Classification of Non-Small Cell Lung Cancer Stage using a Convolutional Neural Network

Aalborg university, 04/02/2019 - 06/06/2019

Master's Thesis in Biomedical Engineering & Informatics
School of Medicine and Health

[Lyons, 2019]

Produced by
Group 10402

Group members:
Josefine Dam Gade & Line Sofie Hald

AALBORG UNIVERSITY
DENMARK

School of Medicine and Health
Biomedical Engineering & Informatics
Niels Jernes Vej 12
9220 Aalborg Øst
http://www.smh.aau.dk

Title

Classification of Non-Small Cell Lung Cancer Stage using a Convolutional Neural Network

Project period

04/02/2019 - 06/06/2019

Project group

10402

Members

Josefine Dam Gade
Line Sofie Hald

Supervisor

Alex Skovsbo Jørgensen

Pages: 80
Appendices: 2
Submission date: 06/06/2019

Abstract:

Lung cancer is the leading cause of cancer-related mortality, of which non-small cell lung cancer is the most common type, that furthermore is divided into four subtypes. The primary tumor of these subtypes can be divided into four stages of cancer. This is performed by pathologists on histopathological images. However, due to the number of cells in a sample and the staining intensities which can vary, inter-observer variability is likely to occur when diagnosing tumor subtypes and stages. Existing methods have classified lung cancer into subtypes using CNNs, but it has not yet been investigated whether a CNN could be used to determine the stage of lung cancer, which will be investigated in this study. Data from The Cancer Genome Atlas has been acquired from the bronchus and lung subset. Data selection and preprocessing were performed to prepare the data and ensure homogeneity in the datasets, and the AlexNet was fine-tuned using a pretrained model. Four different experiments were conducted due to a concern about noisy labels. Furthermore, three WSI-classification approaches to evaluate the predicted labels were performed in each of the experiments. The trained model was tested patch-based and WSI-based. The patches were classified with an accuracy of 0.52±0.04 during the four experiments. The experiments with the best results had an accuracy of 0.56 in all three approaches. Additionally, two of the approaches had the highest accuracies of 0.56±0.01 when classifying WSI-based.

Resume

Lungekræft er den ledende årsag til kræftrelateret dødsfald, hvoraf ikke-småcellet lungkræft er den hyppigste kræfttype. Denne type er ydermere inddelt i fire undertyper, hvoraf den primære tumor af disse undertyper kan inddeles i fire stadier af kræft. Denne detektering bliver lavet af patologer ud fra histopatologiske billeder. Grundet den mulige variation i antallet af celler i en histopatologisk vævsprøve samt variation i staining intensiteten, er der mulighed for at interobservatør variabilitet opstår ved diagnosticering af tumor undertyper og stadier.

Eksisterende metoder har klassificeret undertyper af lungekræft ved brug af CNNs, men det har endnu ikke være undersøgt hvorvidt et CNN kan anvendes til at bestemme stadiet af lungekræft, hvilket vil blive undersøgt i dette studie. Data fra The Cancer Genome Atlas er blevet hentet fra bronchus og lunge sættet. Dataudvælgelse og præprocessering var udført for at forberede dataet til netværket og sikre homogenitet i datasættene. Herefter var AlexNet fine-tunet i de sidste to lag ved brug af initierende vægte fra en prætrænet model. Fire forskellige eksperimenter var udført grundet en bekymring om støjfyldte labels. Derudover var der udført tre WSI-baserede testmetoder for hvert eksperiment til at evaluere de forudsete patch-baserede labels.

De trænede modeller var testet patch-baseret. Patchene var klassificerede med en nøjagtighed på 0.52±0.04 under de fire eksperimenter. Eksperimenterne med de bedste testresultater havde en nøjagtighed på 0.56 i alle tre WSI-baserede klassificeringsmetoder. Ydermere havde to af de WSI-baserede klassificeringsmetoder den højeste nøjagtighed på 0.56±0.01.

Preface

This master's thesis project has been produced by group number 10402 from the 4th semester of the master in Biomedical Engineering and Informatics at Aalborg University. The project has been produced within the period from the 4th of February until the 6th of June 2019.
The scope of this semester was to acquire knowledge at the highest international level within selected research areas of pattern recognition and image analysis, and work on problems which were complex, non-deterministic, and require innovative solutions. The aim of this project was to design, develop, and test an automatic method to classify the stage of non-small cell lung cancer.
The project group would like to thank the supervisor Alex Skovsbo Jørgensen for supervision and guidance throughout the project period.

Reading Guide

The project consists of five chapters, a bibliography, and two appendices. The chapters, sections, subsection, figures, and tables are numbered based on the chapter they appear in. The appendices are indicated with a letter and the figures and tables are indicated with both letters and numbers. Harvard referencing is used in this project, by which a reference is given as squared brackets containing the authors and year of publication. A reference is inserted before a full stop if it refers to the previous sentence, while a reference is inserted after a full stop if it refers to all the sentences up until the previous reference. All references appear in alphabetical order in the bibliography with a maximum of three authors mentioned followed by et al. Abbreviations are used in the project, of which the term is spelled out and followed by the abbreviation in parentheses the first time it is used.

Contents

Introduction 1

Cancer is the second most frequent cause of death worldwide, only outnumbered by cardiovascular diseases. Hereof, approximately 8.9 million people died of cancer-related reasons in 2016. The number of people diagnosed with cancer worldwide has increased over the past century. Between 1990 and 2016, the number of people with cancer has more than doubled, and the number of people dying from cancer-related reasons has increased. However, if this number is standardized according to the increased life expectancy and population size, the number of people dying from cancer per 100,000 citizens has decreased during this period. The number of people having a type of cancer ranged between 0.2% to 2% in different countries worldwide in 2016. Of these, the most common type of cancer was breast cancer, which 0.12% of the population had. Compared to this, tracheal, bronchus, and lung cancer was the fourth most common type of cancer which 0.04% of people had worldwide. However, the leading cause of cancer-related mortality was tracheal, bronchus, and lung cancer, which caused 25.82 deaths per 100,000 citizens. [Roser and Ritchie, 2019]

Lung cancer including throat, esophagus, and bronchus was likewise not the most frequent type of cancer in Denmark, but it was the type of cancer leading to the most cancer-related death incidents. Hereof, it led to 23.6% of all cancer-related deaths in 2016. [Statistik, 2016] Cancer-related mortality has been descending in Denmark, but Denmark was the country with the most cancer-related death incidents compared to the other Nordic countries. Furthermore, death rates obtained by the Organisation for Economic Co-operation and Development (OECD) have been compared for 41 countries from the western part of the world, in which Denmark was the 7th highest ranked with most cancer-related deaths per 100,000 citizens in 2015. [Statistik, 2018]

Cancer is a result of uncontrolled mitosis which leads to the formation of a tumor. The reason for developing cancer is different among the different cancer diseases. [Pecorino, 2012] For lung cancer, the most frequent risk factor contributing to cancer is smoking, which has an impact on three-quarter of these incidents [Brody et al., 2014].
There are different types of lung cancer, which require different treatments. Therefore, the type and stage of lung cancer must be diagnosed as the treatment plan is determined based on this. Furthermore, the prognosis for the patient depends on the stage and type of lung cancer, as some types are more aggressive, cause more metastases, and have lower survival rates. [Houlihan and Tyson, 2012] Thus, diagnosis and treatment of lung cancer early in the course of the disease can be important for the survival rate [McCance et al., 2010].

This leads to the following initiating problem:

Which biomarkers are important to diagnose lung cancer and the stage hereof, and which approaches have been developed to diagnose lung cancer early in the course of the disease?

Background 2

This chapter presents the background for the project, including the anatomy of the lungs and the physiological and anatomical alterations due to lung cancer. Furthermore, the different types of lung cancer are described along with methods for detection, which leads to the problem statement and objectives for this project.

Lung cancer has a five-year survival rate of approximately 15% [Houlihan and Tyson, 2012]. Lung cancer was the fourth most common type of cancer which 0.04% of people suffered from worldwide in 2016 [Roser and Ritchie, 2019], and additionally it was the type of cancer which is associated with the highest cancer-related mortality worldwide [Houlihan and Tyson, 2012; Roser and Ritchie, 2019]. In Denmark, it led to 23.6% of all cancer-related deaths, by which it further more was the most common reason for cancer-related deaths in Denmark [Statistik, 2016].

2.1 Lung Anatomy

The left and the right lung are surrounded by pleural cavities which are located in the thoracic cavity. The plural cavities contain fluid that assists in the optimal function of the lungs during respiratory movement. The location of these are often referred to as the left or right hemithorax, that are separated by the mediastinum, which is a mass of tissue, that surrounds, stabilizes, and supports the trachea, esophagus, thymus, and major vessels. [Martini et al., 2012]

In figure 2.1, the branching of the left bronchus is illustrated. The inhaled air passes through the trachea, which branches into the right and left primary bronchi. These supply air to the right and left lung respectively. The primary bronchi branches into secondary bronchi in each lobe of the lungs. These branch further into tertiary bronchi which supply air to the bronchopulmonary segments. Each bronchopulmonary segment contains several bronchioles, which ultimately branch into terminal bronchioles, that deliver air to the alveoli in a pulmonary lobule. Here, the gas exchange between air and blood takes place in the air-filled pockets within the alveoli. [Martini et al., 2012]

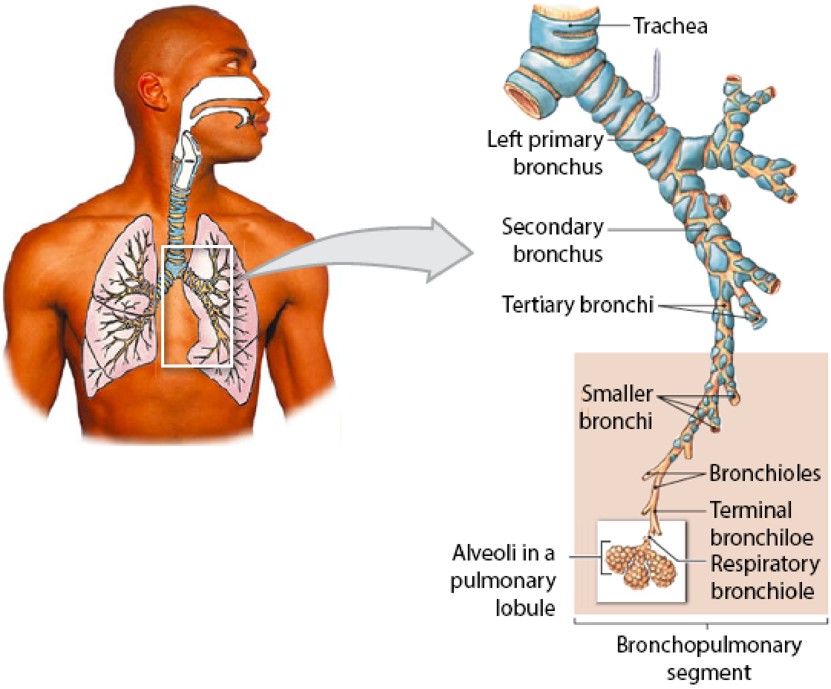
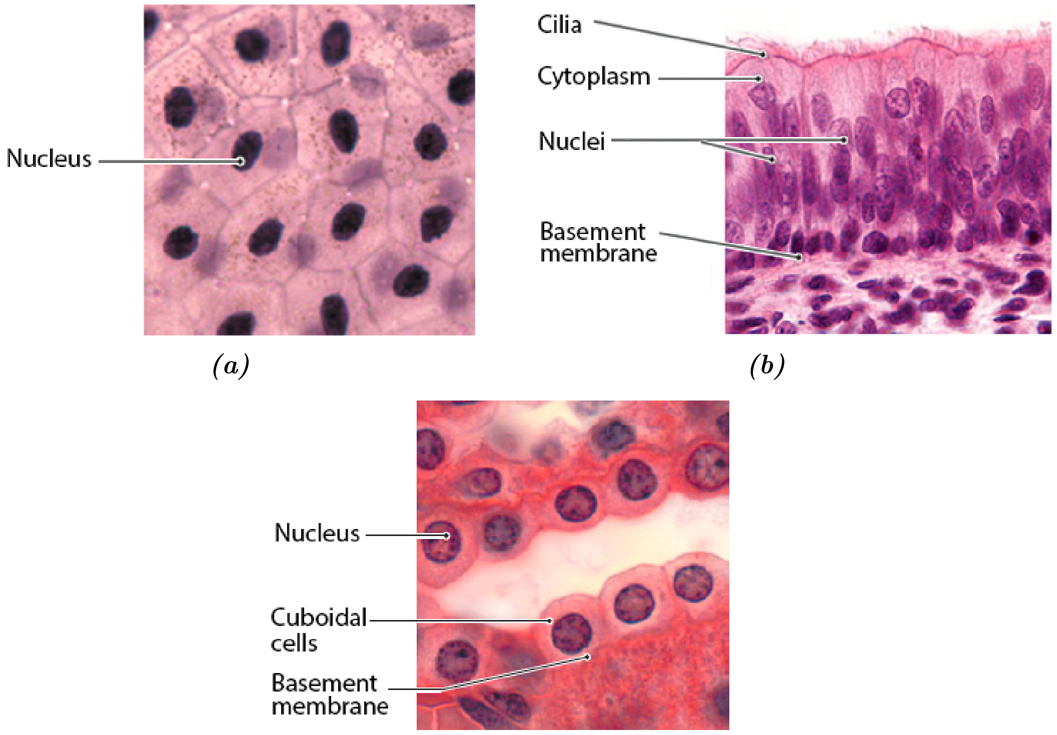


Figure 2.1: *Branches from the trachea to the alveoli. Modified from Martini et al. [2012].*

The epithelial tissue covers the inner and outer surface of the lungs and forms glands. Three examples of epithelial tissue are illustrated in figure 2.2. The epithelial tissue provides physical protection from abrasion, dehydration, and destruction by chemical or biological agents, and produces secretions in the glands, which provide physical protection or temperature regulation. The epithelial cells can be categorized based on the shape of its cells and the number of cell layers between the epithelial basement membrane and exposed surface. The shapes of the epithelial cells are categorized in: Squamous, cuboidal, and columnar, while the number of epithelial cell layers are identified as simple or stratified. When categorizing the shape of the epithelial cell in a sectional view perpendicular to the exposed surface and the basement membrane, the squamous cells appear thin and flat, cuboidal cells appear as small boxes, and columnar cells appear as tall and relatively slender rectangles. [Martini et al., 2012]



(a) *(b)*

(c)

Figure 2.2: *Tissue samples of: Squamous epithelium (a), columnar epithelium (b), and cuboidal epithelium (c). Modified from Martini et al. [2012].*

The number of epithelial cell layers is defined as simple epithelium when only one layer of cells covers the basement membrane, which makes it very thin. This is an advantage in the simple squamous epithelium in the alveoli where gas-exchange happens, as it reduces the time for absorption or diffusion across the epithelial barrier. Simple squamous epithelium that appears in the ventral body cavity, such as the pleural cavity of the lungs, is called mesothelium. An example of a tissue sample of simple squamous epithelium is illustrated in figure 2.2a. If several layers of epithelial cells cover the basement membrane, the cell layers are defined as stratified epithelium, which is thicker and appears on locations where more mechanical or chemical stress occurs. [Martini et al., 2012]
Columnar epithelial cells are densely packed and appear rectangular. Within the respiratory tract, pseudostratified columnar epithelium appears, which includes several types of cells, which have varying shapes and functions, and often possess cilia. As an example, pseudostratified ciliated columnar epithelium appears in the lining of the trachea and bronchi, which is illustrated in figure 2.2b. [Martini et al., 2012]
Cuboidal epithelium cells resemble hexagonal boxes, in which the spherical nuclei are located near the center of the cell and the distance between the nuclei is approximately equal to the height of the epithelium. A collection of epithelial cells that produce secretions is called a gland. An example of cuboidal epithelial cells in the glands is illustrated in figure 2.2c. [Martini et al., 2012]

2.2 Lung Cancer

Most lung cancer cases are associated with smoking, but for one-quarter of these cases this relation is not present. Among these non-smoking related cases, the cause of lung cancer is a combination of environmental and genetic risk factors, especially in Asia where the food may be cooked indoor on cooking stoves with poor ventilation. [Brody et al., 2014] However, smoking is the greatest promoter for lung cancer due to the 81 carcinogens and the chronic inflammation which often follows. The biological products which fight the inflammation also cause DNA damage, which thereby causes an increased mutation range. [Pecorino, 2012]

Cancer is a result of an alteration in the DNA, which can be smaller alterations in a single chromosome pair or larger chromosomal aberrations. The accumulation of the mutation happens over time, which is the reason why the risk of cancer increases with age. The mutations adapt to the microenvironment of the cell as the growth of the mutated gene advantages compared to the neighboring genes. The accumulation of the mutation occurs when the cells defense mechanism, which repairs the DNA, fails. The mutations, which are not repaired in the DNA before mitosis, which is the division of a cell, passes permanently to the daughter cell. This can result in severe DNA damage which induces apoptosis, which is a mechanism where the cell kills itself, that happens to protect the body from cell transformation. When the tumor starts growing at a distant site, new formations of blood supply are developed to provide oxygen and nitrogen to the cells, which is called angiogenesis. In angiogenesis, the new blood supply to the tumor is formed growing from pre-existing blood supply. [Pecorino, 2012]

The three main steps that contribute to the number of cells in an individual are cell proliferation, apoptosis, and differentiation. Cell proliferation is the most obvious, in which a cell is divided into two daughter cells. Apoptosis influences the number of cells, as they are reduced due to abnormalities in the cells. Differentiation is a process of which the cell can enter an inactive phase of cell growth during the process, which may affect the number of cells. When the DNA in the cells related to cell growth is changed, unregulated cell growth can occur. The cell growth is regulated by positive and negative molecular factors. There are two major types of mutated genes which influence cancer: Oncogenes and tumor suppressor genes. An oncogene is a mutated gene in which the protein product is produced in increased quantity or activity, and thereby is dominant in the initiation of tumor formation. In the oncogenes, either the protein production is higher within the cells or the activity of the altered products is increased by which they become dominant and easier initiate tumor formation. The tumor suppressor genes inhibit both the growth factor and tumor formation. [Pecorino, 2012]

Cancer is characterized by uncontrolled mitosis which leads to the formation of a tumor. Normally, mitosis is initiated by environmental factors outside the cell. A signal is transmitted to the cell and the nucleus by which the gene expression is regulated so that proteins essential for mitosis are produced. [Pecorino, 2012] In the prophase of mitosis, spindle fibers are extended between a centriole pair while the chromosomes are coiled tightly, as illustrated in figure 2.3. During the late prophase, DNA has been replicated by which copies of the chromosomes, called chromatids, are generated and attached to the duplicate at a single point, called the centromere. These are attached to spindle fibers forming chromosomal microtubules. During the metaphase, the chromosomes move to the metaphase plate. After this the centromeres split in the anaphase and the daughter chromosomes are pulled towards the opposite ends of the cell along the chromosomal microtubules. During the telophase, the nuclear membrane reforms, the chromosomes start to uncoil, and cleavage furrow if formatted. Thereafter, cytokinesis can begin, in which the cells are fully divided into two daughter cells. [Martini et al., 2012]

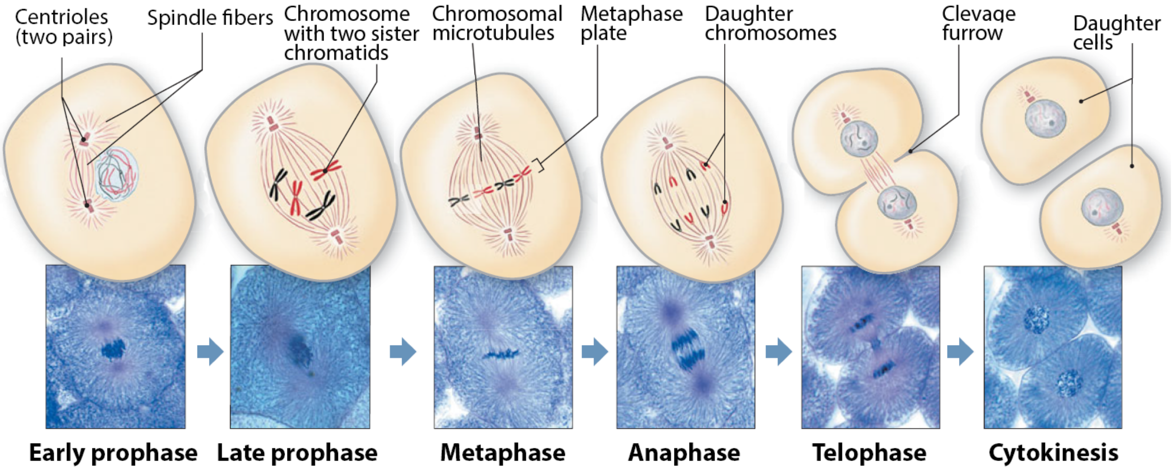


Figure 2.3: *The stages of a mitosis into two new daughter cells [Martini et al., 2012].*

Tumor necrosis is another process which kills tumor cells characterized by membrane disruption and slows the tumor growth [Martini et al., 2012; Pecorino, 2012]. Thereby, tumor necrosis affects the rate of tumor growth [Pecorino, 2012]. A tumor can either be benign or malignant, of which the benign tumor does not metastasize, but some can be life-threatening due to its locations. However, the malignant tumor does not remain encapsulated and will thereby infiltrate normal tissue and result in metastasis. The onset of the tumor infiltration is highly dependent on the type of cancer. Some types of lung cancer infiltrate normal tissue early in the course of the lung cancer, while the infiltration happens late in the course for some types of lung cancer, as described in table 2.1. [Pecorino, 2012]

Metastasis happens when the cancer has spread from the primary site to other parts of the body which then becomes secondary sites as illustrated in figure 2.4. By intravasation, cancer cells penetrate into the bloodstream, in which they are spread singly or as emboli in the direction of the blood. Therefore, cancer in different parts of the body has particular organs in which the metastasis first occurs, as the dimensions of the tumor cells are larger than the capillaries, the tumor cells get stuck in the first capillaries that they encounter. Lastly, by extravasation, the tumor cells escape the bloodstream or lymphatic vessel by which they can migrate to the surrounding stroma. However, only very few tumor cells which enter the bloodstream survive and thereby cause metastases in other parts of the body. [Pecorino, 2012]

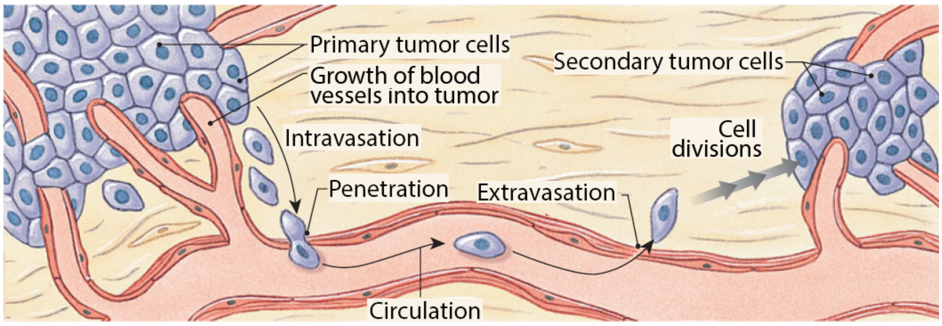


Figure 2.4: *Metastasis from primary tumor to secondary tumor cells. Modified from Martini et al. [2012].*

Lung cancer is, according to the World Health Organization (WHO), classified in four major histological types: Squamous cell carcinoma (SCC), adenocarcinoma, small cell lung carcinoma, and large cell carcinoma. These lung cancer types are grouped into two main categories: Small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). [Houlihan and Tyson, 2012] An overview of the SCLC and NSCLC subtypes is depicted in table 2.1.

