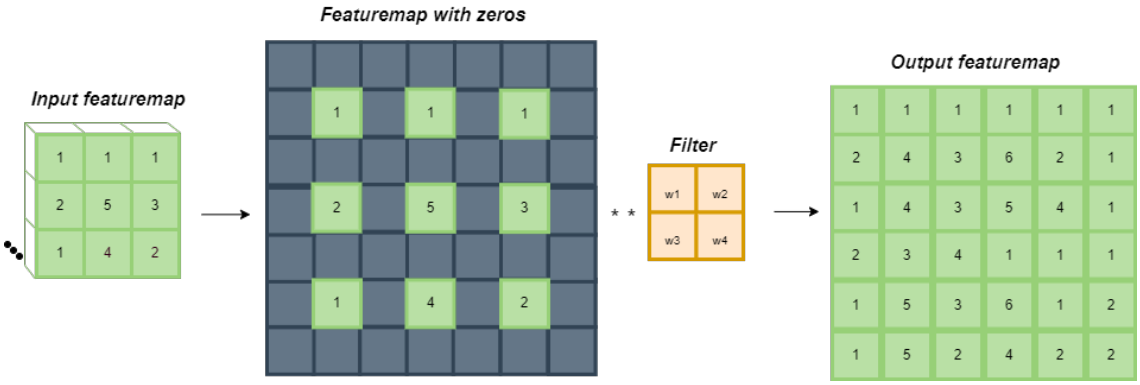


Figure B.2. Depiction of a transposed convolution, where zeros are inserted and a 2x2 filter is applied to yield the output featuremap

The generated feature maps from the convolution and transposed convolution, was input to the Rectified Linear Unit (ReLU) activation function.

ReLU

The ReLU activation function is a non-linear function used widely in deep neural networks, and transforms the input featuremap values, as depicted in equation B.1 and figure B.3 [Nwankpa et al., 2018].

$$f(x_i) = \begin{cases} x_i & if \;\; x_i \geq 0 \\ 0 & if \;\; x_i < 0 \end{cases} \tag{B.1}$$

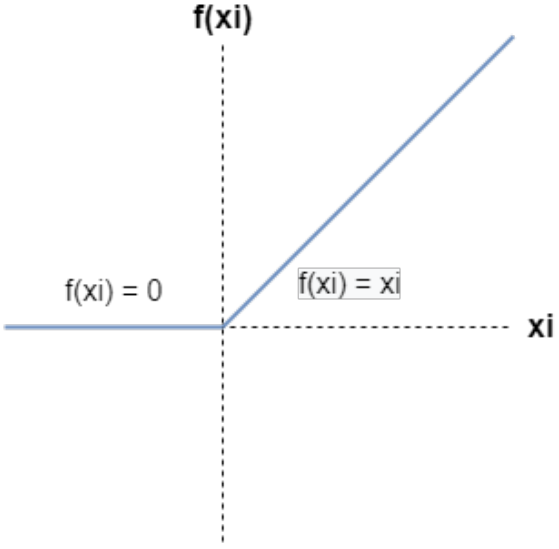
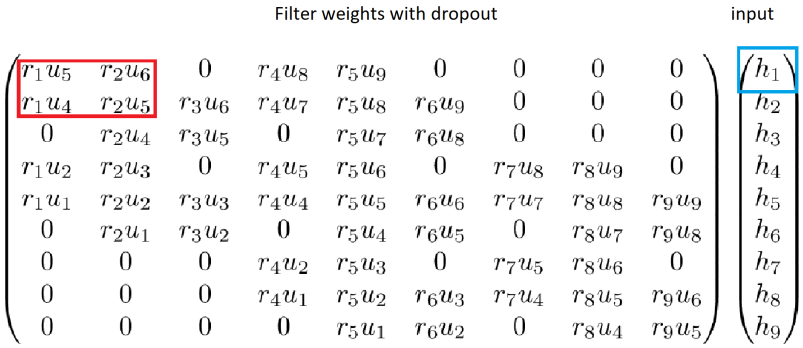


Figure B.3. ReLU activation function

The ReLU function thresholds the input below the bias value, thus input values below zero are set to zero. The bias value effectively shifts the activation function to the left or right, which can be important for a successful model. [Chollet et al., 2018]

Dropout

Dropout is a function that regularizes a CNN, by forcing non-trainable weights during backpropagation as a means of mitigating overfitting. Dropout in CNNs works by randomly multiplying the filter weights with 0, effectively preventing a convolution with that weight and the concerning area in the input. [Srivastava et al., 2014] However, this does not mean that the weight is not trainable, which is exemplified by Reinholdt [2019]. In figure B.4, the denotion r represents the dropout value, multiplied to the weights u, which in turn is multiplied to the input h. Given that the r1 value (red square) is 0, only the first column is disregarded, however the same weights are present at other entries in the matrix, which makes the weights trainable, but disregards the first value of the input. The effect is adding noise, which also prevents overfitting.[Reinholdt, 2019]



$$\begin{pmatrix} r_1u_5 & r_2u_6 & 0 & r_4u_8 & r_5u_9 & 0 & 0 & 0 & 0 \\ r_1u_4 & r_2u_5 & r_3u_6 & r_4u_7 & r_5u_8 & r_6u_9 & 0 & 0 & 0 \\ 0 & r_2u_4 & r_3u_5 & 0 & r_5u_7 & r_6u_8 & 0 & 0 & 0 \\ r_1u_2 & r_2u_3 & 0 & r_4u_5 & r_5u_6 & 0 & r_7u_8 & r_8u_9 & 0 \\ r_1u_1 & r_2u_2 & r_3u_3 & r_4u_4 & r_5u_5 & r_6u_6 & r_7u_7 & r_8u_8 & r_9u_9 \\ 0 & r_2u_1 & r_3u_2 & 0 & r_5u_4 & r_6u_5 & 0 & r_8u_7 & r_9u_8 \\ 0 & 0 & 0 & r_4u_2 & r_5u_3 & 0 & r_7u_5 & r_8u_6 & 0 \\ 0 & 0 & 0 & r_4u_1 & r_5u_2 & r_6u_3 & r_7u_4 & r_8u_5 & r_9u_6 \\ 0 & 0 & 0 & 0 & r_5u_1 & r_6u_2 & 0 & r_8u_4 & r_9u_5 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ h_8 \\ h_9 \end{pmatrix}$$

Figure B.4. Convolution operation of filter weights (u) with dropout (r) and an input (h)

Max Pooling

The max pooling operation takes in a feature map and outputs the maximum value of a specified window of that feature map, as shown in figure B.5. This is beneficial as it retains the information from the best match of the convolution in addition to reducing the amount of feature map values. Effectively, this gives a greater field of view for subsequent filters, which enables a combination of relevant high- and low-level features, such as a prostate gland and cell shapes, for the classification of stroma, benign or cancer. [Chollet et al., 2018]

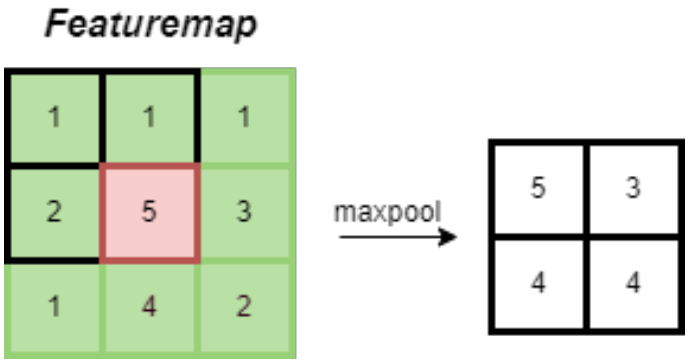


Figure B.5. maxpool operation, where a 3x3 feature map is reduced to a 2x2 feature map retaining the largest values from the maxpooling operation

All maxpooling layers of the CNN_{dual} were defined with a stride of 2, where the output was the maximum value of and area on the input image of size 4x4. The resulting dimensions were half the input featuremaps width and height.