Type of lung cancer		Frequency	Growth rate	Metastasis
SCLC	Small cell carcinoma	13%-15%	Very rapid	Very early
NSCLC	Large cell carcinoma	10%-15%	Rapid	Early
	Squamous cell carcinoma	30%	Slow	Late
	Adenocarcinoma	35%-40%	Moderate	Early

Table 2.1: *Overview of the frequency out of all lung cancer cases, growth rate, and metastasis of the different lung cancer subtypes. Modified from McCance et al. [2010].*

As depicted in table 2.1, the most common types of lung cancer are SCC and adenocarcinoma, of which SCC occurs in the epithelial cells and adenocarcinoma occurs in the glandular cells [Pecorino, 2012].

2.2.1 Small Cell Lung Cancer

SCLC represents approximately 13%-15% of all lung cancer cases but accounts for approximately 25% of all lung cancer-related deaths and thereby has the worst prognosis. SCLC is categorized into two subtypes: Small cell carcinoma and combined small cell carcinoma. [Houlihan and Tyson, 2012; McCance et al., 2010] Hereof, combined small cell carcinoma is a variant of small cell carcinoma with a component of any histologic subtype of NSCLC [Houlihan and Tyson, 2012]. SCLC is a neuroendocrine tumor, which arises from basal neuroendocrine cells, and is often located in the central airways and has a size of 6 to 8 μm [Houlihan and Tyson, 2012; McCance et al., 2010].

The cell growth happens rapidly in SCLC and tends to metastasize early and widely, as depicted in table 2.1. If SCLC is untreated it is the most aggressive clinical course of any pulmonary tumors, with a median survival time of two to four months after diagnosis. Treatment with chemotherapy, radiation, or both, benefits approximately 90% of patients, but most patients relapse within 2 years. [Houlihan and Tyson, 2012; McCance et al., 2010] Chemotherapy inhibits DNA metabolism, and thereby blocks DNA synthesis in the rapidly division of cancer cells, by targeting vulnerabilities in the cancer cells, while radiation usually are delivered to the tumor and reacts with the water inside the cells which result in damage to the DNA [McCance et al., 2010; Pecorino, 2012].

2.2.2 Non-Small Cell Lung Cancer

NSCLC represents 75%-85% of all lung cancer cases and includes three main subtypes: Squamous cell carcinoma (SCC), adenocarcinoma, and large cell carcinoma [Houlihan and Tyson, 2012; McCance et al., 2010; Pecorino, 2012]. An overview of the growth rate and metastasis arising from these three subtypes of NSCLC are depicted in table 2.1.

Adenocarcinoma is the most common subtype of lung cancer and represents approximately 35%-40% of all lung cancer cases [Houlihan and Tyson, 2012; McCance et al., 2010]. This type of tumor arises from the glands and is usually smaller than 4 cm. Surgical removal of the tumor is often possible, but since metastasis occurs early in the course of the cancer, the five-year survival rate is below 15%, which chemotherapy can be used to increase. [McCance et al., 2010]

SCC accounts for approximately 30% of all lung cancer cases and involves squamous epithelium which has been replaced with malignant squamous cells. This subtype tends to be slow growing and may take years to develop from carcinoma to clinically evident tumor [Houlihan and Tyson, 2012; McCance et al., 2010]. These tumors are typically located centrally close to the hilus, which is a shallow indentation where the blood vessels and nerves reach the lymph node. Furthermore, they project into the bronchi, and tend not to metastasize until late in the course of the disease. Therefore, surgical treatment is preferred, but total surgical removal of the tumor may not be possible if metastasis has happened. Chemotherapy can additionally be used to improve survival and quality of life but has limited effectiveness against SCC. [McCance et al., 2010]

Large cell carcinoma represents approximately 10%-15% of all lung cancer tumors, and is thereby the least common of the NSCLC tumors [Houlihan and Tyson, 2012; McCance et al., 2010]. These tumors are fast-growing and metastasis occurs early in the course, by which surgical removal is limited to palliative procedures, and neither radiation nor chemotherapy increases survival. The cancer cells of this type of tumor are generally larger than leukocytes and contain large, darkly stained nuclei, but show none of the histological findings of SCC or adenocarcinoma, and are therefore diagnosed through a process of exclusion. [McCance et al., 2010]

2.3 Diagnosis of Lung Cancer

An accurate diagnosis and staging of lung cancer is essential, as this is used to determine the treatment plan and thereby may impact the prognosis [Houlihan and Tyson, 2012]. Therefore, it is essential for long-term survival, that lung cancer is diagnosed and treated early in the course of the disease [McCance et al., 2010].

2.3.1 Preliminary Diagnosis

To make a diagnosis, different clinical examinations are performed with the aim of detecting the presence of a primary lung cancer, determining the cell type in the tumor tissue, and perform staging of the tumor. The evaluations include a review of the patient history and a physical examination. During the physical examination, noninvasive methods including sputum cytology or image modalities such as chest x-ray, computed tomography (CT), or positron emission tomography (PET) can be utilized when searching for abnormalities. Chest x-ray and CT scans provide anatomical information, whereas PET scans provide functional tumor information. When combining CT and PET it is possible to achieve a higher level of accuracy in the evaluation of the lung cancer stage. [Houlihan and Tyson, 2012; McCance et al., 2010]

2.3.2 Extraction of Histological Samples

If the noninvasive tests indicate abnormalities in terms of growth factors, signaling, or changes to the oncogenes and tumor suppressor genes, diagnostic tests are performed. These can be performed through a biopsy of the mediastinal lymph nodes, a transthoracic needle aspiration (TTNA), which is performed through the chest wall, or a transbronchial needle aspiration (TBNA) during a bronchoscopy. A bronchoscopy enables direct visualization of endobronchial abnormalities and direct sampling through exfoliation or aspiration of the abnormalities. An endobronchial biopsy can provide histologic samples and TBNA can be performed to sample lymph nodes in the mediastinum and peribronchial areas. [Houlihan and Tyson, 2012; McCance et al., 2010]

Histological samples are extracted following a clinical suspicion of lung cancer. The extraction of tissue samples is performed as exfoliation or aspiration, which is placed on a glass slide, and analyzed through microscopy by pathologists. Through exfoliation, cells that naturally are rejected from the body or cells that can be exfoliated from the surface of the epithelial tissue are collected. Through aspiration, cells are aspirated using a fine needle, which is useful as neoplastic cells are more loosely connected than normal cells and therefore are more easily aspired. [Ejersbo et al., 2014; Houlihan and Tyson, 2012] The aim of preparation of histological and cytological samples is to transfer a representative and well-preserved amount of the tissue sample to a glass slide to enable a precise diagnosis to be made [Ejersbo et al., 2014].

To be able to investigate the histological samples on a computer, photomicrography is performed by attaching a digital camera to a microscope from which images can be recorded using a computer software [Feldman and Wolfe, 2014]. This can be done as whole slide images (WSIs), which allows high-resolution digitization of the entire glass slides [Bándi et al., 2019].

2.3.3 Preparation and Staining of Tissue Samples

To examine the tissue sample, multiple steps are performed, as illustrated in figure 2.5. After extraction of the tissue sample from the patient's body, autolysis begins, which is an intracellular deterioration process. Therefore, an important step when processing tissue samples is fixation. The primary purpose of fixation is to stabilize the enzymes and proteins, and to disable microorganisms, and thereby stop autolysis and preserve the tissue. This can be performed on the glass slide with the sample using physical and chemical fixation methods. Physical fixation includes augmentation of a fixative with heat and microwave, and cryopreservation. Chemical fixations can either be additive or non-additive: The additive form cross-links by adding themselves to the tissue and forming chemical bonds, and non-additive denatures and dissociate water molecules from the tissue proteins and thereby dehydrates the tissue sample. [Qidwai et al., 2014] The preparation of the histological samples are highly important as the lack of fixation, preparation, and staining can lead to misdiagnoses, due to autolysis [Ejersbo et al., 2014]. Different fixatives can be used depending on the most important histological features, and thereby significantly impact the ability to make a diagnosis [Qidwai et al., 2014].

After fixation, the histological sample is stained to highlight the biomedical features of interest [Duraiyan et al., 2012; Kalyuzhny, 2016]. Hematoxylin and eosin (H&E) staining can be applied to the tissue sample with the aim of displaying the contrasting colors between the nucleus and cytoplasm and thereby differentiate cellular components. H&E staining is a process with multiple steps. First, the tissue sample must be cleaned for paraffin wax, to which three stations of clearance is used to enable staining in an aqueous hematoxylin solution. Clearance is insoluble with water, after which two different alcohols are used before the sample is exposed to water. Hematoxylin stains the nuclear chromatin along with other acidic cellular elements, as it binds to the DNA and carboxy groups of proteins in the nuclear chromatin [Llewellyn, 2009]. There exist multiple hematoxylin stains which each produce different colors and patterns. An example of this difference is the determination of the nuclear color, which depends on the mordants used: Iron mordants produce blue-black nuclei and aluminum mordants produce blue-purple nuclei. The unbound hematoxylin is then rinsed off with water followed by acid alcohol. After this, either Erosin B or Erosin Y dye is applied of which Y is more likely used in anatomical pathology. Erosin has a three-color effect with three shades of pink which enhances the appearance of red blood cells, collagen, and smooth muscles. After this, the histological sample is yet again rinsed with different types of alcohol. [Feldman and Wolfe, 2014]

Based on the stained tissue samples, pathologists can diagnose the type of lung cancer. A tumor is classified as adenocarcinoma if it has histologic features such as acinar and papillary formation. Furthermore, in small biopsy specimens, micropapillary and lepidic might be seen in histological samples, but these can be hard to recognize in cytologic material. On the contrary, if the tumor shows hallmarks such as keratinization, which is a superficial layer of cells filled with keratin, and presence of intercellular bridges, it should be classified as SCC. However, a problem with classification of the subtypes can arise in paucicellular specimens and in poorly differentiated carcinomas, in which no clear histologic difference is present. An example is solid-type adenocarcinoma, in which some of the features which are normally squamoid such as nested appearance and glassy eosinophilic can appear, but without keratinization or intracellular bridges. In these cases of poorly differentiated carcinoma, it is recommended to use immunohistochemical (IHC) markers to help in the classification process. [Moreira and Saqi, 2015]

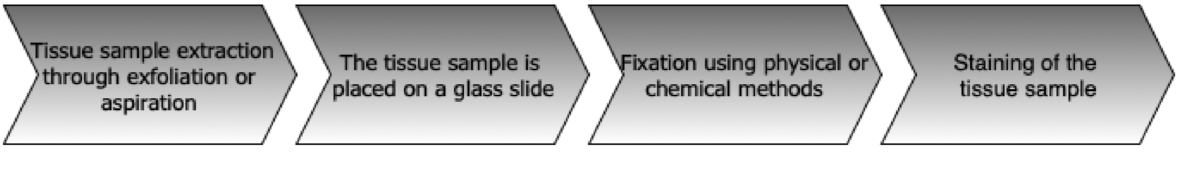

Tissue sample extraction through exfoliation or aspiration
The tissue sample is placed on a glass slide
Fixation using physical or chemical methods
Staining of the tissue sample

Figure 2.5: *The overall process of extraction of tissue samples, preparation, and staining.*

IHC staining is a method used for processing of the histological samples, that utilizes monoclonal and polyclonal antibodies. This is a product of the immune system, in the detection of specific antigens, and involves a specific antigen-antibody reaction. IHC staining is widely used in medical research and clinical diagnosis of cancer to detect tumor antigens. It has advantages compared to traditional special enzyme staining techniques, which only identify a limited number of proteins, enzymes, and tissue structures. [Duraiyan et al., 2012; Kalyuzhny, 2016]

2.3.4 Stages of Lung Cancer

The purpose of clinically staging lung cancer is to facilitate an accurate and thorough description of the extent of the lung cancer in order to determine the optimal treatment [Houlihan and Tyson, 2012]. Through staging, the tissue sample is examined for histological patterns of cancer by a pathologist, which conventionally is done under a microscope. This may be assisted by IHC analysis of protein expression. [McCance et al., 2010]
When staging lung cancer, different modalities can be utilized, including PET, CT, chest x-ray, and cytologic or histologic tissue analysis. Hereof, PET, CT, and chest x-ray are used when investigating the location of possible tumors and metastases. [Houlihan and Tyson, 2012; McCance et al., 2010] Thereby, they can be used to describe the extent of the lung cancer, of which solid tumors are classified into four groups: Stage I-IV. Information from a histological sample regarding the number, shape, and stain color of the tumor cells, are valuable in diagnosing of the subtype and stage of the lung cancer. Thereby, a histological analysis can provide an indication of the prognosis for the patient, as the prognosis is highly dependent on the subtype and stage of the lung cancer, and can be used to determine the treatment plan, which is different for the subtypes and stages. [Houlihan and Tyson, 2012]

There are different stage classifications for SCLC and NSCLC respectively, which are used to indicate the prognostic groups as the overall survival rate is different for the patients due to the effect of the tumor type. Hereof, SCLC has a rapid tumor growth rate and metastasis occurs very early in the course of this type of lung cancer, as illustrated in table 2.1, which results in a low survival rate. SCLC is staged as either limited or extensive disease, in which the limited stage is confined to one hemithorax without pericardial or pleural effusion and the extensive stage is all other presentations of the disease. Of these two stages, the limited stage is more curable than the extensive stage of the SCLC, which affects the survival rate. [Houlihan and Tyson, 2012] For the subtypes of NSCLC, both large cell carcinoma and adenocarcinoma metastases early, but the tumor growth rate is rapid for large cell carcinoma and moderate for adenocarcinoma, as illustrated in table 2.1. On the contrary, SCC has a slow tumor growth rate and metastases late in the course of the lung cancer, and may thereby take years to develop from carcinoma to clinically evident tumor. Due to these differences, the stage of the NSCLC is a major indicator for the prognosis and treatment for lung cancer, of which the encapsulation of the tumor affects the possibility of surgical removal. [Houlihan and Tyson, 2012] The subtypes of NSCLC are staged using the TNM classification, where T denotes the extent of the primary tumor, N indicates nodal involvement, and M describes the extent of metastasis [Houlihan and Tyson, 2012; McCance et al., 2010]. The primary tumor is subdivided into four categories (T1–T4) which reflects the size, location, and local involvement. If no primary tumor is evident, it is denoted (T0), while a tumor denoted as (T1-T3) is size dependent and a tumor denoted as (T4) can have any size but depends on the location. The lymph node involvement is subdivided into: No regional lymph node metastasis (N0), ipsilateral peribronchial, ipsilateral hilar lymph nodes, and/or intrapulmonary nodes (N1), ipsilateral mediastinal and/or subcarinal lymph nodes (N2), and contralateral mediastinal, contralateral hilar, ipsilateral or contralateral scalene, or supraclavicular lymph nodes (N3). Metastatic spread is denoted as either absent (M0) or present (M1). [Houlihan and Tyson, 2012]

2.4 Assistive Methods to Detect Lung Cancer

Pathologists are currently evaluating histological samples visually, which gives rise to a challenge as there can be millions of cells in one sample. Furthermore, stained tissue samples vary in color due to different medical facilities and research laboratories. This makes the process of diagnosis time consuming, highly influenced by subjectivity, and has shown to cause significant inter-observer variability when diagnosing tumor subtypes and stages. [Khosravi et al., 2017; Mishra et al., 2017] In addition, lung cancer is a heterogeneous disease with a large diversity in the genetics and phenotypes both intra-tumor and inter-tumor [Houlihan and Tyson, 2012; Khosravi et al., 2017] This heterogeneity is reflected in the different cellular and molecular changes. An accurate diagnosis of the type and stage of lung cancer has to be made as the treatment is based on it, which impacts the prognosis of the patient and ultimately the long-term survival [Houlihan and Tyson, 2012; McCance et al., 2010]. This indicates a need for assistive techniques when diagnosing lung cancer patients, but still facilitate challenges with computational pathology. This is due to the tissue samples varying in color, shape, and batch effects due to different staining protocols at the medical facilities and research laboratories. [Khosravi et al., 2017; Mishra et al., 2017]

Different studies have investigated methods for detection and staging of lung cancer in histopathological WSIs. The primary focus has been to distinguish SCC and adenocarcinoma, which are the two most prevalent types of lung cancer. The process of distinction between these two types of NSCLC is not always clear, by which it becomes a challenging and time-consuming process for pathologists. [Coudray et al., 2018] In existing work for tumor classification, different image techniques have been used, such as threshold with region growing, k-means, otsu, and morphological features such as area and shape structure [Mishra et al., 2017]. Furthermore, studies have investigated the use of machine learning and neural networks with the aim to detect cancer more accurately, which have been performed on the WSIs or smaller patches [Bándi et al., 2019; Coudray et al., 2018; Khosravi et al., 2017; Mishra et al., 2017; Yu et al., 2016].

Databases containing tissue images have been constructed to investigate and analyze cancer tissue samples. Two of these are The Cancer Genome Atlas (TCGA) and The Stanford Tissue Microarray Database (TMAD). TCGA is a research program that aims for providing a data sharing platform for the cancer research community, which enables the researchers to improve the precision of medicine in oncology [NCI, 2019a,b]. TCGA is a part of the Genomic Data Commons (GDC) from the National Cancer Institute (NCI) [NCI, 2019a,b], which have been used in several studies which have applied methods to distinguish SCC and adenocarcinoma [Coudray et al., 2018; Khosravi et al., 2017; Yu et al., 2016]. TMAD is a public resource containing annotated tissue images and associated information. This data can be used by researchers worldwide to design, view, score, and analyze the tissues. [Marinelli et al., 2007] The data from TMAD have been used in several studies investigating the distinction of cancer subtypes [Khosravi et al., 2017; Yu et al., 2016].

A study by Coudray et al. [2018] investigated the classification of SCC and adenocarcinoma in patches from WSIs acquired from the TCGA using a convolutional neural network (CNN). With the constructed CNN, the study achieved an area under the curve (AUC) of 0.97 when classifying between the two types of lung cancer. The results from the CNN were compared to classifications performed by three pathologists, of which 50% of the misclassified WSIs by the network were also misclassified by at least one of the pathologists. Furthermore, of the WSIs in which at least one of the pathologists misclassified the slide, 83% of these cases were correctly classified by the network. [Coudray et al., 2018]

A study by Yu et al. [2016] made a combination of conventional thresholding and image processing with machine learning techniques to distinguish malignancy from normal adjacent tissue in adenocarcinoma and SCC WSIs from TCGA and cropped slide images from TMAD. The images from TCGA and TMAD were tiled into patches. To make this distinction, a total of seven classifiers using random forest, support vector machines (SVM) or naive Bayes classification were tested, which had an average AUC of 0.81. For this, 80 quantitative features were used for SCC, including features of the nuclei (sum of the variances, difference of the variances, and correlation coefficient of adjacent pixels), radial distribution of pixel intensity, and intensity mass displacement of the cytoplasm. 240 features were used for adenocarcinoma, including texture features of the nuclei (sum of the entropies, difference of the variances, and angular second moment), edge intensity of the nuclei, texture features of the cytoplasm and intensity distribution of the cytoplasm. Hereof, the three best-performing classifiers were the classifier trained with SVM with a Gaussian kernel, the classifier trained with random forest using conditional inference trees, and the classifier trained with Breiman's random forest. Furthermore, it was desired to distinguish adenocarcinoma from SCC. For this, the classifiers which performed the best were trained with SVM's with Gaussian kernel and random forest classifiers respectively. To facilitate extraction of features specific to the tumor nucleus and cytoplasm, these were segmented prior to feature extraction. The study found that the top quantitative features were texture in the tumor nucleus and cytoplasm and radial distribution in the pixel intensity. [Yu et al., 2016]

A study by Deniz et al. [2018] used two widely known network architectures, AlexNet and VGG16, to distinguish benign and malignant tissue in histopathological images of breast cancer. They used transfer learning to investigate three different approaches to their classification problem. In the first approach, feature extraction was utilized to extract feature vectors from the first fully connected layer in both AlexNet and VGG16 using pre-trained models, after which these feature vectors have been concatenated. In the second approach, feature extraction was utilized in the same way as in the first approach, but for the second fully connected layer in both AlexNet and VGG16. In the third and last approach, the AlexNet was fine-tuned to represent this classification problem. It was found that the approach using fine-tuning of the AlexNet outperformed the two other approaches. Hereof, the accuracy using images with a magnification level of 40x was 84.87% in the first approach, 84.58% in the second approach, and 90.96% in the third approach. [Deniz et al., 2018]

The constructed networks have been developed as it is hard for pathologists to diagnose cancer tissue from normal tissue, and even harder to divide a type of cancer into subtypes [Bándi et al., 2019; Coudray et al., 2018; Khosravi et al., 2017; Mishra et al., 2017; Yu et al., 2016]. CNNs have been widely used in deep learning approaches for medical images as they can handle very large datasets and detect complex patterns. This can be used to classify biomedical structures and patterns with high accuracy, which is highly important for medical applications [Chollet, 2018; Khosravi et al., 2017; Ngo et al., 2016]. The stage of the tumor is furthermore important to diagnose, as the treatment and prognosis depend on it. So far, the studies have focused on dividing lung cancer into the subtypes: SCC and adenocarcinoma. However, to our knowledge, CNNs which automatically classifies the stage of the histological sample remain uninvestigated. [Houlihan and Tyson, 2012]

2.5 Problem Statement

Lung cancer was the leading cause of cancer-related mortality worldwide in 2016 which caused 25.82 deaths per 100,000 citizens [Roser and Ritchie, 2019], and is divided into four major histological types: SCC, adenocarcinoma, small cell lung carcinoma, and large cell carcinoma. It is important that the cancer type and its stage is diagnosed correctly by a pathologist, as each subtype of lung cancer has a different prognosis and is treated differently. Furthermore, it is important that lung cancer is diagnosed at an early stage, as the five-year survival rate is at 15%. [Houlihan and Tyson, 2012] Approaches have been proposed in order to distinguish SCC and adenocarcinoma, which are the most frequent types of lung cancer. For this, image analysis techniques have been used including CNNs. However, it has not been investigated whether a CNN could be used to determine the stage of lung cancer. Therefore, this could be interesting to study as the stage of the lung cancer additionally has an impact on the prognosis for the patient.

How can a CNN be developed to detect the stages of lung cancer in H&E stained whole slide images from TCGA?

To examine the problem statement, the following objectives will be investigated:

- Develop a preprocessing method for the H&E stained images
- Develop and train a CNN to detect the stages of lung cancer
- Perform quantitative analysis of the CNN output
- Validate the output from the CNN

Methods 3

This chapter presents a description of the data selection, data preprocessing, the architecture of the CNN, and the training, validation, and test hereof.