Fully Connected layer

A fully connected layer works different than a convoultional layer. The term fully connected means that each input to a neuron is the sum of all weights and bias' from all preceeding neurons, and thus every neuron is connected. [Chollet et al., 2018] this is depicted in figure B.6

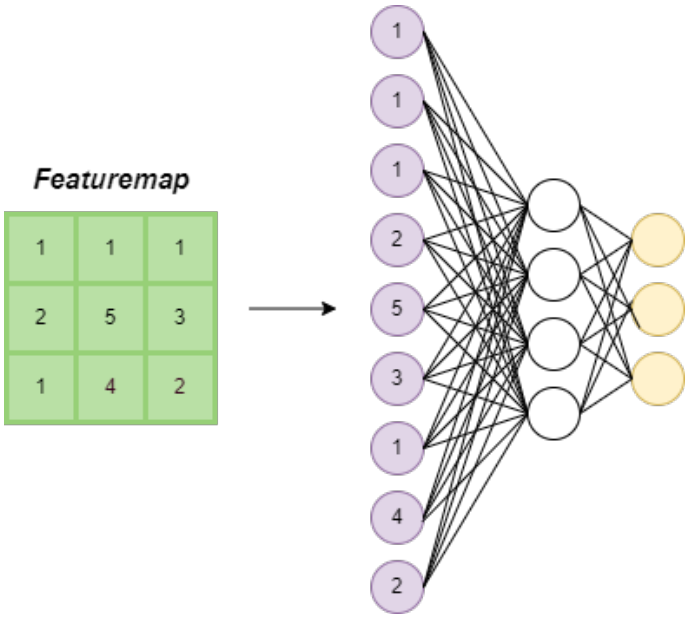


Figure B.6. Fully connected layer example, where the feature map values (purple) are flattened and all neurons are connected to the proceeding layer, concluding the architecture with three neurons (yellow)

The output *Y(x)* of a fully connected layer is given by the product between an input vector x and a weight matrix W with an added bias b, which is all wrapped in an activation function [Chollet et al., 2018], which for this study was ReLU, denoted as g in equation B.2

$$Y(x) = g(xW + b) \tag{B.2}$$

For this study the fully connected layers are concluded with a softmax activation function, for classifying stroma, benign or cancer patches.

Softmax

The softmax activation function is used in a single label multiclass classification problem, to compute a probability between 0 and 1, where the sum of all outputs equals to 1. [Nwankpa et al., 2018] The last layer of the fully connected output of CNN_{dual} contained three neurons, were each neuron represented a certain class. As such, applying the softmax to the output neurons yielded the probability that the concerning input patch belonged to one of three classes.

The softmax activation function is given by:

$$s(x_i) = \frac{exp(x_i)}{\sum_j exp(x_j)} \tag{B.3}$$

For i=1..n values within every neuron in the last layer, the exponential value is calculated and divided by the sum of j=1..n exponential values, resulting in a probability prediction for every output neuron, corresponding to stroma, benign or cancer.

B.0.1 Optimization and Backpropagation

For the network to perform better at mapping a H&E patch input to the expected reconstruction targets and classifications, adjustments to the weights and bias' was done based on the output of both loss functions in equation 4.2 and 4.1. The adjustments was made by computing the gradient of the loss, with respect to the weights and bias', finding the direction the individual parameter should be adjusted in.

Denoting the weights and bias' of the network as, W, the total loss can be formulated as in equation B.4:

$$loss = f(W) \tag{B.4}$$

The derivative of $f(W)$ describes the gradient and thus the curvature of the parameters in W, which is useful when finding a minimum of a function. Since W is a collective term for all weights and bias' in a network, it is evident that increasing the performance of a network means adjusting every weight and bias by their individual gradient, propagating backwards from the last layer to the first layer, deriving the magnitude of contribution of each parameter to the final loss, such that the loss decreases, which can be described as $\frac{\partial L}{\partial W}$. As the contributing operations are chained together, backpropagation uses the chain rule to compute the derivative of the individual parts. Mathematically, this means that

the derivative of a composite function is constituted by the derivative of the individual functions. [Chollet et al., 2018]

After the gradient is calculated the weights and bias' are updated proportionally in the direction of the gradient by using mini batch stochastic gradient descent (mini-batch SGD), shown in figure B.7.