The main purpose of the methods used in this project was to develop a CNN that can detect the stage of lung cancer in H&E stained WSIs from TCGA. To accomplish this, the pipeline of this project, which is illustrated in figure 3.1, includes five steps: Data selection, data preprocessing and dataset construction, patch extraction, training of the CNN, and test of the trained model. To achieve this, MATLAB was used during the data selection, while the remaining steps were conducted using programming in Python using TensorFlow.

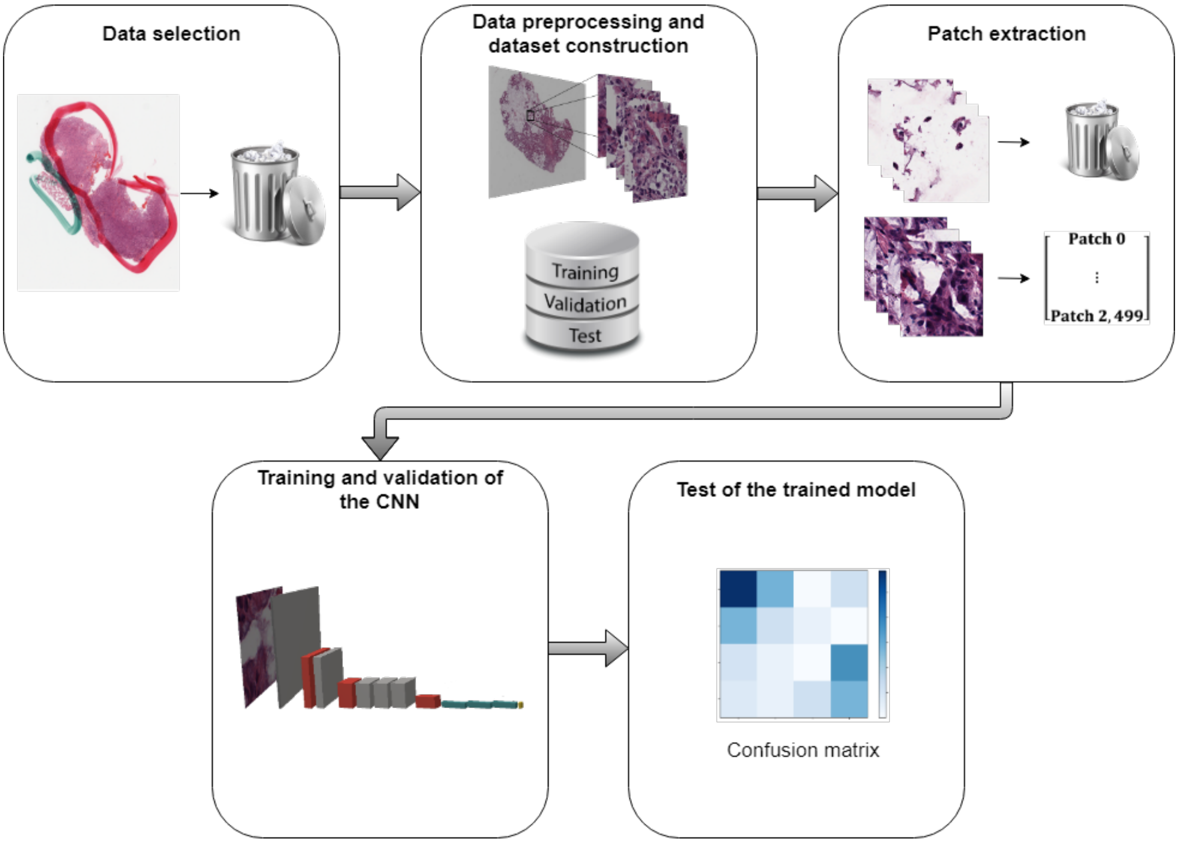

Data selection
Data preprocessing and dataset construction
Training
Validation
Test
Patch extraction
Patch 0
⋮
Patch 2, 499
Training and validation of the CNN
Test of the trained model
Confusion matrix

Figure 3.1: *The pipeline of the project including data selection and preprocessing, construction of the dataset, training of the CNN, and test of the trained model.*

The first step in figure 3.1 illustrates the data selection, in which data with artifacts and biases not related to the tissue sample, were removed. In the second step, illustrated in figure 3.1, the data was preprocessed by extracting subimages from the WSIs, after which the data was split into three datasets: Training, validation, and test. In the third step in figure 3.1, patches were extracted from the subimages and patches which did not contain a certain amount of the tissue sample were excluded. In the fourth step illustrated in figure 3.1, a CNN was trained with the constructed training and validation datasets. As illustrated in the fifth step in figure 3.1, the trained model was tested, and the results hereof were evaluated.

3.1 Data Selection

The data consisted of histopathological images with metadata, acquired from the bronchus and lung subset of TCGA (https://portal.gdc.cancer.gov). The image data was acquired as WSIs stored in SVS files, which is a single-file format of stacks of Tagged Image File Format (TIFF) files. Metadata for the WSIs was acquired from a JavaScript Object Notation (JSON) files for clinical and biospecimen information, respectively. In these JSON files, metadata associated with each WSI was stored and included subject demographics, WSI-based annotation of lung cancer subtype and stage, and percentage of tumor nuclei in the WSI. The conversion of the SVS files and the data selection is illustrated in figure 3.2.

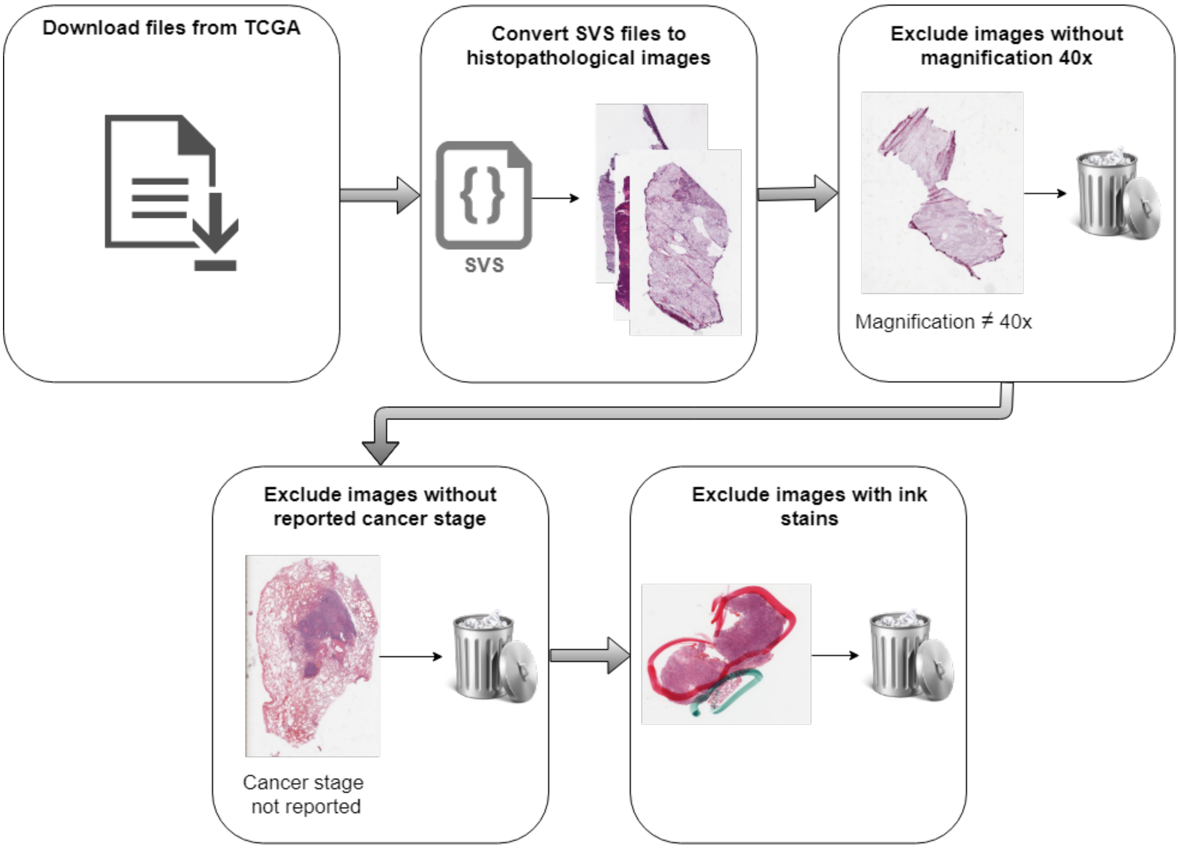

Download files from TCGA
Convert SVS files to histopathological images
SVS
Exclude images without magnification 40x
Magnification ≠ 40x
Exclude images without reported cancer stage
Cancer stage not reported
Exclude images with ink stains

Figure 3.2: *Exclusion of data to ensure quality of the data used.*

In TCGA, data for 1,026 subjects with a total of 3,220 WSIs with metadata were available. These were attempted to be downloaded, as illusrated in the first step in figure 3.2. The demographics for all the subject available in TCGA, including the number of subjects, number og WSIs, mean age at diagnosis, and gender, are illustrated in table 3.1.

Subjects	Gender (Male/Female)	Mean age (years±std)	Age not reported	WSIs
1,026	615/411	66.2±9.4	41	3,220

Table 3.1: *Subject demographics including the number of subjects, gender, mean age at diagnosis and standard deviation, subjects with no age reported, and the total number of WSIs for all the subject.*

The acquired data was converted from SVS files with the highest magnification to WSIs using MATLAB, which is illustrated in the second step in figure 3.2. To ensure homogeneity in the input images, only WSIs acquired with a magnification level of 40x were used in this project. Therefore, the magnification levels were extracted from the SVS files and data with other levels were excluded, as illustrated in the third step in figure 3.2. Furthermore, it was desired to classify the data based on the cancer stage, by which the cancer stage of the WSI has been extracted from the metadata. Based on this, WSIs which did not have a reported cancer stage were excluded, as illustrated in the fourth step in figure 3.2. Lastly, during the data selection, the quality of the data was ensured by performing a visual inspection of the converted WSIs. Throughout the visual inspection, it was found that some WSIs had ink stains on them, as illustrated in the fifth step in figure 3.2. Since the ink stains were not related to the tissue structures, the WSIs with ink stains were additionally excluded. Through the visual inspection, it was additionally discovered that a great variation occurred among the WSIs, with regards to staining intensity and amount of tissue. An example of a WSI with a great contrast is illustrated in figure 3.3a, an example of a WSI with less contrast is illustrated in figure 3.3b, and an example with multiple tissue examples in one WSI is illustrated in figure 3.3c.

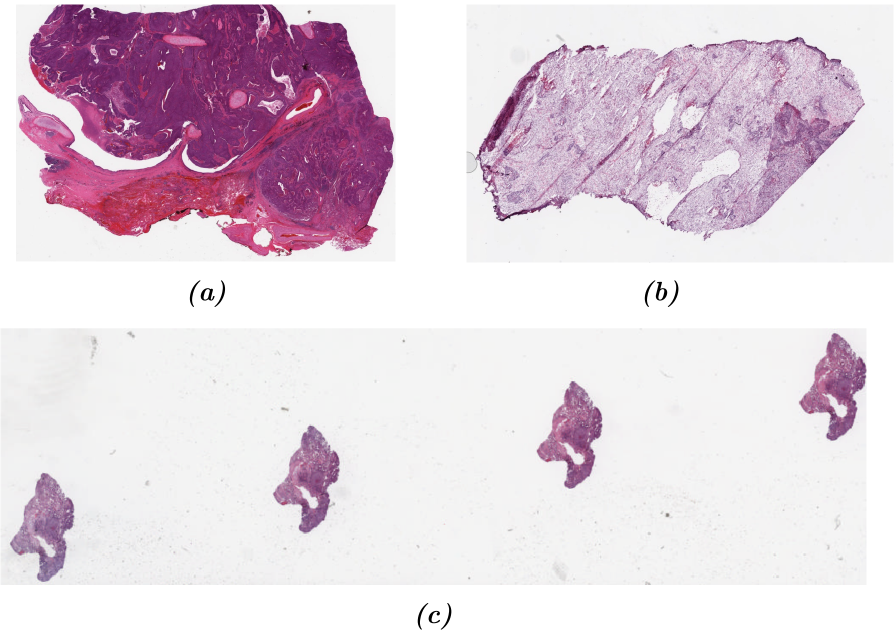


Figure 3.3: *Examples of a WSI with great contrast (a), a WSI with less contrast (b), and a WSI with multiple tissue samples(c).*

The number of WSIs was reduced from 3,220 WSIs which were available in TCGA to 549 after the exclusion of WSIs. The exclusion process is illustrated in figure 3.4. Initially, download errors occurred when the data was acquired from TCGA, resulting in a reduction by 1,530 WSIs. Additionally, as a result of the data selection, 1,117 of the acquired WSIs were excluded as these were not recorded with a magnification level of 40x. Finally, 24 WSIs were excluded due to two additional exclusion criteria which included missing information about cancer stage and ink stains on the WSIs. Due to the exclusion criteria, one WSI was excluded due to missing information about the cancer stage in the metadata, and 22 WSIs were excluded due to ink stains on the WSIs. Finally, one WSI was excluded due to both ink stains and missing information about the cancer stage.

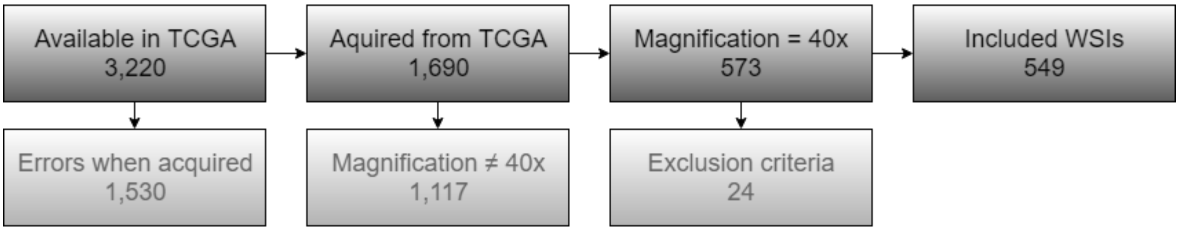


Figure 3.4: *Exclusion of WSIs described from the available data in TCGA to the data used in this project. The boxes in the first row illustrate the included WSIs, while the faded boxes in the second row illustrate the exclusion of WSIs and the reasons hereof.*

Ink stains on the WSIs were the main reason for excluding WSIs in the final exclusion step, illustrated in figure 3.4. Examples of three WSIs excluded due to ink stains are illustrated in figure 3.5. In these, different amounts and types of ink stains are illustrated. In figure a, multiple colors of ink stains are drawn on the tissue. In figure b, small ink stains appear on the tissue, assumable due to errors during the staining process. Finally, in figure c, ink stains which do not cover the tissue is illustrated. All the excluded WSIs due to ink stains, are illustrated in appendix B.

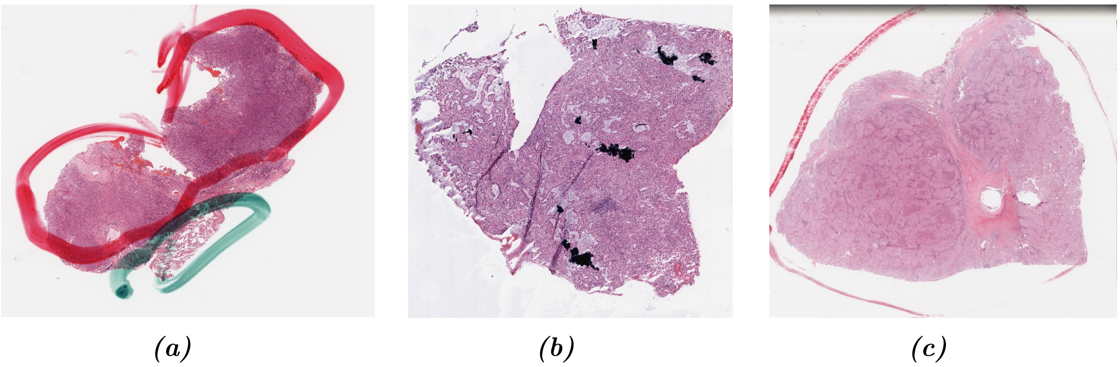


Figure 3.5: *Examples of WSIs with ink stains which were excluded. WSI with red and green ink stains on the tissue sample (a), WSI with black ink stains on the tissue sample (b), and red ink stains surrounding the tissue sample (c).*

The subject demographics of the remaining subjects after the exclusion process, are depicted in table 3.2.

Subjects	Gender (Male/Female)	Mean age (years±std)	Age not reported	WSIs
377	212/165	65.8±9.3	0	549

Table 3.2: *Subject demographics including the number of subjects, gender, mean age at diagnosis and standard deviation, subjects with no age reported, and the total number of WSIs for all the subject.*

The WSIs acquired from TCGA were NSCLC of the subtypes: SCC and adenocarcinoma. The remaining data consisted of a total of 549 WSIs, in which all four cancer stages (I-IV) were represented in both subtypes of NSCLC. The image data information for the acquired TCGA data are illustrated in table 3.3.

	Squamous cell carcinoma	Adenocarcinoma	Total
Subjects	173	213	383
WSIs	236	337	573
Stage I	116	197	313
Stage II	85	79	164
Stage III	31	46	77
Stage IV	4	13	17
Not reported	0	2	2

Table 3.3: *Information about the image data including the number of subjects, WSIs, and cases of the four different stages depicted for the two different subtypes of NSCLC and as the total amount.*

3.2 Data Preprocessing and Dataset Construction

The main obstacle when using CNNs is overfitting, which occurs when the model becomes too good at representing the data it has been trained on, as it will not perform as good on the test set, which represents data the trained model has not been exposed to. Therefore, the dataset used for a neural network is split into three datasets: Training, validation, and test set. This enables a validation set to be included to validate the model, by which it is evaluated based on multiple sets before exposed to the test set. Overfitting can occur when too few samples are presented to the network, by which it is not possible to make a generalized model. [Chollet, 2018]

Data preprocessing is an important step to make the data more manageable for the network and to ensure the quality of the data presented to the network [Chollet, 2018]. The preprocessing steps and the construction of datasets for this project are illustrated in figure 3.6.

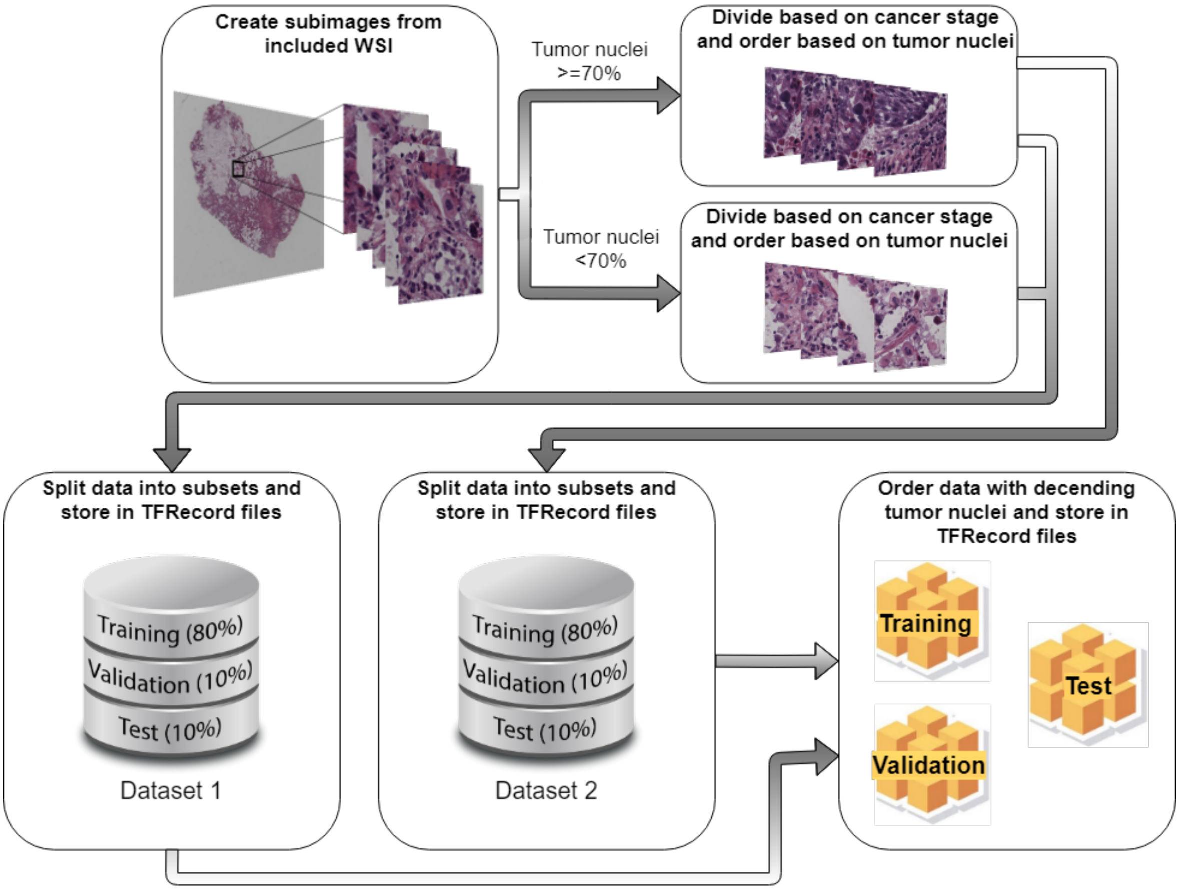

Create subimages from included WSI
Tumor nuclei >=70%
Tumor nuclei <70%
Divide based on cancer stage and order based on tumor nuclei
Divide based on cancer stage and order based on tumor nuclei
Split data into subsets and store in TFRecord files
Training (80%)
Validation (10%)
Test (10%)
Dataset 1
Split data into subsets and store in TFRecord files
Training (80%)
Validation (10%)
Test (10%)
Dataset 2
Order data with decending tumor nuclei and store in TFRecord files
Training
Test
Validation

Figure 3.6: *Initial data preprocessing, including creation of subimages, followed by construction of training, validation, and test datasets as TFRecords.*

The WSIs ranged in size of which the images had a width of 58,465±23,529 pixels and a height of 58,232±40,264 pixels when they were horizontally orientated. It was desired to extract the entire WSIs as TIFF files, but due to the high resolution, it was not possible to maintain this resolution if the WSIs were extracted in full size. Therefore, subimages with the size of 512×512 pixels were generated from the WSIs during preprocessing using MATLAB, as illustrated in the first step in figure 3.6. To ensure equally sized subimages, which were more manageable for the network, the edges were padded with pixels with a value of 255 if the size was smaller than 512×512. The generated images were split into subfolders for each WSI containing the corresponding subimages.

To ensure similarity between the datasets, it was essential that the WSIs with a great or limited representation of tumor nuclei were equally divided between the training, validation and test set. The annotations of the cancer stage were WSI-based, by which there was no guarantee that cancer was represented in all the created subimages. Furthermore, the subimages may even contain a larger amount of benign tissue than malignant tissue, by which an issue of noisy labeling occurred. However, it was more likely that subimages from a WSI with a high percentage of tumor nuclei also contained tumor nuclei within the subimages, by which the tumor nuclei percentage was retrieved from the metadata for each WSI. The tumor nuclei percentage was used to construct two different datasets, by which the potentially noisy labels could be addressed:

1. Training, validation, and test datasets containing all subjects, equally divided into the three datasets.
2. Training, validation, and test datasets containing subjects with WSIs with a tumor nuclei percentage of 70% or more, equally divided into the three datasets.

The two different datasets enabled training experiments with different concentrations of tumor nuclei. The choice of 70% as the limit, was due to a natural separation within the dataset. It was anticipated, that the use of the small dataset in training would result in a better fit of the model and a higher performance. However, a model trained on the large dataset might learn to recognize the lung cancer stage even if the concentration of tumor nuclei was low. Furthermore, some WSIs did not have a tumor nuclei percentage reported. These were only included in the large dataset as the tumor nuclei percentage was unknown and thereby was not guaranteed to be 70% or more. The process of evaluating the tumor nuclei percentage is illustrated in the link between the first and second steps in figure 3.6.

An equal distribution of the four cancer stages (I-IV) should be represented in the training, validation, and test datasets, in both the small and large dataset, to ensure that the model learned these representations and to validate and test the classifications hereof. A subject-based division was chosen to ensure that the model would not validate or test on WSIs from the same subject it had been trained on. Thereby, the data was divided subject-based into a training, validation, and a test dataset, as illustrated in the second step in figure 3.6. An iterative training framework was used in which the network received an input with a high concentration of tumor nuclei in the initial epochs and a lower concentration of tumor nuclei in the last epochs. The tumor nuclei percentage in the WSIs ranged from 0% to 100% and some percentages of tumor nuclei in the WSIs were not reported. However, all WSIs were labeled as either SCC or adenocarcinoma by which some tumor nuclei must be represented in the WSI to make this labeling possible. Therefore, all the WSI were included in the large dataset to get as large a dataset despite the great variance of tumor nuclei concentration among the WSIs. To address the variance in concentration of tumor nuclei, an iterative framework was enabled by ordering the data within each of training, validation and test dataset respectively, starting with the WSIs with the highest tumor nuclei percentage and ending the WSIs with the lowest tumor nuclei percentage followed by the ones not reported in the large dataset. Likewise, in the small dataset the data was ordered within each of the training, validation and test set. This made it possible to load the subimages from the WSIs from a higher to a lower concentration of tumor nuclei within each dataset. It was anticipated that the WSIs with a higher tumor nuclei percentage were more reliable and contained less noise, by which the network would learn there less noisy representations first. The ordering of the data with regards to tumor nuclei percentages is illustrated in the second step in figure 3.6.

The subject-based division of subimages was performed randomly into either the training, validation, or test set with a similar distribution of the four cancer stages (I-IV) and tumor nuclei percentage. The subjects were divided with approximately 80% in training, 10% in validation, and 10% in test, as illustrated in the third step in figure 3.6. The number of WSIs in the three subsets is depicted in table 3.4 and table 3.5. These depict that stage IV is poorly represented in both datasets, which raise concern about the classification of this stage. Included in the datasets were a subject ID, the subimages corresponding to the subject, and the cancer stage of the WSI which was used as the label for all the corresponding subimages.

The Large Dataset				
	Training	Validation	Test	Total
Stage I	229	38	35	302(55.0%)
Stage II	123	18	16	157(28.6%)
Stage III	60	8	6	74(13.5%)
Stage IV	8	3	5	16(2.9%)
Total	420(76.5%)	67(12.2%)	62(11.3%)	

Table 3.4: *Distribution of the 549 WSIs in the large dataset with regards to stage and the Number of WSIs included in the training, validation, and test set respectively.*

The Small Dataset				
	Training	Validation	Test	Total
Stage I	45	5	9	59(55.1%)
Stage II	22	3	3	28(26.2%)
Stage III	12	2	3	17(15.9%)
Stage IV	1	1	1	3 (2.8%)
Total	80(74.8%)	11(10.3%)	16(15.0%)	

Table 3.5: *Distribution of the 107 WSIs in the small dataset with regards to stage and the Number of WSIs included in the training, validation, and test set respectively.*

After the data had been ordered, each dataset was stored separately as TFRecord files, that is Tensorflow's own binary storage format, which is illustrated in the fourth step in figure 3.6. These files included information of subject ID, a path to one or more associated WSIs, cancer subtype, and the cancer stage.

3.3 Patch Extraction

CNNs cannot be applied to very high-resolution images such as gigapixel WSIs, due to high computational cost. Furthermore, it has several drawbacks to apply the CNN for WSI classification, as extensive image downsampling would be required, by which most of the discriminative details such as the shape, size, and texture of the cells and nuclei would be lost. Therefore, it can be preferable to divide the WSI into smaller patches, in which the discriminative details are preserved. [Hou et al.] Thereby, patches were extracted from the subimages, which were created in the preprocessing step, described in section 3.2.

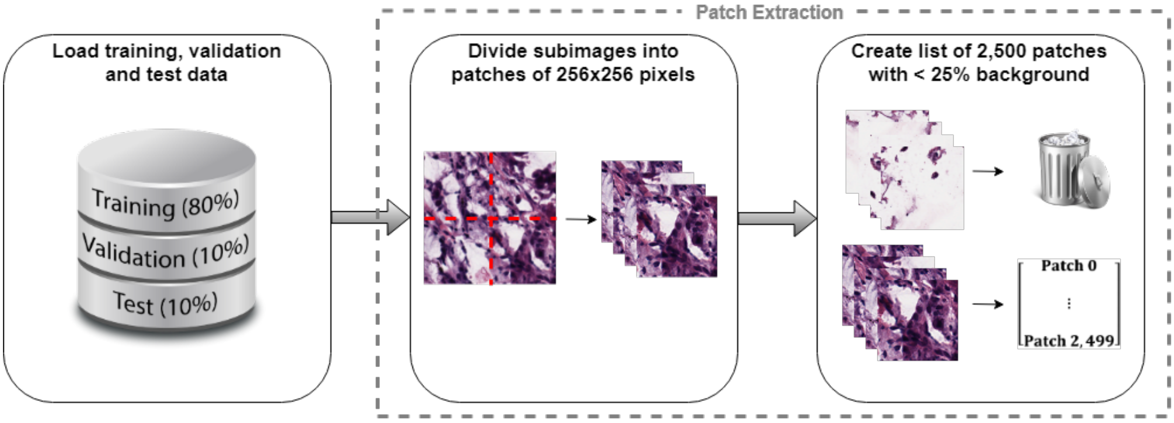
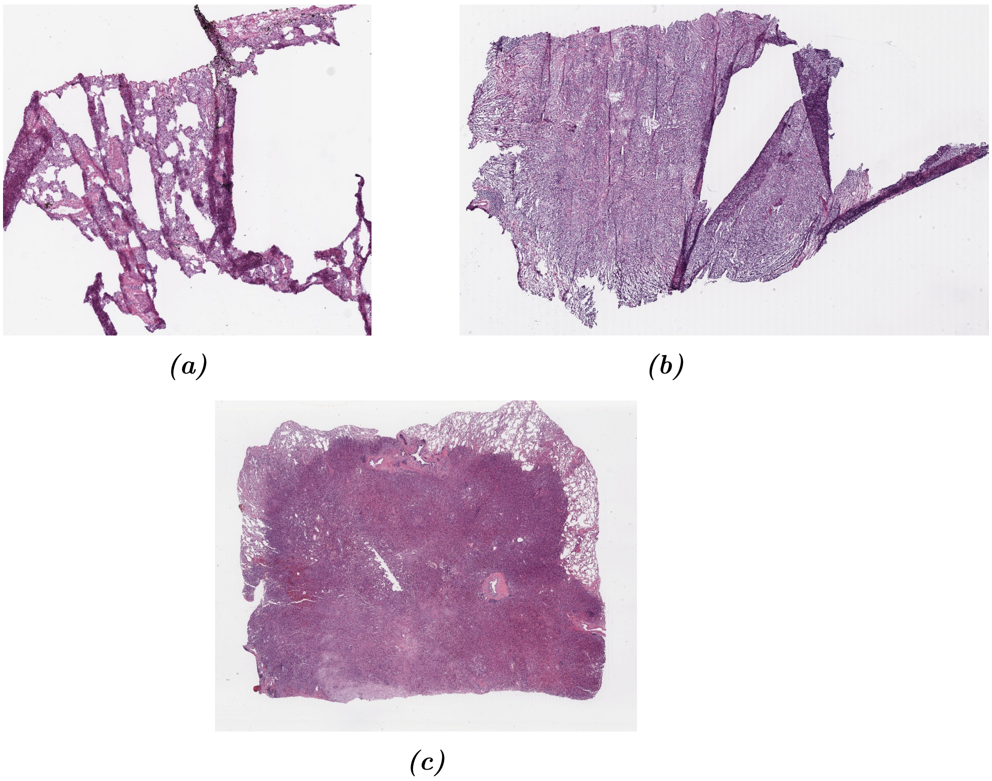

Patch Extraction
Load training, validation and test data
Training (80%)
Validation (10%)
Test (10%)
Divide subimages into patches of 256x256 pixels
Create list of 2,500 patches with < 25% background
Patch 0
⋮
Patch 2, 499

Figure 3.7: *Flowchart of the patch extraction before applying the data to the network.*

The labels and the paths to the subimages used for training, validation, or test were acquired from the datasets, as illustrated in the first step in figure 3.7. The image-data consisted of tiff files of subimages with a size of 512×512 pixels from the WSIs. These were divided into patches of 256×256 pixels, as illustrated in the second step in figure 3.7. Each of these patches were assigned the label of the cancer stage represented in the WSI. The labels extracted from the metadata were associated with the WSIs and not the extracted patches. Therefore, it was unknown whether each of the patches contained the type of malignant tissue of which it was labeled or if the patch only consisted of benign tissue. To make it more possible for the patch to include malignant tissue, a considerable amount of tissue should be represented in the patch before it was included. Therefore, patches which contained more than 25% white background, were excluded, as illustrated in the third step in figure 3.7.

It was investigated how many patches remained from each WSI after this exclusion. The result hereof showed a minimum of 2,139 patches, a maximum of 151,382 patches, and an average of 36,058±31,982 patches with a patch size of 256×256 pixels remaining from the individual WSIs. Examples of a WSI with minimum, maximum, and an average number of remaining patches after exclusion of patches with more than 25% background is illustrated in figure 3.8. The initial sizes of these WSIs were 13,944×12,605 pixels corresponding to 2,682 patches for the WSI with the least extracted patches, 62,832×39,190 corresponding to 37,573 patches for the WSI with an average extracted patches, and 92,802×118,740 pixels corresponding to 168,141 patches for the WSI with the most extracted patches.



(a) *(b)*

(c)

Figure 3.8: *WSIs from which the minimum (a), average (b), and maximum (c) number of patches remained after exclusions of patches with >25% background.*

3.4 Training and Validation of the CNN

Convolutional neural networks (CNNs) are widely used to classify the structures and patterns within medical images. This is due to the ability to handle large datasets with complex patterns. [Chollet, 2018; Khosravi et al., 2017; Ngo et al., 2016] The CNN used in this project was constructed based on the AlexNet. This has been a widely used network which has won the ILSVRC competition in 2012 [Yu et al., 2018]. Furthermore, it has shown good performance when classifying histopathological tissue [Deniz et al., 2018].

3.4.1 Construction of a Convolutional Neural Network

Neural networks are models, which transform the input data to an expected output. An example of a simple neural network architecture is illustrated in figure 3.9. A neural network consists of an input layer followed by one or multiple hidden layers, which give a result in the output layer. The layers consist of nodes, which are illustrated by the circles within the layers in figure 3.9, that are connected between the layers and updated through weights. Through deep learning, the neural network learns the data representation through multiple steps of successive layers with increasing meaningful representation, which is guided through a feedback process. Tens or hundreds of successive layers can be involved in the neural network model, of which the parameters of the layers are learned automatically from exposure to the training data. [Chollet, 2018]

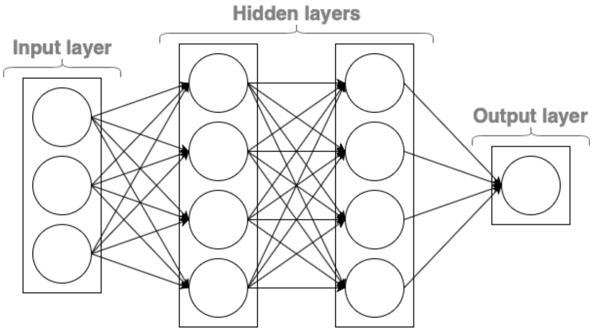


Figure 3.9: *General structure of a neural network with an input layer, two hidden layers, and an output layer. The neurons in each layer are illustrated with circles, which are influenced by the weight of the link between neurons from the different layers.*

Neural networks can be used for purposes such as image segmentation and image classification. For this, images are used as input and through image segmentation, the output is a pixel-drawn mask, whereas the output from image classification is a class label. Different types of neural networks can be applied to encode assumptions of the data depending on the purpose. Hereof, densely connected networks detect relationships between two input features using stacks of dense layer, also called fully connected layers. In fully connected layers, all nodes from the previous layer are connected to all nodes in the current layer. Thereby, fully connected layers learn from previous layers, but fully connected layers would have to learn the same pattern again if it occurred in a different location in an image. CNNs have been used for visual biomedical tasks in which they are typically used for classification tasks.

As illustrated in figure 3.10, the structure of a CNN consists of an input layer, one or more hidden layers, and an output layer. In the first stages of a CNN, convolutional and pooling layers are used. Information is processed in multiple steps through these successive layers, which work as filters. The neurons in each layer are influenced by a given weight and bias, and takes one or more tensors as input and give one or more tensors as output. [Chollet, 2018]

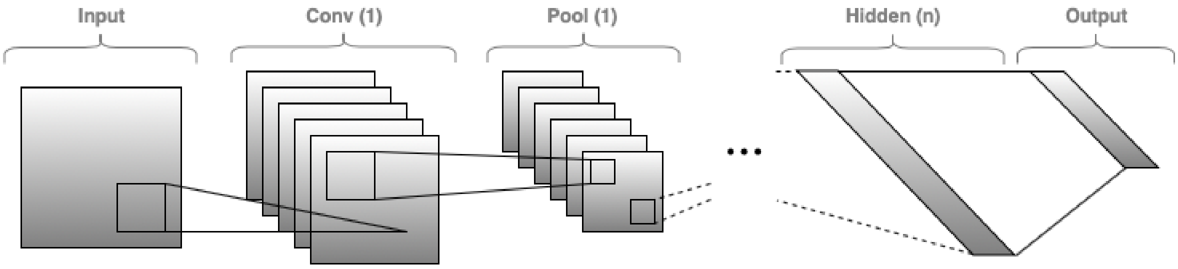


Figure 3.10: *General structure of a CNN, including an input layer, n number of hidden layers including convolutional and pooling layers, and an output layer.*

After detecting the local conjunctions of features in the convolutional layers, pooling layers merge semantically similar features into one. Typically, this is performed by computing the maximum value of a local patch of units in one or a few feature maps, which is called max pooling. Through pooling, the dimension of the feature map is reduced, which causes invariance to small shifts and distortions. In the final layer of a CNN, a softmax layer is often used, which is an activation function that produces an output probability distribution. [Lecun et al., 2015]

The geometrical transformation in the convolutional layer is a discrete convolution, hence the name CNN. [Lecun et al., 2015] In a convolution, a filter is slid over the input feature map with a specified stride. This performs an element-wise matrix multiplication by which the input is transformed to an output feature map. The spatial locations in the output feature map correspond to the locations in the input feature map. [Lecun et al., 2015] An example of a convolution is illustrated in figure 3.11 with a filter size of 3×3. A CNN applies the same geometrical transformation to the different images in the input, by which spatial local patterns are used, which leads to a translation invariant representation. This enables CNNs to process images efficiently using fewer training samples to learn generalized representations, as the learned patterns can be recognized anywhere in the image. [Chollet, 2018]

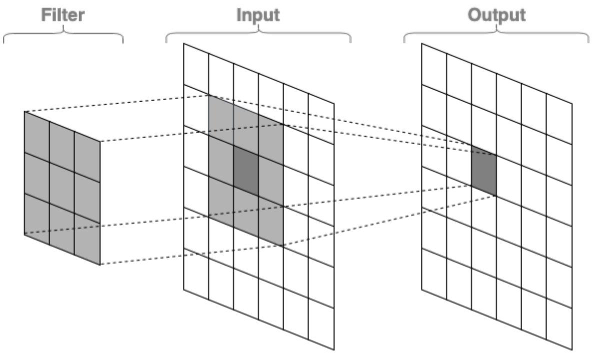


Figure 3.11: *Convolution of an input using a 3×3 filter and the corresponding output.*

The units in a convolutional layer are organized as three dimensional (3D) tensors, called feature maps, which have two spatial axes and one depth axis. In the feature maps, each unit is connected to local patches within the feature maps from the previous layers through a filter bank, which is a set of weights. The weighted sum is then passed through an activation function such as a rectified linear unit (ReLU), which is a non-linear activation function that zeroes out negative values. [Lecun et al., 2015] This enables complex relationships in the data to be learned, as the data achieves non-linearity. The same transformation, as illustrated in figure 3.11, is applied to all patches from the input feature map in the convolution operation. This produces an output feature map consisting of 3D tensors with an arbitrary depth. The arbitrary depth is a parameter of the layers, in terms of filters which encode different aspects of the input data. [Chollet, 2018]

There is a hierarchy of the patterns learned in a CNN, in which the first convolutional layer learns small local patterns such as edges. The second layer learns the larger patterns made from the first convolutional layer, and so on. Thereby, the network increasingly learns more complex and abstract representations. [Chollet, 2018] During the training of a CNN, the learning happens as the weights and biases are optimized several times to find a set of values which parameterize the transformation of information in which the network maps the input to the target. This optimization is performed based on a loss function, which calculates a distance score between the ground truth label and the predicted label of the network. The calculated score is then used in a feedback process which adjusts the value of the weights to minimize the loss. The weights of the layers are often initiated as random values, by which the classification error is expected to be high in the beginning of a training session. This adjustment is performed using the backpropagation algorithm, which applies a chain rule when computing the values of the weights for the neural network. The chain rule propagates backward from the final loss to the weights of the first layer to compute the contribution of the individual parameters to the final loss score. Thereby, the algorithm propagates classification errors backward from the final loss to compute the contribution of the individual parameters to the final loss score. An optimization function reduces the loss over several iterations over batches from the training data by optimizing based on the loss. Each of these optimizations over the training data is called an epoch. An often used optimization function is gradient descent. [Chollet, 2018] Smaller batches of data can be used to make the training set more general and computational efficient [Ioffe and Szegedy, 2015]. Thereby, a gradient over the entire dataset is estimated by the gradient of the loss of a smaller batch [Ioffe and Szegedy, 2015].

3.4.2 Architecture of the AlexNet

The AlexNet architecture, which is used in this project, consisted of five convolutional layers and three fully connected layers, as illustrated in figure 3.12.

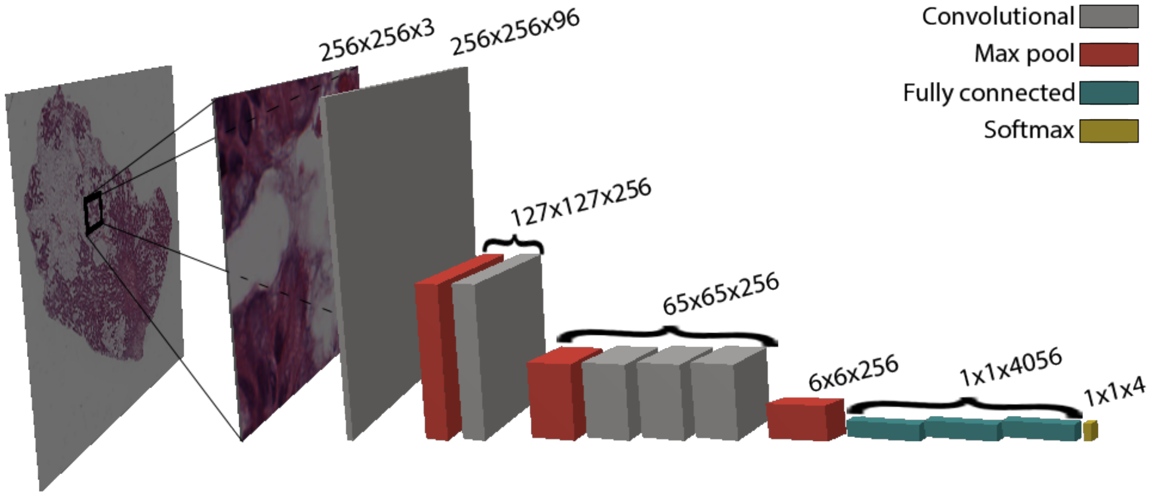


Figure 3.12: *The architecture of the implemented CNN with a patch from a WSI taken as input and probability of the four stages given an output.*

As illustrated in figure 3.12, a 256×256 pixel RGB colored patch from a WSI was given as input along with a label of the represented cancer stage. In the first layer, an 11×11 convolution with stride 4×4 and ReLU was used. To generalize the model, local response normalization has been utilized. Local response normalization implements a contrast enhancement of feature maps, by enhancing local maxima and reducing flat responses. [Krizhevsky et al., 2012] The first convolution was followed by a 3×3 max pooling operation with stride 3×3 and local response normalization. In the second layer, a 5×5 convolution and ReLU was used with two groups, followed by a 3×3 max pooling operation with stride 3×3 and local response normalization. In the third and fourth layer, a 3×3 convolution and ReLU was used with one and two groups respectively. The fifth layer consisted of a 3×3 convolution with two groups and ReLU, followed by a 3×3 max pooling operation.

The complexity of the model was reduced in order to prevent overfitting, by which dropout was used. With dropout, a random number of features were set to zero during the training session, by which they were dropped out of the session, thereby coincidence correlations should be broken. [Chollet, 2018].
In the sixth and seventh layer, a flattening operation was performed which reshaped the output to 6×6×256. This was followed by a fully connection with an output of 4,096 and ReLU, followed by dropout with a dropout rate of 0.5, which indicated the fraction of the features that are zeroed out [Chollet, 2018].

A logits operation was used in the loss function, as the input to the CNN was an RGB colored patch from a WSI, and the input to the last layer thereby did not have a sum of inputs equal to 1. This logits function operated on the unscaled output from earlier layers, which enabled the function to take an input, that did not have a sum of inputs equal to 1. To normalize the input, this operation was followed by a softmax activation function in the final layer, which squished the input to have a sum of inputs equal to 1. The outputs of the softmax function could be interpreted as probabilities for each patch. A cross entropy loss function was applied to the result of the softmax operation. The equation of the cross entropy is stated in equation (3.1) in which t_i is ground truth and s_i is the CNN score for each class i in C.

$$CE = -\sum_{i}^{C} t_i \log(s_i) \tag{3.1}$$

Cross entropy is a measure of the uncertainty of a sample belonging to a given distribution based on a probability value between 0 and 1. The cross entropy loss function is logarithmic. Therefore, a prediction which differs a lot from the ground truth will result in a much higher loss value, and as the prediction approaches the ground truth, the loss value will decrease rapidly. If the prediction was equal to the ground truth, the cross entropy loss would be equal to 0.