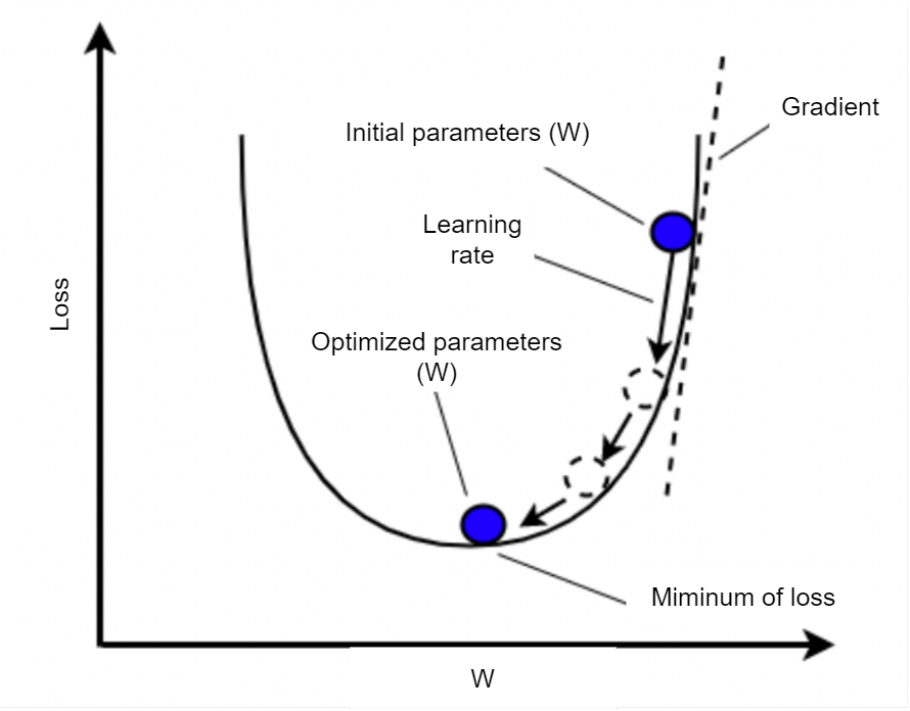


Figure B.7. Conceptual optimization of one parameter. The blue circle are the weights and bias' being moved in proportion by the learning rate and in the direction of the gradient.

Mini-batch SGD uses random samples from the training data, to compute a average gradient from those samples. Then taking the negative of the average gradient from SGD, results in $f(W)$ being moved yielding a lower loss, or to a lower point. [Chollet et al., 2018] Finally, a learning rate is applied that describes the magnitude of the adjustments to the weights and bias', which is formulated in equation B.5

$$W_n = W_{n-1} - LR * \frac{\partial L_n}{\partial W n} \tag{B.5}$$

Optimizer

The principle of the momentum is described in equation B.7, where V_t is the momentum at time t, which is controlled by $0 \leq \gamma \leq 1$, decreasing exponentially with time. The γ value in this project was 0.9 [Kingma and Ba, 2015]

$$V_{t-1} = \left(\frac{\partial L_t}{\partial W_t}\right) + \gamma * \left(\frac{\partial L_{t-1}}{\partial W_{t-1}}\right) + \gamma^2 * \left(\frac{\partial L_{t-2}}{\partial W_{t-2}}\right) .. + \gamma^{n-1} * \left(\frac{\partial L_{t-n}}{\partial W_{t-n}}\right) \tag{B.6}$$

substituting V_{t-1} to equation B.5 yields the following equation

$$W_t = W_{t-1} - \gamma V_{t-1} + LR * \frac{\partial L_t}{\partial Wt} \tag{B.7}$$

The RMSprop keeps a moving average of the squared loss with respect to the weights and bias' in the network, which is then squared and divided with the gradient. RMSprop introduces a β and ϵ term [Bushaev, 2022], which in this study was 0.99 and 10^{-7} respectively. The RMSprop is calculated as:

$$S_{dWt} = \beta S_{dWt-1} + (1-\beta)\left(\frac{\partial L_t}{\partial W_t}\right)^2 \tag{B.8}$$

The effect of RMSprop is moving fast for gradients moving towards a minimum, and slower for gradients moving away from the minimum. Formulating the final Adam optimization algorithm by implementing the RMSprop component gives the following equation:

$$W_t = W_{t-1} - \gamma V_{t-1} + LR * \frac{\left(\frac{\partial L_t}{\partial Wt}\right)}{\sqrt{S_{dWt} + \epsilon}} \tag{B.9}$$

B.0.2 Principle component analysis

PCA is a multi variate analysis method, for reducing the dimensions of a dataset, transforming the data into a set of uncorrelated principle components. The principle approach of PCA, is explained in the following steps.

- Step 1: Calculate the mean of the dataset
- Step 2: Shift data such that mean value is on top of origin
- Step 3: Find best fitting line by maximizing the sum of squared distances (eigenvalue) from the projected points to the origin
- Step 4: Calculate the variance around the origin for every principle component
- Step 5: Repeat step 1-4 for all dimensions
- Step 6: Reduce dimensionality

For large datasets with multiple dimension the singular value decomposition (SVD) can be used to compute the variance of the principle components:

$$A_{mxn} = U_{mxm} D_{mxn} V^T_{nxn} \tag{B.10}$$

Where A is the matrix containing the original data, that can be decomposed into U, D and V^T. U is an orthonormal matrix, that does a rotation of the data, such that the covariance is reduced to zero, which means that the data is uncorrelated. D is the diagonal matrix, containing the square roots of the eigenvalues and V^T is the transposed

matrix containing the eigenvectors, describing the direction of the principle components. To calculate the variance of every principle component, the eigenvectors in V^T are scaled with the corresponding eigenvalues in D, resulting in a size ordered description of the variance from largest to lowest. These can be sorted, such that the desired amount of variance can be retained, while the rest is disregarded. [Semmlow and Griffel, 2004] Figure B.8 shows Step 1-6, where 2 dimensions is reduced to a one dimensional sub-space.

The steps can bee seen in figure B.8

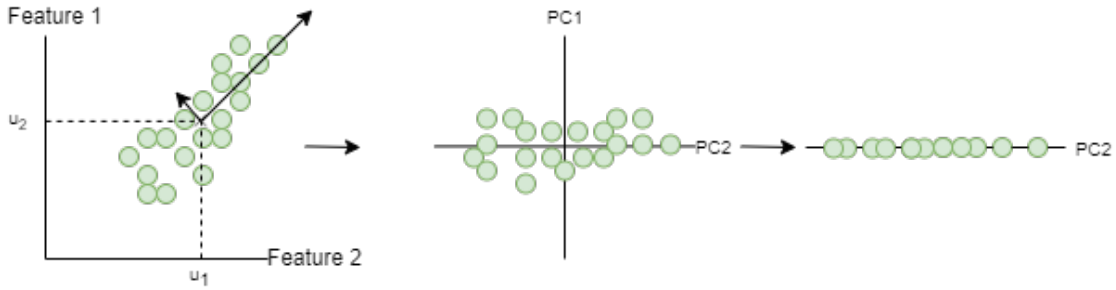


Figure B.8. The PCA showed in three separate images for dimensionality reduction. In the left most image, step 1-2 are done calculating the mean (u1, u2) of the dataset. In the middle image the data is rotated and the eigenvalues of the principle components are calculated, as in step 3-4. Finally, the data is reduced by choosing PC2, as it retained the most variance.

B.0.3 Unsupervised K-means clustering

Unsupervised learning is a tool for finding patterns and transformation in data, which does not require a target and one such tool is called the K-means clustering. K-means clustering is an unsupervised method for grouping similar data points to reveal patterns in data. To achieve this k-means searches for (k) number clusters within the dataset. The k refers to the number of centroids randomly initialized by the algorithm, where every data point is assigned to which ever centroid is closest. The K-means was used as it provided an opportunity for searching for a fixed number of sub-groups in feature space. Since the K-means searches for similar points in feature space, the resulting sub-groups would also be naturally occurring. The principles of the K-means algorithm is elaborated in the following steps:

- Step 1: Initialize k random centroids
- Step 2: Calculate the euclidean distance from every data point to every centroid
- Step 3: Classify samples closest to a centroid as belonging to that cluster.
- Step 4: Calculate mean vector of all clusters respectively and move centroids to that point
- Step 5: Repeat Step 2-4 until convergence

[Bharadwaj et al., 2021]

The algorithm is converged either when the predefined number of iterations is achieved, or when the mean vector stabilizes i.e when the sum of the squares of the distances of the data points within the given cluster to the centroid was at a minimum. [Bharadwaj et al., 2021] Practically, the sckit-learn python package K-means was used to perform the unsupervised clustering.