The eighth and last layer was fully connected, followed by a softmax cross entropy function with logits over the final feature map. Thereby, the patch was classified into four mutually exclusive classes as the output: Stage I, II, III, and IV. Hereof, the patch was assigned four probabilities, one for belonging to each of the four stages.

3.4.3 Transfer learning

It can be beneficial when training a new model to use knowledge about learned features from a previous model by which transfer learning has been used in this project. Furthermore, transfer learning might have made the training session faster and optimized the training to get a better performance. Ideally, the initial performance should be better than when using initial random weights, the model should learn faster and achieve higher performance. However, it is not always obvious that a better performance is achieved with transfer learning before the trained model is tested. [Soria et al., 2009] The model used for the transfer learning should be very general enabling the spatial hierarchy of the features to act as a generic model and hence its features can be useful in many computer vision problems [Chollet, 2018]. Therefore, a model trained on the ImageNet data with 1,000 classes was used for the transfer learning as this was trained on a large dataset with a great number of labels by which it should be a generic model.

A pre-trained network can be used for either feature extractions or fine-tuning, of which feature extraction uses representations from a previous network to extract features of interest from new samples and then train a new classifier from scratch. However, it was desired to slightly adjust the representations of the model to make it more relevant for the current problem by which it was chosen to use fine-tuning. [Chollet, 2018] With this method, it was possible to only retrain the last two layers in the network, as the earlier layers have learned more robust and generic features, such as visual edges, colors, and textures, and the last layers then learned the more specialized features, such as cell or cell nuclei type, which is needed for classification of lung cancer stages.

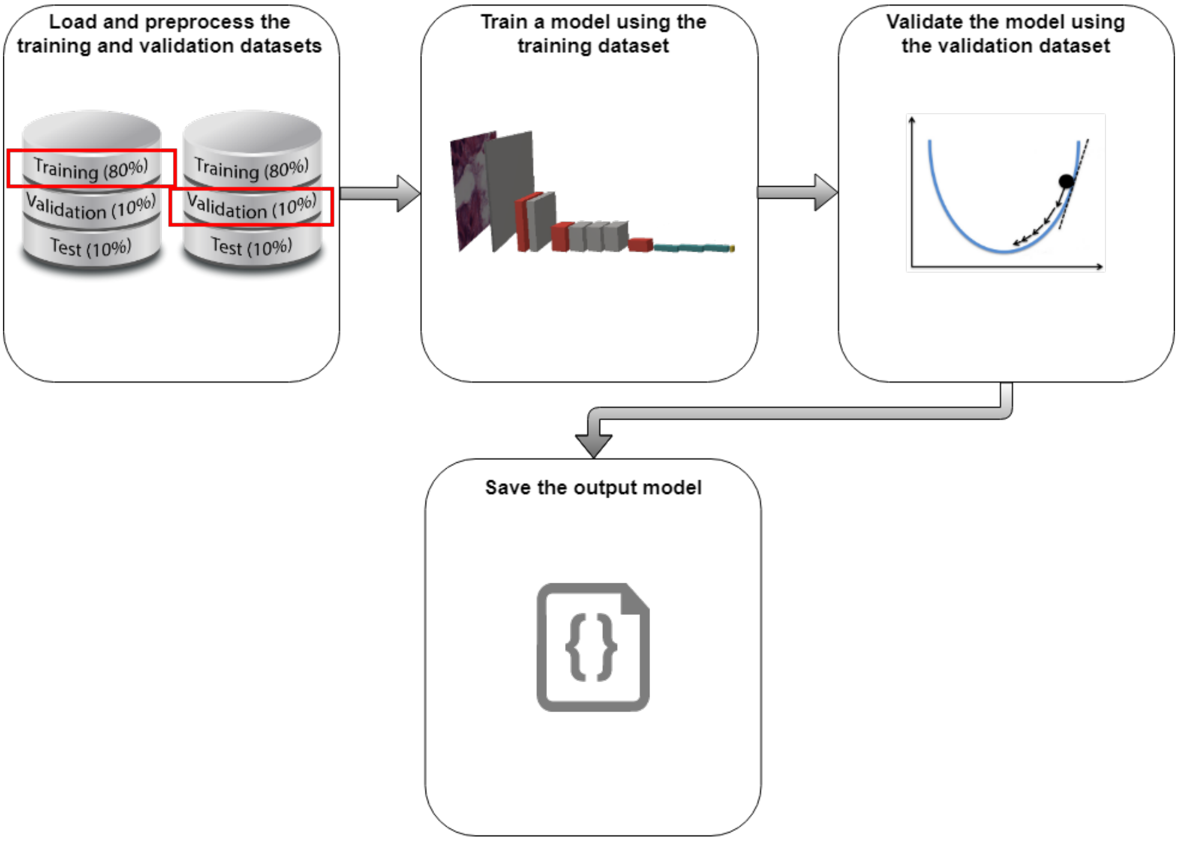

Load and preprocess the training and validation datasets
Training (80%)
Validation (10%)
Test (10%)
Training (80%)
Validation (10%)
Test (10%)
Train a model using the training dataset
Validate the model using the validation dataset
Save the output model
{}

Figure 3.13: *Flowchart of the training and validation.*

After the patch extraction, the WSIs from the training and validation dataset were given as input to the network, as illustrated in the first step in figure 3.13. These were used as input to fine-tune a model of the AlexNet by training the last two layers, as illustrated in the second step in figure 3.13. Thereby, the more abstract representations of the data were slightly adjusted to make the model relevant for the current problem. Thereby, a new classifier was trained based on parameters inherited from the prior model.

As illustrated in the third step in figure 3.13, the trained model was validated using gradient descent and a cross entropy loss function, and the weights of the fine-tuned layers were updated. The initial weights and biases for the CNN was a replication of the initial model described in Krizhevsky et al. [2012], before they trained using data augmentation and using initialization of non-zero biases. Through fine-tuning, the weights of the last two layers of the network were then changed during the optimization using gradient descent. This optimization was dependent on the learning rate of 0.01 and the cross entropy loss function.

Training Experiments

To address the concern about noisy labels, two different training experiments were conducted for each of the two datasets, resulting in four different experiments:

1. Training using 4,000 patches corresponding to 2,000 patches from 2 WSIs per epoch using the large dataset.
2. Training using 4,500 patches corresponding to 250 patches from 18 WSIs per epoch using the large dataset.

3. Training using 4,000 patches corresponding to 2,000 patches from 2 WSIs per epoch using the small dataset.
4. Training using 4,500 patches corresponding to 250 patches from 18 WSIs per epoch using the small dataset.

Experiment 1 and 3 constituted training using the largest number of patches from each WSI to enable a more general representation of the WSI. The number of patches was thereby based on the minimum number of patches remaining after exclusion of patches with more than 25% background. Hereof, a common denominator of 2,000 patches was chosen. This entailed a smaller number of WSIs which could be processed within an epoch due to a TensorFlow limitation of the input capacity. This restricted the CNN in these experiments to get an input of two WSIs per epoch. The WSIs used, were changed in between the different epochs.

Experiment 2 and 4 constituted training using a smaller number of patches from each WSI, but a larger number of WSIs per training. This enabled more variety regarding staining intensity and representation of nuclei in the inputs to each epoch, which then facilitated learning of more variety in appearance of the features. The number of patches was chosen to be 250 patches, which enabled the number of WSIs to be increased to 18 WSIs per epoch. Likewise experiment 1 and 3, the WSIs used in these experiments were changed in between the different epochs.

The training and validation were repeated for 50 epochs in all four experiments, in which the patches for the training and validation batches were randomly selected from each WSI. After the 50 epochs, the trained model was saved as the output of the training and validation experiment, which is illustrated in the fourth step in figure 3.13.

3.5 Test of the Trained Model

Figure 3.14 illustrates the test process in which the performance of the trained model was tested on the test dataset.

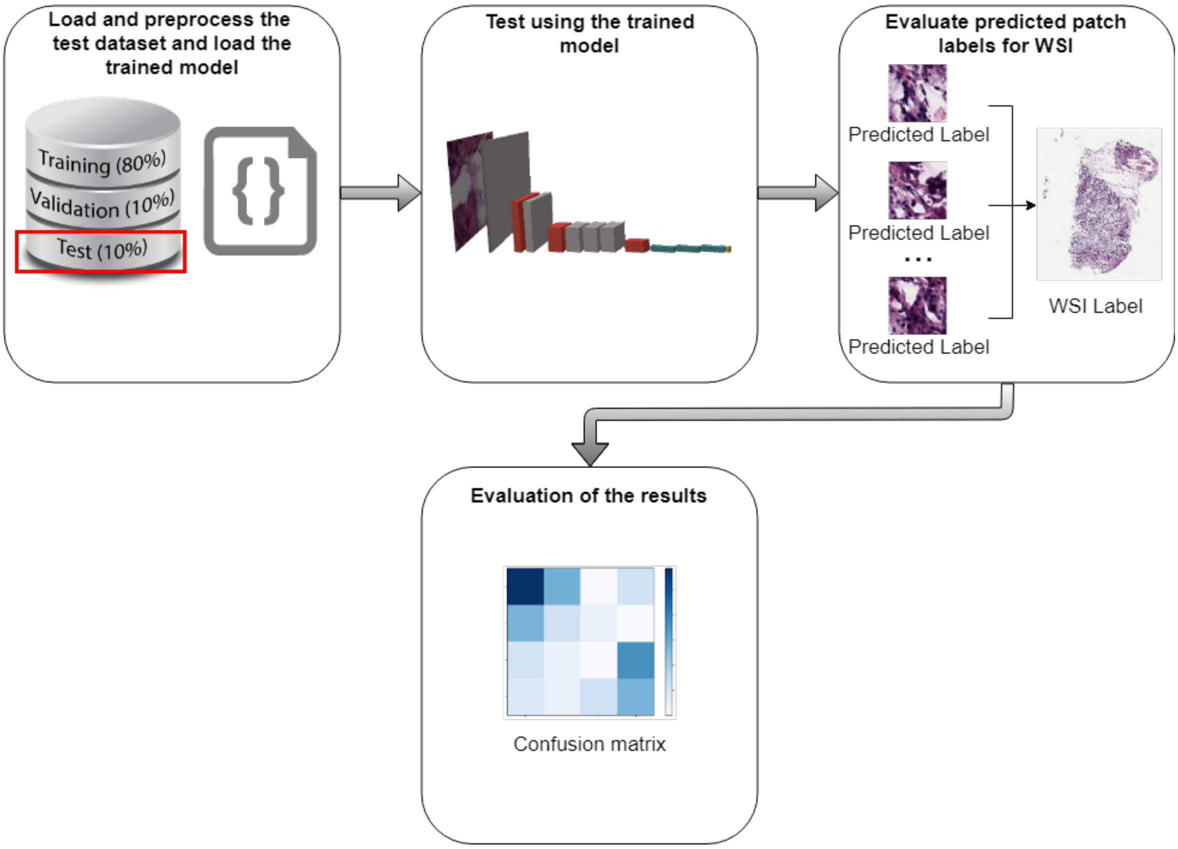

Load and preprocess the test dataset and load the trained model
Training (80%)
Validation (10%)
Test (10%)
{}
Test using the trained model
Evaluate predicted patch labels for WSI
Predicted Label
Predicted Label
...
Predicted Label
WSI Label
Evaluation of the results
Confusion matrix

Figure 3.14: *Flowchart of the test of the trained model in which the third step has been conducted in three different ways.*

Patches were extracted from the test dataset and preprocessed as described in section 3.3. These preprocessed patches were then fed to the network, as illustrated in the second step in figure 3.14, and an image classification was performed using the trained model. As the label extracted from the metadata was given for the WSI, it was desired to classify the data WSI-based. Therefore, the probabilities of the predicted class labels for the patches was evaluated for the entire WSI as illustrated in the third step in figure 3.14. Three different WSI-classification approaches to label the WSI was investigated based on the predicted patch labels:

1. Extract 10% of the predicted labels with the highest probability and provide the WSI with the most frequent label
2. Extract the predicted labels with a probability above a threshold of 0.3 and provide the WSI with the most frequent label
3. Include all predicted labels and provide the WSI with the most frequent label

In the first WSI-classification approach, it was investigated whether it could be beneficial only to include the predicted labels for the patches with the highest accuracy, as these should be the most reliable. Thus, the labels assigned to the WSIs could be more comparable when a fixed number of patches were included, by which all the final labels for the WSIs were estimated based on the same amount of predicted patch labels. Therefore, this WSI-classification approach was conducted with 10% of the predicted labels from the patches with the highest probability. Based on these, a label was given to the WSI-based on the most frequent predicted patch label.

In the second WSI-classification approach, it was investigated whether it could be beneficial to use a threshold for the accuracy of the included predicted patch labels. Thereby, the number of predicted patch labels used to assign a label for the WSI did not matter, but it was ensured that all the included patches had an accuracy which was considered reliable classifications of the predicted patch labels. Thereby, this WSI-classification approach was conducted with the predicted labels of the patches which were classified with a probability above the threshold of 0.3. Based on these, a label was assigned to the WSI-based on the most frequent predicted patch label.

In the third WSI-classification approach, it was investigated whether it could be beneficial to include all the predicted labels for the patches as this gave a more general representation of the WSI, which overall should be predominant of the true label of the cancer stage. Thereby, this WSI-classification approach was conducted using all the predicted patch labels. Based on these, a label was assigned to the WSI-based on the most frequent predicted patch label regardless of the accuracy.

To evaluate the overall performance of the trained model, the accuracy of the patch-based predictions performed by the test were evaluated to obtain a general indication of how well the test performs on the individual patches. Furthermore, performance matrices were produced to evaluate the performance of the three WSI-classification approaches along with the individual patch-classifications, as illustrated in the fourth step in figure 3.14.

Results 4

This chapter presents the results from the four different training and validation experiments, and the results from the tests of these. Furthermore, the observations in the results are described.

4.1 Training and Validation Results

During training and validation, four different experiments were conducted:

1. Training using 4,000 patches corresponding to 2,000 patches from 2 WSIs per epoch using the large dataset.
2. Training using 4,500 patches corresponding to 250 patches from 18 WSIs per epoch using the large dataset.
3. Training using 4,000 patches corresponding to 2,000 patches from 2 WSIs per epoch using the small dataset.
4. Training using 4,500 patches corresponding to 250 patches from 18 WSIs per epoch using the small dataset.

The four experiments were performed for 50 epochs during each training and validation, with different batch sizes of randomly selected patches. The models were updated based on the cross entropy loss of the predicted patch labels from the validation set.

4.1.1 Experiment 1

In the first experiment, training and validation were performed for 50 epochs and with a batch size of 4,000 patches corresponding to 2,000 patches from 2 WSIs per epoch. These WSIs were from the large dataset, which included all the data after the data selection.
The cross entropy loss for the training and validation is illustrated in figure 4.1 for each epoch. The graph of the cross entropy loss indicated a convergence in the model fit, as the curve decreased with a steep slope during the initial epochs. Furthermore, a great deviation in the initial epochs was observed between the training and validation loss.

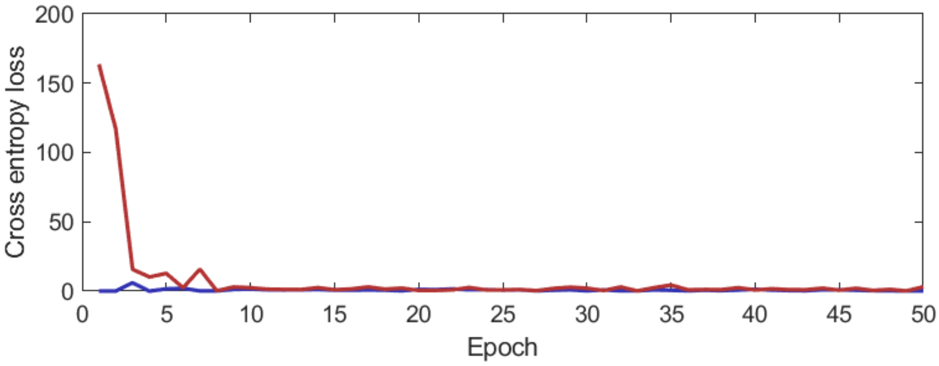


Figure 4.1: *Cross entropy loss for each epoch during the first experiment. The blue curve indicates the training loss and the red curve indicates the validation loss.*

Due to the great deviation between the loss in the initial epochs for training and validation respectively, an additional graph was constructed for the loss between zero and ten, as illustrated in figure 4.2. After approximately four epochs for the training loss and eight epochs for the validation loss, the curves approximately converged. The loss curves oscillated around a mean loss of approximately 1.0 for the training epochs, and 1.5 for the validation epochs, which indicated that the validation generally had a higher loss than the training during the epochs. Furthermore, the oscillations of the validation loss were larger than the oscillations for the training loss.

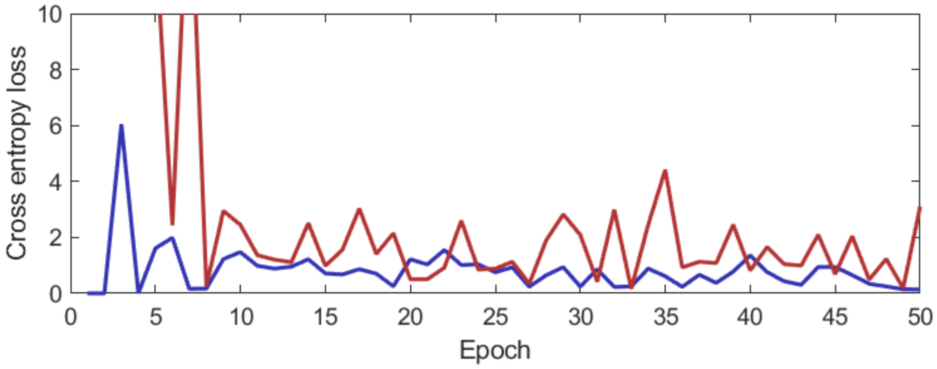
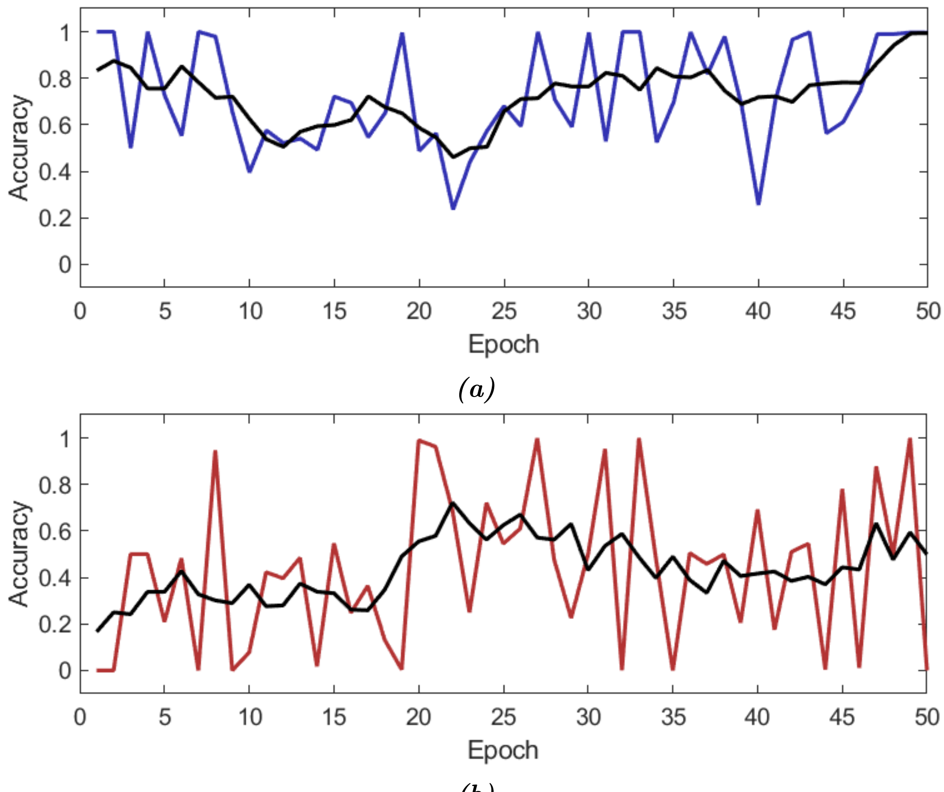


Figure 4.2: *Cross entropy loss for each epoch during the first experiment in which a limit for the y-axis was set between zero and ten. The blue curve indicates the training loss and the red curve indicates the validation loss.*

The accuracies of the training and validation epochs during the first experiment is illustrated in figure 4.3. Large deviations in the accuracy of both the training and validation were present between epochs. For the training epochs, the minimum accuracy was 0.2, mean accuracy was 0.7 and maximum accuracy was 1.0. For the validation epochs, the minimum accuracy was 0.0, mean accuracy was 0.4 and maximum accuracy was 1.0. During the training and validation, the accuracies had large smooth oscillations, when investigating the moving average, as illustrated in figure 4.3. These indicated that the accuracies during the training epochs slightly decreased during the first 25 epochs after which a slight increase was observed until the 50th epoch, as illustrated in figure 4.3a. On the contrary, the accuracies during the validation epochs increased until the 25th epoch after which the accuracies decreased until the 50th epoch, as illustrated in figure 4.3b. Furthermore, it was observed that the moving average curve of the accuracies in the training generally was higher than the moving average curve of the accuracies in the validation.



(a)

(b)

Figure 4.3: *Accuracy for each epoch during the training (a) and validation (b) from experiment 1. The blue curve illustrates the accuracy during training, the red curve illustrates the accuracy during validation, while the black curves illustrate a five-point moving average.*

4.1.2 Experiment 2

In the second experiment, training and validation were performed for 50 epochs and with a batch size of 4,500 patches corresponding to 250 patches from 18 WSIs per epoch. Likewise the first experiment, the WSIs used, were from the large dataset, which included all the data after the data selection.

The cross entropy loss for the training and validation is illustrated in figure 4.4, which, likewise the first experiment, indicated a convergence in the model fit.