PCA with stroma C

This appendix visualizes the two most explaining principle components, when stroma patches are used in CNN_{dual} prediction, from where the PCA is conducted on the feature space. In figure C.1, the leftmost and rightmost figure shows the normalized features from a PCA conducted on CNN_{dual} trained with and without grade labels respectively. It is evident that the variance of feature 1 is explained by stroma in both cases, since the corresponding features ranges from 0-1, while GS3, GS4 and benign tissue features are clustered on top of each other. As such, for investigating the aim of this study, the stroma patches were disregarded.

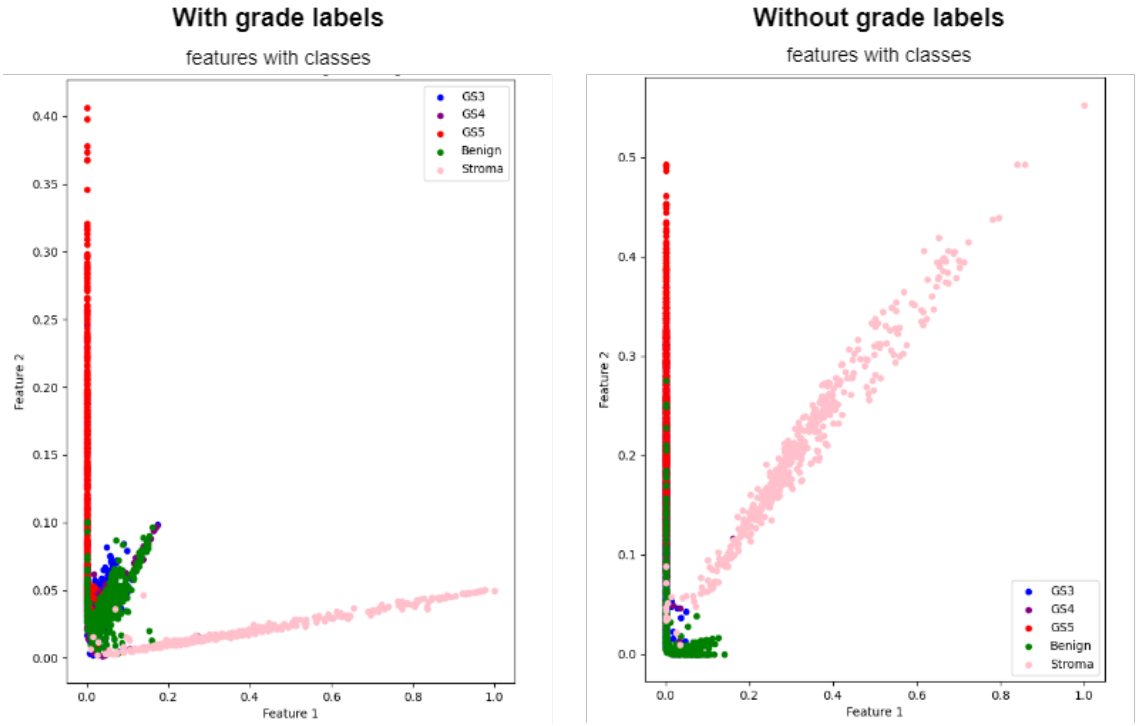


Figure C.1. The two most explaining features using CNN_{dual} with (left) and without (right) grade labels, with classes assigned to the features.

Portfolio D

This portfolio contains reflections along with documentation of the methods used for completing the master thesis of biomedical science and informatics.

D.0.1 Planning the project

Throughout this section it is described how the master thesis was planned to obtain a workflow that is efficient and agile, which is constituted by a time schedule and weekly planning of the entities of the time schedule. For the master thesis a time schedule was used to obtain an overview of the projects components. The components were arranged in a chronological order such that one component would be finished before the next one was initialized. A section of the time schedule is shown in figure D.1, were the time consuming entity was analysis of feature space in parallel with test and validation of the network.