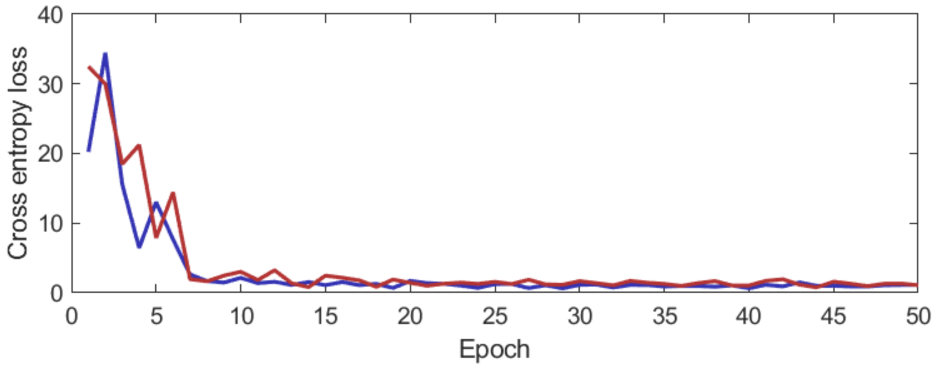


Figure 4.4: *Cross entropy loss for each epoch during the second experiment. The blue curve indicates the training loss and the red curve indicates the validation loss.*

Due to the great deviation between the loss between the initial epochs and the following epochs, an additional graph was constructed for the loss between zero and ten, as illustrated in figure 4.5. The losses resembled those from the first experiment with regards to convergence of the curves and oscillation size after the flatten tendency. After approximately seven epochs, both curves had converged, which approximately was the same number of epochs as in the first experiment for the validation losses, but later for the training losses. After this flattening, the curves oscillated around a mean loss of approximately 1.2 for the training epochs, and 1.5 for the validation epochs, which indicated that the training generally had a slightly higher loss than the validation.

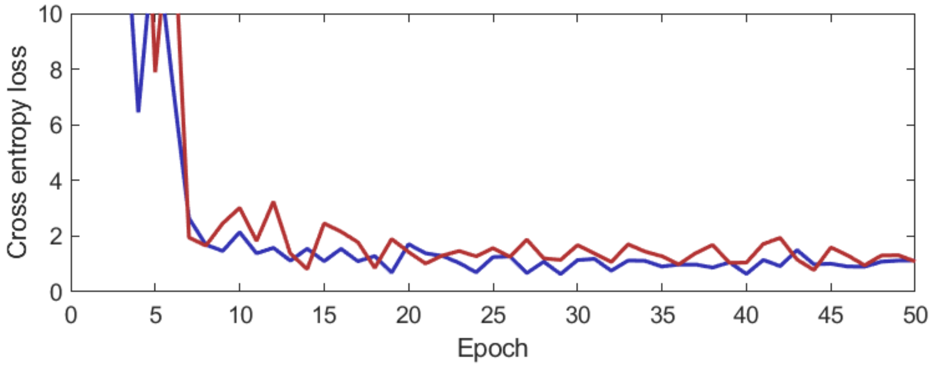
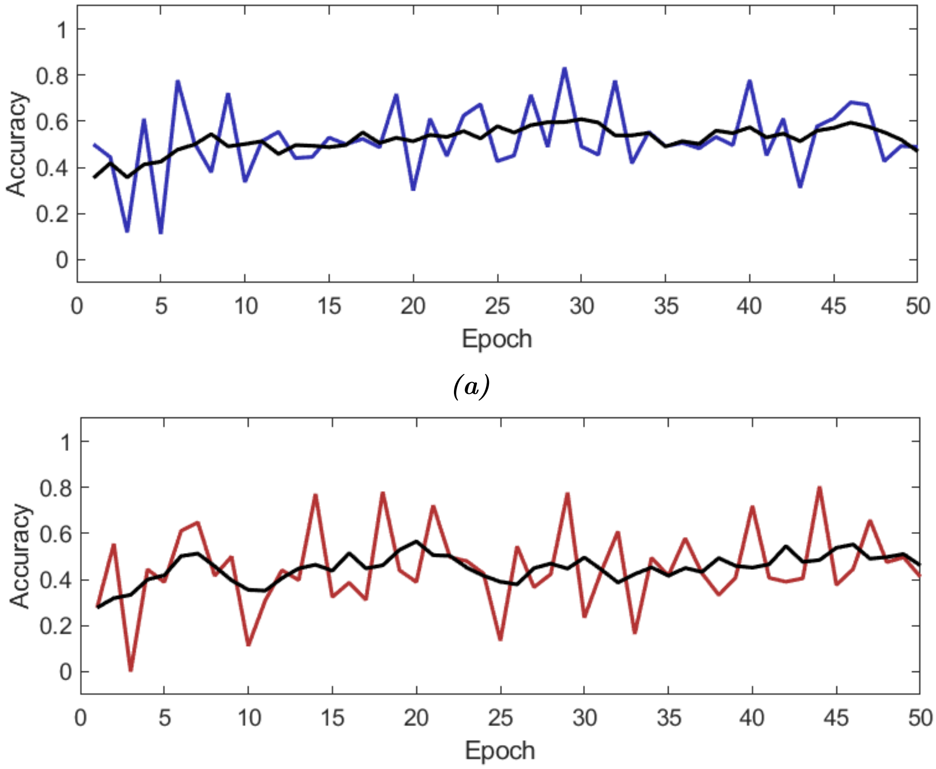


Figure 4.5: *Cross entropy loss for each epoch during the second experiment in which a limit for the y-axis was set between zero and ten. The blue curve indicates the training loss and the red curve indicates the validation loss.*

The accuracies of the training and validation during the second experiment is illustrated in figure 4.6. Large deviations in the accuracy of both the training and validation were present between epochs. However, the oscillations caused by these deviations were smaller in the second experiment than the oscillations observed in the first experiment. For the training epochs, the minimum accuracy was 0.1, mean accuracy was 0.5 and maximum accuracy was 0.8. For the validation epochs, the minimum accuracy was 0.0, mean accuracy was 0.5 and maximum accuracy was 0.8. This indicated that the accuracies of the second experiment in general was lower than those in the first experiment. A general trend of a slightly increasing accuries of the epochs was observed in the moving average of the accuracy during the first 25 epochs of the training. After this, no increasing or decreasing tendency was observed, as illustrated in figure 4.6a. Furthermore, no general tendency was observed during the validation epochs, as illustrated in figure 4.6b. It was observed that the accuracies of the training and validation was approximately at the same level during the epochs.



(a)

(b)

Figure 4.6: *Accuracy for each epoch during the training (a) and validation (b) from experiment 2. The blue curve illustrates the accuracy during training, the red curve illustrates the accuracy during validation, while the black curves illustrate a five-point moving average.*

4.1.3 Experiment 3

In the third experiment, training and validation were performed as described in section 4.1.1. These WSIs were from the small dataset, which included data with a concentration of tumor nuclei of 70% or more.

The cross entropy losses for the training and validation are illustrated in figure 4.7 for each epoch. Likewise the first two experiments, the graph of the cross entropy loss indicated a convergence in the model fit. However, the model fit occurred in later in the training during this experiment.

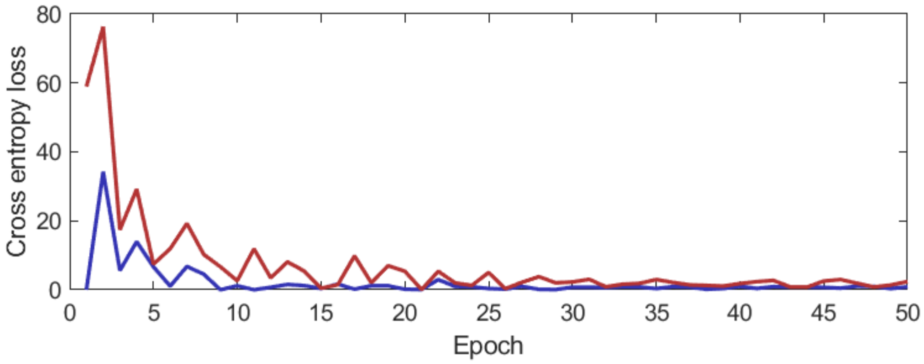


Figure 4.7: *Cross entropy loss for each epoch during the third experiment. The blue curve indicates the training loss and the red curve indicates the validation loss.*

Due to the great deviation in the loss between the initial epochs and the following epochs, an additional graph was constructed for the loss between zero and ten, as illustrated in figure 4.8. The losses of the third experiment showed greater oscillations during the validation epochs and a later tendency to converge than the past two experiments. However, the losses of the training in the third experiment resembled those in the past two experiments but with a slightly later tendency to converge. After approximately 9 epochs for the training and 29 epochs for the validation curve, the curves had approximately converged, but with a more evident flattening tendency of convergence within the training epochs than the validation epochs. After this flattening, the curves oscillated around a mean loss of approximately 0.7 for the training, and 1.9 for the validation, which indicated that the validation generally had a higher loss than the training during the epochs.

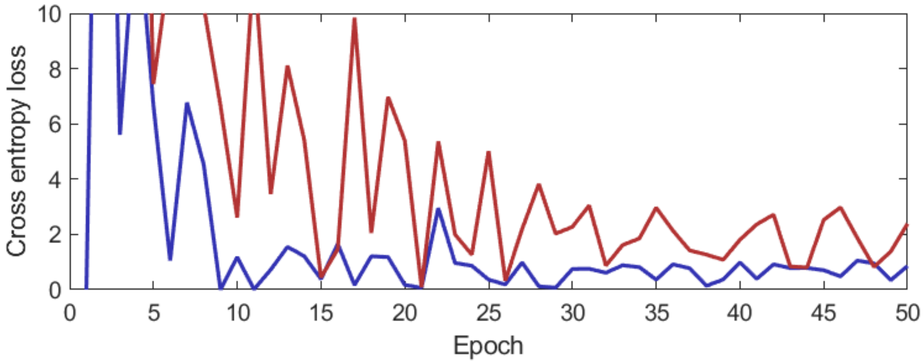
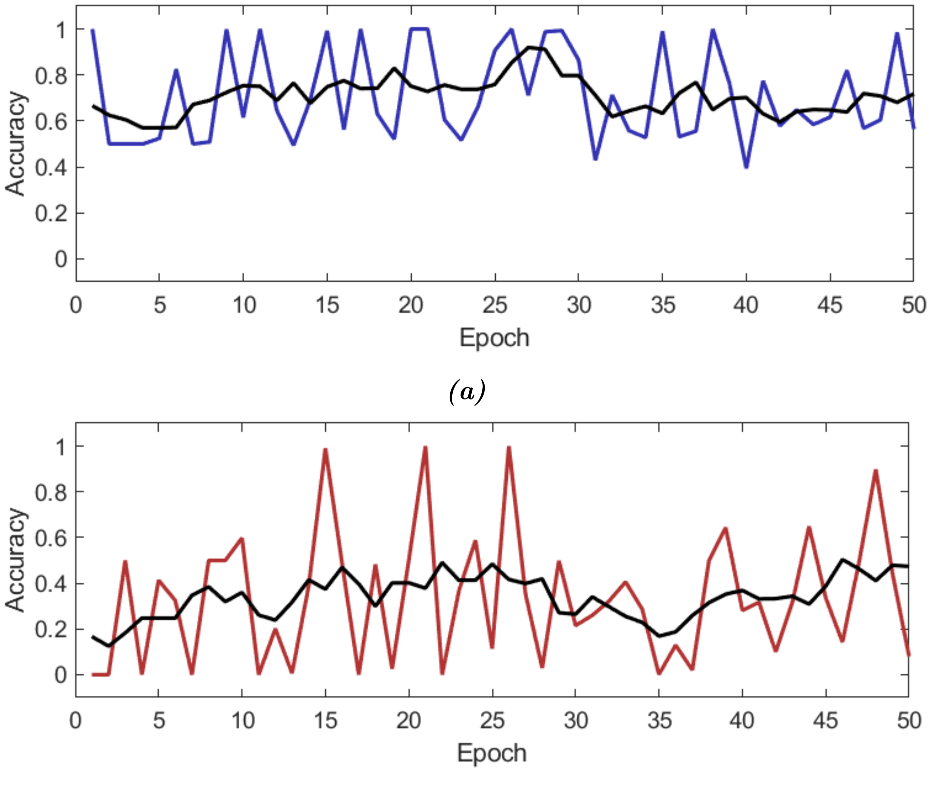


Figure 4.8: *Cross entropy loss for each epoch during the third experiment in which a limit for the y-axis was set between zero and ten. The blue curve indicates the training loss and the red curve indicates the validation loss.*

The accuracies of the training and validation during the third experiment is illustrated in figure 4.9. Large deviations in the accuracies of both the training and validation were present between epochs, likewise the past two experiments. For the training epochs, the minimum accuracy was 0.4, the mean accuracy was 0.7 and the maximum accuracy was 1.0. For the validation epochs, the minimum accuracy was 0.0, the mean accuracy was 0.3 and the maximum accuracy was 1.0. Likewise the first experiment, the accuracies of the epochs in training and validation had large smooth oscillations when investigating the moving average. Thereby, a slight tendency of increasing until epoch 25 was observed in the training accuracies after which a slightly decreasing tendency was observed, as illustrated in figure 4.9a. The tendencies of decreasing and increasing accuracy varied more in the validation, as the accuracies had a slight tendency of increasing until the 25th epoch, after which a slightly decreasing tendency was observed until the 35th epoch. Finally, a slightly decreasing tendency was observed until the 50th epoch, as illustrated in figure 4.9b. Furthermore, it was observed that the moving average curve of the accuracies of the training generally was higher than the moving average curve of the accuracies of the validation. The expectation of a better performance due to the concentration of tumor nuclei within the WSIs was not satisfied, as both the loss and accuracy were not markedly better than the results from the first experiment, which used the same number of patches and WSIs per epoch.



(a)

(b)

Figure 4.9: *Accuracy for each epoch during the training (a) and validation (b) from experiment 3. The blue curve illustrates the accuracy during training, the red curve illustrates the accuracy during validation, while the black curves illustrate a five-point moving average.*

4.1.4 Experiment 4

In the fourth experiment, training and validation were performed as described in section 4.1.2. Likewise the third experiment, the WSIs used, were from the small dataset, which included data with a concentration of tumor nuclei of 70% or more.

These WSIs were from the small dataset, The cross entropy loss for the training and validation are illustrated in figure 4.7 for each epoch. Likewise the previous experiments, the graph of the cross entropy loss indicated a convergence in the model fit.

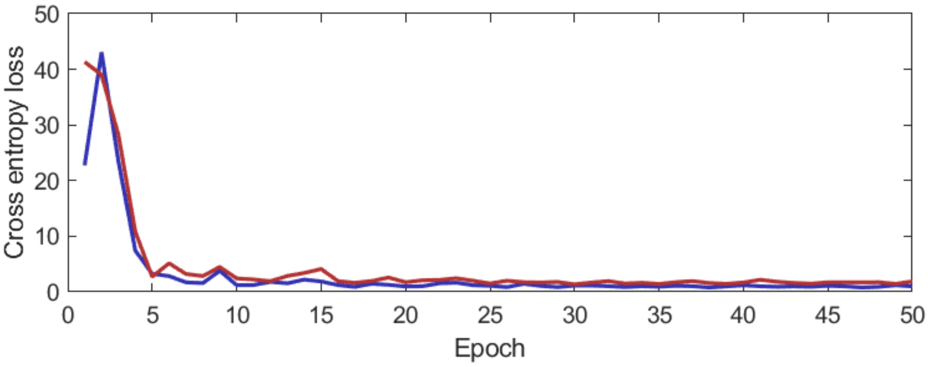


Figure 4.10: *Cross entropy loss for each epoch during the fourth experiment. The blue curve indicates the training loss and the red curve indicates the validation loss.*

Due to the great deviation between the loss between the initial epochs and the following epochs, an additional graph was constructed for the loss between zero and ten, as illustrated in figure 4.11. These curves resembled the previous loss curves from experiment 1 and 2. However, tendency of convergence of the loss during the training and validation epochs in this experiment did in general occur earlier than the previous three experiments. After approximately five epochs, both curves had a tendency to flatten, and thereby approximately converged. After this flattening, the curves oscillated around a mean loss of approximately 1.3 for the training epochs, and 2.1 for the validation epochs, which indicated that the training generally had a slightly higher loss than the validation during the epochs.

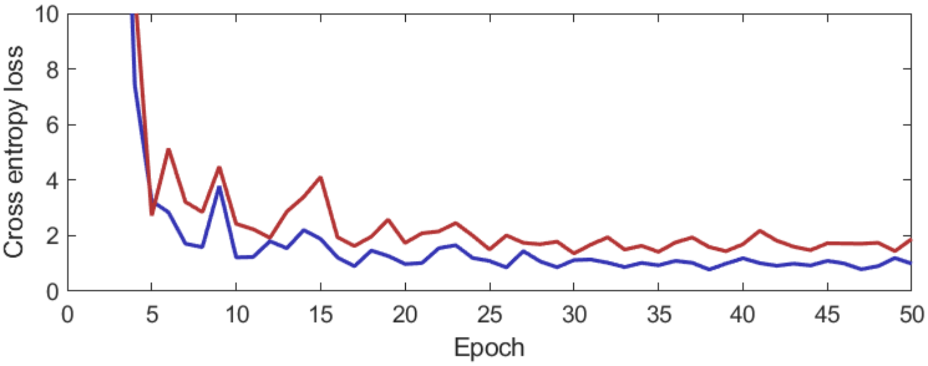
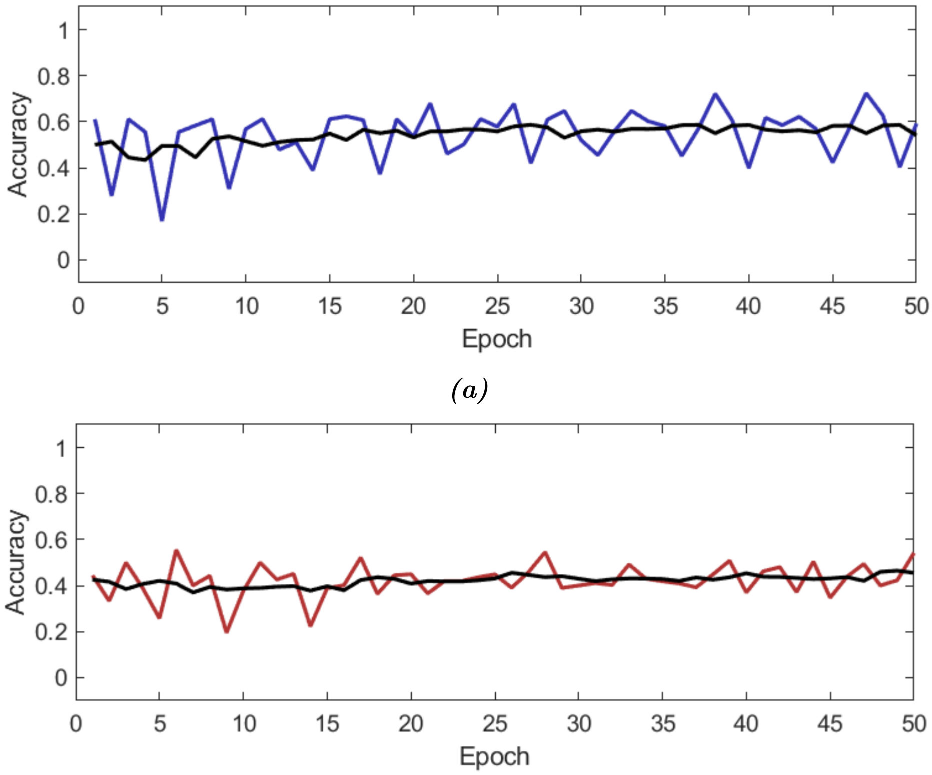


Figure 4.11: *Cross entropy loss for each epoch during the fourth experiment in which a limit for the y-axis was set between zero and ten. The blue curve indicates the training loss and the red curve indicates the validation loss.*

The accuracies of the training and validation epochs during the fourth experiment is illustrated in figure 4.12. Deviations in the accuracies of both the training and validation were present between epochs. However, the oscillations were smaller in the fourth experiment than in the remaining three experiments. For the training epochs, the minimum accuracy was 0.2, mean accuracy was 0.5 and maximum accuracy was 0.7. For the validation epochs, the minimum accuracy was 0.2, mean accuracy was 0.4 and maximum accuracy was 0.6. Likewise the second experiment, when investigating the moving average of the accuracies, a slight tendency of increasing in accuracies was observed during the first epochs in the training, while no tendency of decreasing or increasing was observed in the validation. However, the slightly increasing tendency ended after 20 epochs in this experiment, which was five epochs earlier than in the second experiment. The accuracies of the training and validation epochs are illustrated in figure 4.12a and figure 4.12b. Furthermore, it was observed that the moving average curve of the accuracies of the training epochs generally was higher than the moving average curve of the accuracies of the validation epochs. Both the loss and accuracy were not markedly better than the results from experiment 2, which used the same number of patches and WSIs per epoch. Thereby, the expectation of a better performance due to the concentration of tumor nuclei within the WSIs was not satisfied, likewise for experiment 3.



(a)

(b)

Figure 4.12: *Accuracy for each epoch during the training (a) and validation (b) from experiment 4. The blue curve illustrates the accuracy during training, the red curve illustrates the accuracy during validation, while the black curves illustrate a five-point moving average.*

4.2 Test Results

Image classification was performed patch-based and WSI-based on the test datasets from both the small and the large dataset with the trained models. Three different WSI-classification approaches were investigated in the test when assigning a label for the WSI, based on the predicted patch labels:

1. Extract 10% of the predicted labels with the highest probability and provide the WSI with the most frequent label
2. Extract the predicted labels with a probability above a threshold of 0.3 and provide the WSI with the most frequent label
3. Include all predicted labels and provide the WSI with the most frequent label

These three WSI-classification approaches were investigated for all four trained models from the 50th epoch with 2,494 patches from each WSI, which was the lowest number of patches available from the WSIs in the test datasets after the exclusion of background patches. Furthermore, the accuracies of the WSI-based labels from the three WSI-classification approaches were compared to the accuracy of the patch-based predictions performed by the test. Additionally, confusion matrices were produced to evaluate the performance of the three WSI-classification approaches and the patch-based classification. For the test of experiment 1 and 2, a total number of 62 WSIs were used, distributed as 35 stage I, 16 stage II, 6 stage III, and 5 stage IV WSIs. For the test of experiment 3 and 4, a total number of 16 WSIs were used, distributed as 9 stage I, 3 stage II, 3 stage III, and 1 stage IV WSIs.

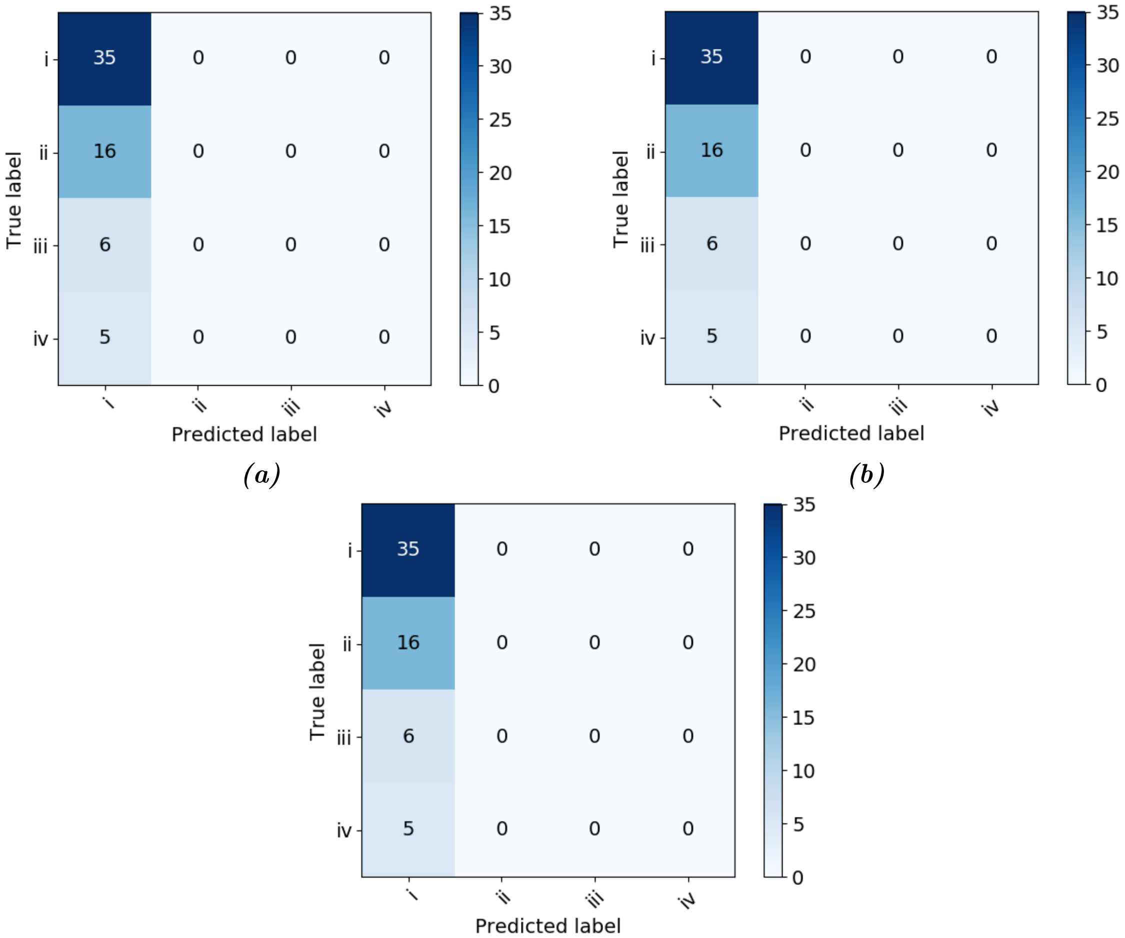
4.2.1 Experiment 1

The accuracies for the WSI-based classifications with the three different WSI-classification approaches using the trained model from experiment 1 are depicted in table 4.1. The WSIs used in this test were acquired from the large dataset.

	Accuracy
Approach 1	0.56
Approach 2	0.56
Approach 3	0.56
Patch-based	0.56

Table 4.1: *Accuracy from the three different WSI-classification approaches in the first experiment along with the accuracy of the patch-based classification.*

The accuracies were equal for the three WSI-classification approaches, which furthermore were equal to the accuracy of the patch-based classification. The distribution of the assigned WSI-labels from the three WSI-classification approaches is illustrated in the confusion matrices in figure 4.13, which indicated the performance of the test. From the confusion matrices it was evident that the WSIs only were classified as stage I in all three WSI-classification approaches. Therefore, the 35 WSIs which were stage I, were also classified correctly. Since no other stages were classified, the 27 WSIs which were not stage I were classified incorrectly in all three WSI-classification approaches.



(a)

(b)

(c)

Figure 4.13: *Confusion matrices based on WSI-labels from WSI-classification approach 1 (a), WSI-classification approach 2 (b), and WSI-classification approach 3 (c), in which the darkness of the blue colors indicate the frequency of a predicted class. The color bar indicates the colors corresponding to the frequency of the classification of the given class label.*

Furthermore, the distribution of the predicted patch-based labels were investigated to evaluate whether this test only predicted stage I labels or other stages as well. This distribution is illustrated in the confusion matrix in figure 4.14. The results from this confusion matrix verify that the majority of the patches have been classified as stage I, while only 65 patches were classified as stage II or III. This result hereof, was that 87,228 patches were correctly classified, while 67,400 patches were incorrectly classified, hence the accuracy of 0.56.

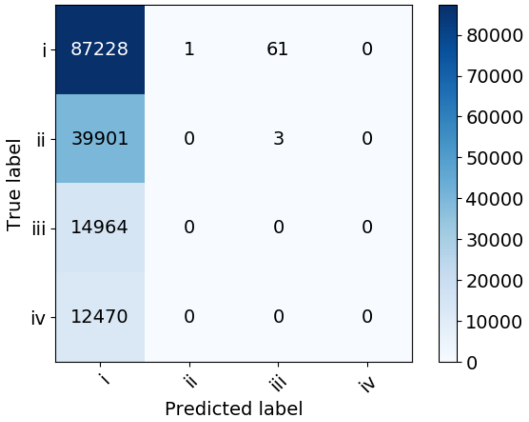


Figure 4.14: *Confusion matrices of predicted patch-based labels, in which the darkness of the blue colors indicate the frequency of a predicted class. The color bar indicates the colors corresponding to the frequency of the classification of the given class label.*

4.2.2 Experiment 2

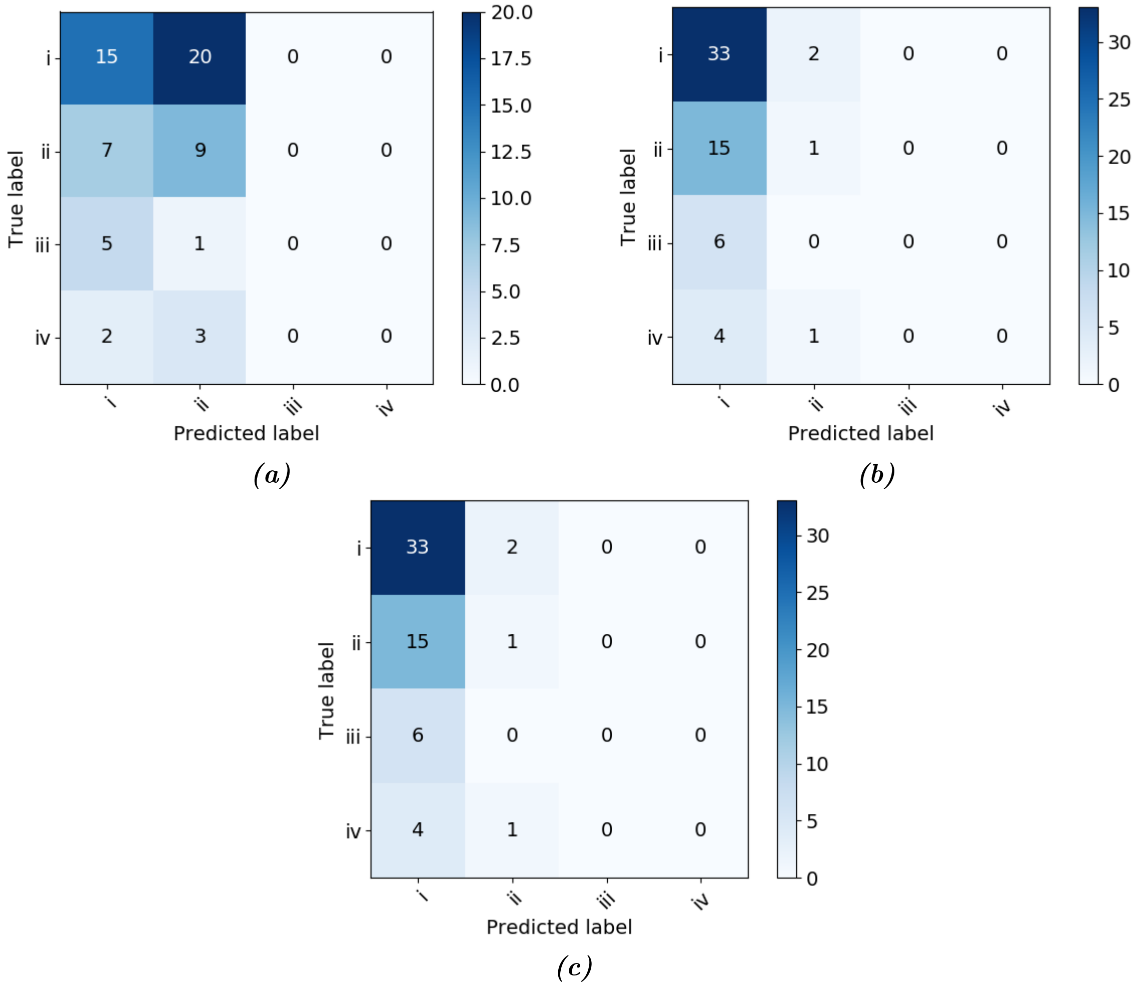
The accuracies for the WSI-based classifications with the three different WSI-classification approaches using the trained model from experiment 2 are depicted in table 4.2. Likewise experiment 1, the WSIs used in this test were acquired from the large dataset.

	Accuracy
Approach 1	0.39
Approach 2	0.55
Approach 3	0.55
Patch-based	0.46

Table 4.2: *Accuracy from the three different WSI-classification approaches in the second experiment along with the accuracy of the patch-based classification.*

The accuracies from the three WSI-classification approaches ranged between 0.39 and 0.55, of which both WSI-classification approaches 2 and 3 had an accuracy of 0.55. The accuracy of WSI-classification approach 2 and 3 were slightly larger than the accuracy of the patch-based classification, while the accuracy of WSI-classification approach 1 was slightly smaller than the accuracy of the patch-based classification.

The performance of the test is illustrated in the confusion matrices in figure 4.15, in which it was evident that the WSIs had only been classified as stage I or II in the three WSI-classification approaches. In the first WSI-classification approach, the WSIs predominantly were classified as stage II, while they predominantly were classified as stage I in WSI-classification approaches 2 and 3. In the first WSI-classification approach 24 WSIs were classified correctly as either stage I or II, while 38 WSIs were classified incorrectly. In the second and third WSI-classification approaches, 34 WSIs were classified correctly as either stage I or II, while 28 WSIs were classified incorrectly. Thereby, the WSIs in the test of this experiment were generally classified more diverse than the WSI-labels in experiment 1.



(a)

(b)

(c)

Figure 4.15: *Confusion matrices based on WSI-labels from WSI-classification approach 1 (a), WSI-classification approach 2 (b), and WSI-classification approach 3 (c), in which the darkness of the blue colors indicate the frequency of a predicted class. The color bar indicates the colors corresponding to the frequency of the classification of the given class label.*

The distribution of the predicted patch-based labels was investigated in a confusion matrix, which is illustrated in figure 4.16. Based on this, it was possible to evaluate whether some stages were more represented in the patch-based labels. The results from this confusion matrix demonstrated that approximately one-third of the patches, corresponding to 105,101, were classified as stage I, of which 58,720 patches have been classified correctly as stage I. Furthermore, 11,957 patches were classified correctly as stage II, 85 patches were classified correctly as stage III, and only 9 patches were classified correctly as stage IV. This resulted in a total of 70,771 patches being correctly classified and 83,857 patches being incorrectly classified, hence the accuracy of 0.46. Thereby, the larger diversity compared to experiment 1, indicated in the results from the three WSI-classification approach, was likewise demonstrated for the patch-based classifications.

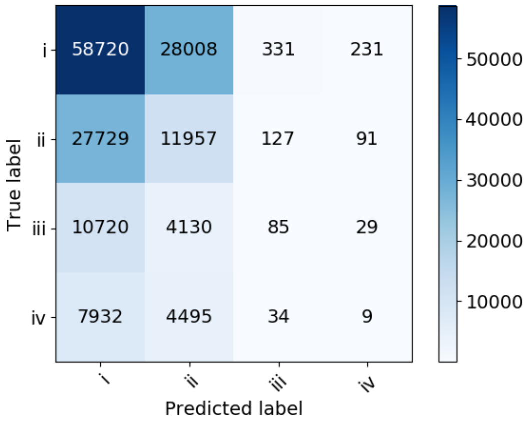


Figure 4.16: *Confusion matrices of predicted patch-based labels, in which the darkness of the blue colors indicate the frequency of a predicted class. The color bar indicates the colors corresponding to the frequency of the classification of the given class label.*

4.2.3 Experiment 3

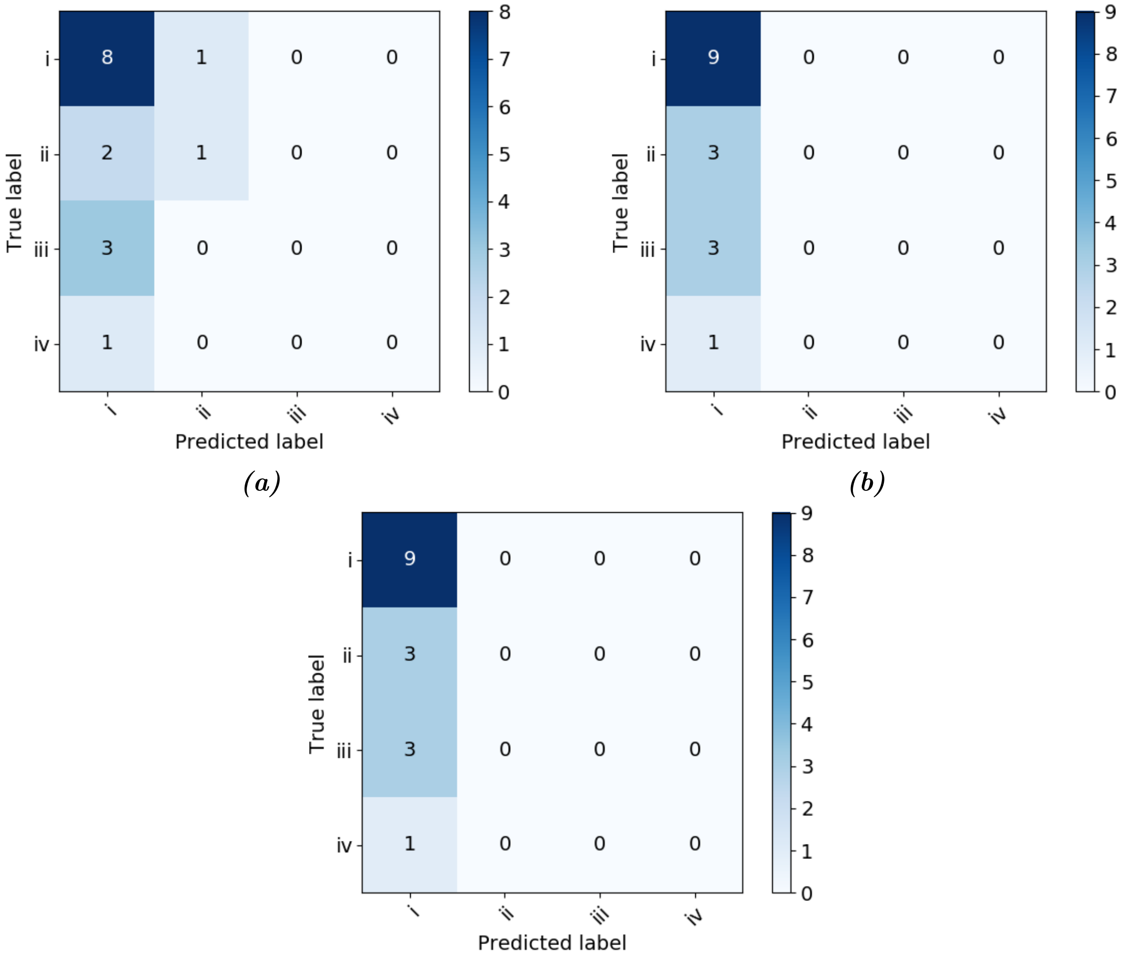
The accuracies for the WSI-based classifications with the three different WSI-classification approaches using the trained model from experiment 3 are depicted in table 4.3. The WSIs used in this test were acquired from the small dataset.

	Accuracy
Approach 1	0.56
Approach 2	0.56
Approach 3	0.56
Patch-based	0.51

Table 4.3: *Accuracy from the three different WSI-classification approaches in the third experiment along with the accuracy of the patch-based classification.*

The accuracies were equal for the three WSI-classification approaches, which all were slightly larger than the accuracy of the patch-based classification. To investigate differences from the WSI-classification approaches, confusion matrices were produced, which are illustrated in figure 4.17. From the confusion matrices it was evident that the WSIs predominantly were classified as stage I for the first WSI-classification approach, similarly to the second and third WSI-classification approach in experiment 2. However, the WSIs were only classified as stage I in the second and third WSI-classification approach, similarly to the three WSI-classification approaches in experiment 1.

In the first WSI-classification approach, nine WSIs were classified correctly as either stage I or II, while seven WSIs were classified incorrectly. In the second and third WSI-classification approach, the nine WSIs which were stage I were classified correctly, while the seven which were not stage I were classified incorrectly. Thereby, the classified WSI-labels in the test of this experiment were generally more diverse than the WSI-labels in experiment 1, but less diverse than the WSI-labels in experiment 2.



(a)

(b)

(c)

Figure 4.17: *Confusion matrices based on WSI-labels from WSI-classification approach 1 (a), WSI-classification approach 2 (b), and WSI-classification approach 3 (c), in which the darkness of the blue colors indicate the frequency of a predicted class. The color bar indicates the colors corresponding to the frequency of the classification of the given class label.*

A confusion matrix of the predicted patch-based labels has been produced to investigate whether the distribution of these indicated a predominant classification of stage I, likewise the WSI-classification approaches. The confusion matrix is illustrated in figure 4.18. The results from this confusion matrix verified that the majority of the patches have been classified as stage I with 35,081 predicted patch-labels, while the second most classified was stage II with 4,210 predicted patch-labels. Furthermore, 613 patches were classified as stage III and no patches were classified as stage IV. This resulted in a total of 20,412 correctly classified patches and 19,492 incorrectly classified patches, hence the accuracy of 0.51. The trend of a larger diversity compared to experiment 1 and smaller diversity compared to experiment 2, indicated in the results from the three WSI-classification approach, was furthermore demonstrated for the patch-based classifications.

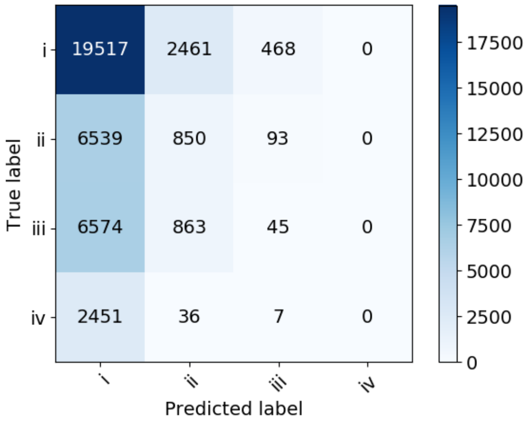


Figure 4.18: *Confusion matrices of predicted patch-based labels, in which the darkness of the blue colors indicate the frequency of a predicted class. The color bar indicates the colors corresponding to the frequency of the classification of the given class label.*

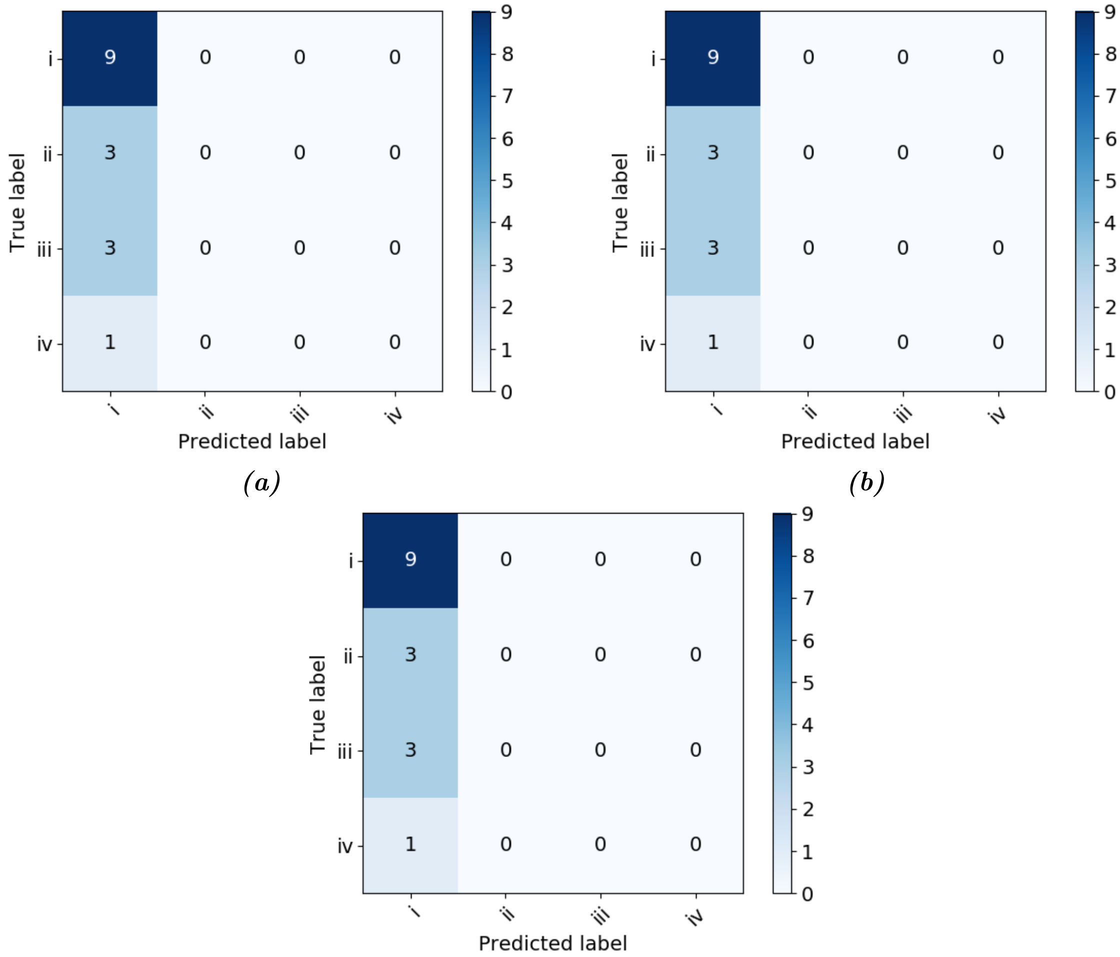
4.2.4 Experiment 4

The accuracies for the WSI-based classifications with the three different WSI-classification approaches using the trained model from experiment 4 are depicted in table 4.4. Likewise experiment 3, the WSIs used in this test were acquired from the small dataset.

	Accuracy
Approach 1	0.56
Approach 2	0.56
Approach 3	0.56
Patch-based	0.54

Table 4.4: *Accuracy from the three different WSI-classification approaches in the fourth experiment along with the accuracy of the patch-based classification.*

The accuracies were equal for the three WSI-classification approaches, which all were slightly larger than the accuracy of the patch-based classification. The performance of the three WSI-classification approaches in this test is illustrated in the confusion matrices in figure 4.19. From the confusion matrices it was evident that the WSIs only were classified as stage I in all three WSI-classification approaches. Therefore, the nine WSIs which were stage I were also classified correctly, but no other stages were classified. Thereby, the seven WSIs which were not stage I were classified incorrectly in all three WSI-classification approaches. Thereby, the classified WSI-labels in the test of this experiment were generally less diverse than the WSI-labels in experiment 2 and 3, but resembled classification of WSI-labels in experiment 1.