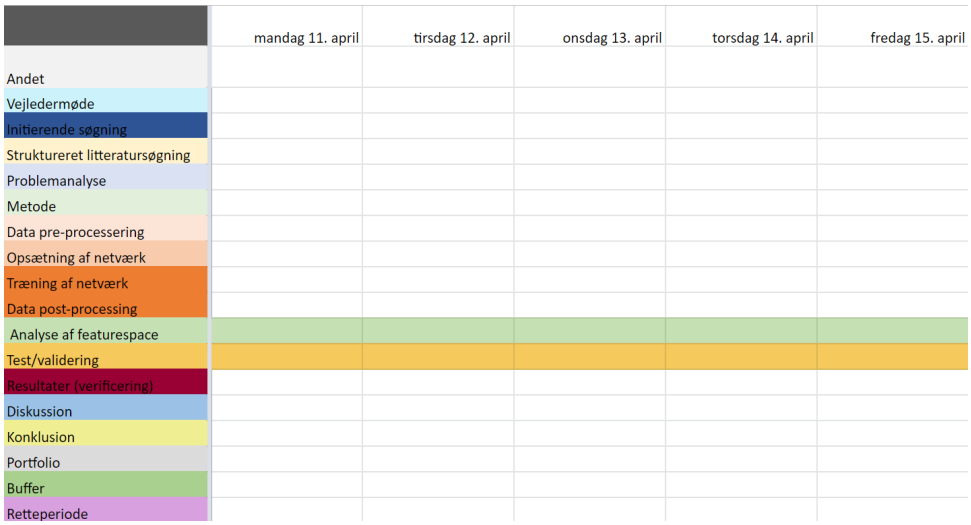


Figure D.1

This time schedule was developed using backcasting, which is a method for distributing enough time for every component of the project. [Holgaard et al., 2014] This is exemplified from figure D.1, where it is deemed important to work on two components simultaneously, to mange the projects time consuming parts within the time frame of the master thesis. Another important aspect of backcasting is the buffer component. This is implemented such that unforeseen tasks can be handled without the project components overlapping and crosses deadlines. Unforeseen tasks were typically present after a meeting with my supervisor, who had given feedback to the work done at the time. At the start of every week, the time schedule was revisited to plan in more detail how the concerning project

component is accommodated. Within the weekly schedule every hour was planned in detail to always keep track of the work that had to be done. The weekly schedule of week 18 is shown in figure D.2, as an example of how that week was planned. In addition, agendas out of the time frame of the project was also written in the weekly schedule, always ensuring that the important constituents of the weeks were readily available.

Tid Uge 18	Mandag	Tirsdag	Onsdag	Torsdag	Fredag	Lørdag	Søndag
8:15-12:00	Ret metode afsnit færdigt	Skriv det sidste på metode og begynd på at dokumentere resultater	Skriv det sidste på metode og begynd på at dokumentere resultater	Skriv det sidste på metode og begynd på at dokumentere resultater	Dokumenter resultater	Dokumenter resultater	Alle resultater skal være lagt ind i rapporten (ikke kommenteret fuld på endnu)
12:30-16:15	Ret metode afsnit færdigt	Skriv det sidste på metode og begynd på at dokumentere resultater	Skriv det sidste på metode og begynd på at dokumentere resultater	Skriv det sidste på metode og begynd på at dokumentere resultater	Dokumenter resultater	Dokumenter resultater	Alle resultater skal være lagt ind i rapporten (ikke kommenteret fuld på endnu)
Andet				Møde med netcompany (10-11 Teams)			

Figure D.2

D.0.2 Experience gained from the masters of biomedical science and informatics

Throughout the master thesis i have worked alone to understand what the benefits and disadvantages are as a contrast to working in groups. In relation to prepare a scientific work, decisions to use and implement certain methods are solely dependent on the author, which is a great responsibility and very difficult. It is difficult since the consequence of a less effective or wrong decision could be to redo previous work. And when there is only one author, few setbacks increases the work amount substantially, which in turn effects the quality of the work when there is a deadline. This is opposed to working in groups, where there is more room for mistakes, since the workload can be spread among group members. A benefit from working alone is that i am forced to work on the components of the project that i consider difficult as opposed to primarily working on for example methods and implementation. This ensures that i gain experience and thus improve my skills in for example writing a problem analysis. The fact that i have experienced working in groups and alone has taught me how to manage and complete relatively big projects by myself, which gives great confidence that i can transition from being a student into a workplace, where greater responsibility is expected along with the ability to work with different people.