(a)

(b)

(c)

Figure 4.19: *Confusion matrices based on WSI-labels from WSI-classification approach 1 (a), WSI-classification approach 2 (b), and WSI-classification approach 3 (c), in which the darkness of the blue colors indicate the frequency of a predicted class. The color bar indicates the colors corresponding to the frequency of the classification of the given class label.*

To investigate the predicted patch-based labels which the three WSI-classification approaches were evaluated from, a confusion matrix of the predicted patch-based labels was produced, which is illustrated in figure 4.20. This enabled an evaluation of whether only stage I labels had been predicted or other stages as well. The results from the confusion matrix demonstrated that other stages than stage I have been classified, but the majority of the patches have been classified as stage I with 36,125 predicted patch-labels, while 3,769 patches were classified as stage II, only 2 patches were classified as stage III, and 8 patches were classified as stage IV. As a result hereof, 21,526 patches were classified correctly and 18,378 patches were classified incorrectly, hence the accuracy of 0.54. Thereby, a larger diversity in the classified patches than the results from experiment 1 was demonstrated. Furthermore, the trend of a smaller diversity compared to experiment 2 and 3, indicated in the results from the three WSI-classification approach, was furthermore demonstrated for the patch-based classifications.

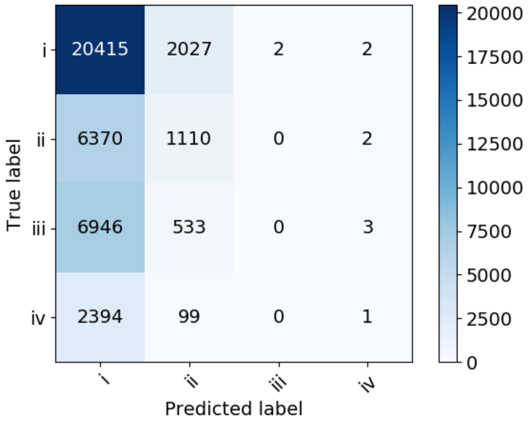


Figure 4.20: *Confusion matrices of predicted patch-based labels, in which the darkness of the blue colors indicate the frequency of a predicted class. The color bar indicates the colors corresponding to the frequency of the classification of the given class label.*

4.2.5 Comparison of the Three WSI-Classification Approaches

The accuracy from three different WSI-classification approaches to classify the WSIs using patch-based predicted labels from all four experiments are depicted in table 4.5. In this, it is depicted that WSI-classification approach 2 and 3 had the overall highest accuracy.

	Accuracy				
	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Mean (±std)
Approach 1	0.56	0.39	0.56	0.56	0.52±0.09
Approach 2	0.56	0.55	0.56	0.56	0.56±0.01
Approach 3	0.56	0.55	0.56	0.56	0.56±0.01
Patch-based	0.56	0.46	0.51	0.54	0.52±0.04

Table 4.5: *Accuracy from the three different WSI-classification approaches from the four experiments along with the accuracy of the patch-based classification.*

Synthesis 5

5.1 Discussion

In the present project, a CNN to classify the stage of lung cancer was developed, trained, and tested using H&E stained WSIs from TCGA. Two different datasets were constructed based on the tumor nuclei percentage of which the large dataset contained all the WSIs, while the small dataset only contained WSIs which contained 70% or more of tumor nuclei. Two training approaches have been conducted in which the first approach included 2,000 patches from two WSIs, while the other approach included 256 patches from 18 WSIs. These approaches were conducted with each of the two constructed datasets. The CNN was constructed based on the AlexNet, and fine-tuning of the AlexNet model was performed using the pretrained weights and biases from the initial model described in Krizhevsky et al. [2012], before they trained using data augmentation and using initialization of non-zero biases. The fine-tuning was performed by retraining the last two layers to adjust the model to classify the lung cancer stage from WSIs. The results and observations from the data selection, training, validation, and test, and considerations regarding the methods used in the project will be discussed in the following subsections.

5.1.1 Main Results

Four different experiments were conducted in which the first two experiments used the large dataset which included all the data after the data selection. The last two experiments used the small dataset which included data after the data selection in which the WSIs had a concentration of tumor nuclei of 70% or more. Furthermore, the number of WSIs and patches from these varied in the experiments. In the first and third experiment, training and validation were performed for 50 epochs and with a batch size of 4,000 patches corresponding to 2,000 patches from 2 WSIs per epoch. In the second and fourth experiment, training and validation were performed for 50 epochs and with a batch size of 4,500 patches corresponding to 250 patches from 18 WSIs per epoch. It was assumed that the training and validation loss curve would be descending during the initiating epochs and after this would converge as the model was fully updated. Additionally, the accuracy curves was assumed to increase during the epochs for the four experiments. Based on the number of input WSIs, it was assumed that the accuracy and loss curves would appear more smooth in the second and fourth experiment compared to the first and third experiment, as patches from multiple WSIs were presented to the network per epoch. Thereby a more general input of cell nuclei were fed to the network by which the noisy patches should be balanced. Furthermore, it was assumed that the last two experiments would have a better performance than the experiment with the equivalent number of WSIs and patches. This is due to a better representation of tumor nuclei which should be represented in the separate patches due to the reported tumor nuclei percentages.

During all four of these experiments, a convergence in the fit of the model was observed. This was due to the loss curve decreasing with a steep slope during the initial epochs after which the loss curves after which the loss curves began to converge. In all four experiments,

the losses of the validation were slightly higher than the losses in the training. In the third experiment, greater oscillations occurred especially in the loss of the validation. However, as the differences between the training and validation losses were not remarkable in the four experiments, overfitting was not considered an issue.
As a convergence was observed within the loss curve for the four experiments, it might have been beneficial to stop the training of the model earlier to prevent possible overfitting. Of the four models trained, three of them should have ended the training session after approximately 10 epochs, while a number of 50 epochs was more suitable for the third experiment.

Oscillations between the epochs were observed in the accuracies during training and validation for all four experiments. These oscillations might have been caused by the great variation within the dataset due to staining intensities which were observed in the visual inspection in data selection. Furthermore, the great variation with regards to the concentration of tumor nuclei may have caused the oscillations.
The performances of the first and third experiment due to the accuracies were approximately identical when considering the minimum, maximum, and mean accuracy with a slightly better performance in experiment 1. This was unexpected as the patches presented in the third experiment was from WSIs with a tumor nuclei percentage with 70% or more. Thereby, the chance of the model being present for patches which included tumor nuclei should be greater for all the patches used in the third experiment and highest during the initial epochs in the first experiment. The oscillations were especially a problem in experiment 1 and 3, as only patches from two WSIs were fed to the model per epoch, by which the weights might have been adjusted based on patches from two WSIs in one epoch and then updated based patches from two markedly different WSIs in the following epoch. Thereby, overshooting of the weights in different directions might have occurred. Due to this, both increasing and decreasing tendencies were observed during the training and validation accuracy curves. The reason for the accuracy curves both increasing and decreasing might be caused by randomness, due to the number of WSIs presented per epoch. Thereby the variation of staining intensities within the patches and associated classes presented for the network is very limited per epoch in these experiments.
On the contrary, the second and fourth experiments were similar with regards to the accuracies and the tendencies hereof. A slight tendency of increasing accuracies in the training epochs were observed in these experiments. However, no tendency of either increase nor decrease was observed during the validation epochs. Furthermore, the oscillations were smaller during experiment 2 and 4 compared to the accuracies of experiment 1 and 3. This might have been due to the number of WSIs included as input in experiment 2 and 4. Since more WSIs were used in each epoch, the weights were updated based on a more diverse amount of data. However, oscillations still occurred in these experiments, as some epochs might have been presented to some WSIs while the next epoch might have been presented to predominantly different WSIs due to the variations of staining intensity within the dataset.

When testing the trained model, the results from all three WSI-classification approaches indicated that the models from all four experiments have been highly affected by the skewed datasets, by which only stage I or stage I and II WSI-labels have been classified during the WSI-classification approaches. In general, the main obstacles when training the model have been the limitation of patches which could be used per epoch due to the the TensorFlow limit with regards to input capacity. The model was thereby updated based on two or 18 WSIs respectively in the four experiments, and only a small percentage of patches available for each of the WSIs. As not all WSIs were included per epoch, a poor representation of the classes were given as input per epoch. This resulted in some epochs only including few or none WSIs from the less represented classes within the dataset. Thereby it was hard for the model to learn features from all the classes.

Three WSI-classification approaches were applied for the test of the model in which a final label was assigned to each WSI. This was due to the concern about noisy labels by which it was investigated how postprocessing could be performed to manage these and give a correct WSI-based classification.
The three WSI-classification approaches had similar accuracies when using the classified patch labels in evaluation of the assiciated WSI. However, as slightly higher accuracy were presented in the WSI-classification approach in which all patches were used in the evaluation and in the WSI-classification approach in which patches predicted with a probability above 0.3 were used in the evaluation. This might be due to a greater number of patches included in the evaluation of these WSI-classification approaches than the WSI-classification approach which only included 10% of the predicted patch class labels. The final WSI-classifications were predominantly given as stage I for all approaches in all experiments and in some of these solely stage I WSI-classifications were made. However, patches were not always classified as stage I, but as the most frequently patch label were given as stage I, the likelihood for a WSI to be classified as this stage were greater than the remaining stages. With regards to the classified patches in the different experiments, the greatest diversity were explored in the second experiment. This could be anticipated, as cancer stage III and IV were more well-represented in the large dataset which was used in this experiment. Furthermore, in this experiment the greatest number of WSIs are used, by which a greater variability is represented in the different epochs.
The first first WSI-classification approach gave the lowest accuracies when classifying the WSI. In this WSI-classification approach, the final label for the WSI was based on 10% of the classified patch labels with the highest probabilities. Thereby, it was controlled that only 10% of class labels with the most probable classification were included in the decision of the final WSI label. Furthermore, with this WSI-classification approach it was ensured that the decision of the final label for the WSI was based on a certain amount of patch labels. However, it was unknown how much malignant tissue which was included in the classified patches. If a patch only included a minimal amount of malignant tissue or only benign tissue, a patch label might have been given with a high probability as another class. Furthermore, the final label for the WSI was based on a fixed number of patches, by which it was not influenced by the magnitude of the probabilities for the predictions, but only that it belongs within the 10% highest predicted probabilities for the individual WSI. Thereby, the predicted probabilities were not controlled with this WSI-classification approach. With this WSI-classification approach, it appears to be a too poor representation of the WSI included in the evaluation to give a general presentation of the WSI.
The second WSI-classification approach were one of the approaches which resulted in the highest probability when classifying the WSI. With this WSI-classification approach, all labels with a predicted probability above a threshold of 0.3 were included. If a patch contained few tumor nuclei, it might have been classified correct, but with a lower probability. Thereby, it might have been possible to include more patches, as the number of included patches

was not limited, but controlled by a probability-based threshold. However, with this WSI-classification approach, it was uncertain how many predicted patch labels which were included in the decision of the final WSI label. Thereby, the final WSI label could have been impossible to give if no patch labels were predicted with a probability above the threshold. Furthermore, the final WSI label could have been given based on a single or very few patch labels, which would not have assigned a general representation of the WSI.
The third WSI-classification approach were additionally one of the approaches which gave the highest probability when classifying the WSI. With this WSI-classification approach, all predicted class labels were included in the decision of the final WSI label. Inclusion of all predicted class labels enabled a general representation of how well the WSI was classified. However, with this WSI-classification approach, a probability of approximately zero was acknowledged as a reliable proposal for the patch label and thereby the final WSI label. It might have been beneficial to include the predicted probability in the decision by giving the predicted labels with a higher probability a higher influence on the final WSI label.

5.1.2 Data Selection

It was unknown how the information regarding the cancer stage, cancer type, and tumor nuclei percentage in the metadata was acquired. Thereby, it was unknown whether the lung cancer type and stage were identified by pathologists, researchers, or someone else. Likewise for the tumor nuclei percentage, it was unknown how this was obtained from the WSIs and whether this was done by the same person who identified the lung cancer type and stage or by an algorithm. Some of the WSIs which were labeled as a certain lung cancer type and stage also had a tumor nuclei percentage of 0% reported, which should not be possible. Therefore, it would have been interesting to retrieve information about the method of which the tumor nuclei percentage was obtained, to understand how WSIs which were classified as lung cancer also were labeled as having no tumor nuclei.
Metastases and primary tumors often contain necrotic tissue after chemotherapy which could be an explanation for the great variation of reported tumor nuclei percentages within the WSIs. Necrotic tissue often contains fewer viable tumor cells compared to solid tumor tissue by which the necrotic tissue could have been reported as containing fewer tumor nuclei than other samples. However, it has been reported that an analysis of mutations will still be possible even though a sample consists of many non-viable tumor cells. [Büttner et al., 2017]
Since the WSIs had annotations of lung cancer type and stage, these were still used as input to the CNN. However, as all data were included in the dataset regardless of tumor nuclei percentage, a great variation within the dataset with regards to the percentage of tumor nuclei was observed. Therefore, it was ensured that the network was fed with the samples which contained the most tumor nuclei first and the WSIs with the least tumor nuclei at last. In this way, the possibility of the network being presented to patches, which contained tumor nuclei in the initial epochs, was higher. This was expected to give a better result, as the initiating patches from the WSIs presented to the network had a higher probability of containing tumor nuclei by which the model more likely would learn the core of the presentation in the initiating epochs.

During the determination of the tumor nuclei percentage, the location of the tumor nuclei must has been known. Thereby, it should have been possible to create a mapping of the locations of the tumor nuclei in the WSIs if information about this was saved in the metadata.

Patches could then be extracted from these locations, and thereby ensure the representation of malignant tissue in the patches used for the CNN. This would reject the concern about noisy labels that occurred in the used datasets, as it was assumed that the ground truth label for the WSI additionally could be used as the ground truth label for each of the patches in the WSI. However, this information was not available in the metadata by which it was only possible to use the percentage as information for the entire WSI.

During data acquisition from the bronchus and lung subset of TCGA, unknown errors occurred during download which corrupted the files. This resulted in a smaller initial dataset than the data available from TCGA. Furthermore, only WSIs with a magnification level of x40 were used in this project, which was retrieved during the extraction of the subimages. This resulted in a reduction of the dataset. A different number of WSIs were available with a magnification level of x20 and x40, as depicted in appendix A. To obtain a larger dataset, WSIs with a magnification level of x20 could have been used in this project. Thereby, a total of 1,117 WSIs could have been available when all data were included, which was twice the number of WSIs with a magnification level of x40. However, as the cancer type and stage were diagnosed based on the appearance of the tumor cells [Moreira and Saqi, 2015], a high resolution of the WSIs was essential by which the edges of the cells appear as distinctly as possible. Furthermore, it was investigated whether the four cancer stages were more equally distributed within the WSIs with a magnification level of x20 compared to the WSIs with a magnification level of x40. However, an approximately identical distribution was observed between the WSIs of the two magnification levels. Furthermore, only a small difference between the number of WSIs available with the different magnification levels was observed in the small dataset. In these, a total number of 107 WSIs with a magnification level of x40 were included and a total number of 150 WSIs with a magnification level of x20 were included. Therefore, as the WSIs with a magnification level of x40 had the highest resolution concurrent with a relatively large dataset available, it was chosen to use these WSIs in the project.

Through visual inspection of the acquired data, it was found that there were ink stains on some of the WSIs. The ink stains might have indicated useful information about locations of the malignant or benign tissue in the WSI, by which they could have been highly useful. If annotations had been available regarding locations of malignant tissue, it would have been possible to only extract the patches with useful information. However, a description of the meaning of these ink stains was not available within the metadata. Thereby, the WSIs with ink stains were excluded from the dataset, as the ink was not biological material and thereby could affect the learned features from the WSI.

To obtain a larger dataset, it might have been possible to automatically extract and use only the patches from the WSIs which did not contain ink. However, the ink stains were in different colors and transparency, which resembled the variations observed regarding staining intensity in the WSIs. Thereby, it might be favorable to crop the WSIs or remove the patches with ink stains manually through a visual inspection of all the patches in the WSIs. However, given that a large dataset was still available without the WSIs containing ink stains, it was chosen to exclude the WSIs and not single patches.

5.1.3 Training and Validation of the CNN

The ground truth for the data used in this project was WSI-based, by which it was uncertain whether malignant tissue was present in all the patches used as input for the CNN. Furthermore, it was uncertain whether the patches included tumor tissue from the labeled stage in combination with tumor tissue from earlier stages or benign tissue. Thereby, mixed labels could additionally have been essential when labeling the patches. This raised concern about noisy labels, as all patches inherited labels from the WSI they had been extracted from. During the training session, the model was trained and validated on patch-based labels and not WSI-based. Thereby, the weights of the CNN were updated based on patches. It might have been considered to adjust the training and validation to update based on all the extracted patches from the WSI, as this would minimize the concern about noisy labels. However, this was incorporated in the test of the trained model, which performed patch-based predictions of labels that were used to assign a predicted label for the associated WSI.
To reduce the noisy labels, it might have been beneficial to train the model and save the performances for the patches. Thereby, patches that are misclassified with a high probability could have been excluded from the dataset. If a patch is misclassified with a high probability, it might contain more benign tissue or other stages of cancer than the one labeled as, by which noisy labels occur. Additionally, to get a more robust training and validation session, patches with a high entropy could have been ignored. Thereby, the model would solely be trained on the patches which more certainly have been given the correct ground truth label. Finally, WSIs or patches from a WSI could have been excluded due to the accuracy from an epoch. Some of the epochs had an accuracy of zero in the training and validation respectively. Based on this it might have been possible to observe a tendency within the data which often was misclassified. Thereby, it might have been possible to reconstruct the datasets in which the patches or WSIs with a tendency of misclassification would have been excluded before training the final model.

The acquired data from TCGA did not have an equal distribution of subjects with the four different lung cancer stages (I-IV). This skewed distribution was not addressed in this project, which may have affected the CNNs ability to learn a representation of the less represented lung cancer stages. Hereof, stage I represented 55.0% of the entire dataset, while stage IV only represented 2.9% in the large dataset, while stage I represented 55.1% and stage IV represented 2.8% in the small dataset. This problem could have been addressed by applying sample weighting during training, which would scale the loss by a given value based on the prevalence of a given class within the dataset. It is anticipated, that a change of the cost for the underrepresented class or an over- and undersampling of the different classes would have led to a better fit of the model. Oversampling could have been used to randomly duplicate samples from the underrepresented classes and thereby resulting in a larger sample size of these classes [Ramentol et al., 2011]. However, this may lead to overfitting of the model as the same patches would be presented to the model multiple times. On the contrary, samples from the overrepresented class could randomly have been deleted using undersampling, which would enable a smaller ratio between the sample size from the different classes [Ramentol et al., 2011]. Yet, this approach might lead to the removal of samples containing important variability of the predictive class.

Overfitting is the main obstacle when using CNNs [Chollet, 2018]. This obstacle can be addressed using different methods. A method used in this project to avoid overfitting was dropout. Dropout reduces the complexity of a model by dropping out a number of random features by setting them to zero during the training session [Chollet, 2018]. The dropout rate was chosen at 0.5 for the sixth and seventh layer, which were fully connected. This dropout rate supported the recommendation in fully connected layers in Hinton et al. [2012]. This recommendation assumes that hyper-parameter tuning is performed to determine the best dropout rate. However, it was chosen to keep a dropout rate at 0.5, as the focus of improving the model remained in the input data, as this was assumed a greater issue. Other methods of regularization instead of dropout could have been applied such as L1 and L2 regularization. In these, a cost is added to the loss function which only allows small values for the weights. However, overfitting did not appear significant during the initial training sessions. Thus, further regularization was not considered necessary in this project.

No augmentation was used during training of the CNN. Data augmentation applies random, realistic transformations to the existing training samples to make them differ from the original samples and imitate a larger dataset, which can be presented to the network [Chollet, 2018]. Data augmentation can be used in biomedical images to simulate realistic data variations, by which the network will learn variance in the data representation [Chollet, 2018]. Thereby, this could have been used to achieve more robustness of the trained model. It could have been interesting to include flip and mirror augmentation to the input data since the tissue can appear orientated in different directions. As the WSIs had a variation with respect to staining intensities, it would additionally have been interesting to include augmentations of the intensities of brightness and contrast. Furthermore, it could have been useful to use data augmentation for WSIs with the cancer stage III and IV to imitate a larger dataset and thereby reduce the ratio between the sample size of the different classes.

5.1.4 Test of the Trained Model

The three WSI-classification test approaches were performed on a fixed number of patches, extracted from the WSIs. However, the patches that contained less than 25% background and thereby were selected for further processing in the tests varied a lot within the datasets. In the test data using the large dataset, the average of possible patches to select from the WSIs for further processing was 27,858 patches, of which the WSI with the smallest number of patches had 2,494 patches while the WSI with the highest number of patches had 114,027 patches. Additionally, in the test data from the small dataset, the average of possible patches to select from the WSIs for further processing was 28,910 patches, of which the WSI with the smallest number of patches had 4,828 patches while the WSI with the highest number of patches had 107,781 patches. This might have been an issue, as a WSI in some cases were tested on most of the available patches, while other WSIs only were tested based on a small percentage of the patches available for further processing. Therefore, it might be interesting to test on all patches available from the WSIs to get a more general presentation of the WSI. Furthermore, the three WSI-classification approaches were only based on the probability of the classified patch labels, of which it might have been beneficial to additionally include the entropy. Thereby, the final WSI label would not only be based on whether a label was correct or incorrect, but also on the difference of the predicted label and the ground truth.

A way to illustrate the classification of the patches from a WSI could be through a map over a WSI with patches in colors according to performance. If this was performed on multiple WSIs with approximately the same percentage of tumor nuclei, it might have been possible to indicate tendencies of the patches which were classified correct or incorrect.

5.1.5 Future Investigations

Transfer learning has been used in this project to fine-tune an existing model which initially were trained on the ImageNet dataset. It was anticipated, that fine-tuning of a pre-trained network would lead to higher accuracy of the trained model. However, this is not always the case [Soria et al., 2009]. Therefore, it would be interesting to train the CNN end-to-end and compare the trained model with the one trained with fine-tuning. When training the model end-to-end, the low-level features would not have been inherited from the learned features from the pretrained model, which might differ from the low-level features found in histopathological images compared to images from ImageNet. Based on an end-to-end trained model, it would be possible to evaluate whether a negative transfer has occurred. If the performance in the end-to-end trained model is better than the model trained with transfer learning in this project, negative transfer has occurred.

In this project, the classification of the four cancer stages (I-IV) was based on patches from a WSI. Another approach could be to segment the nuclei and classify the stages based on this. In this project, the tumor nuclei percentage within a WSI is known, by which this could have been used as ground truth for the nuclei segmentation. During the segmentation of the nuclei, the shape, size, and spatial distribution can be detected, which plays an essential role in the context of a disease such as cancer grading. Thereby, classification on the basis of segmented nuclei could potentially increase the classification accuracy when detecting lung cancer stages. [Pang et al., 2010; Salvi and Molinari, 2018] Segmentation approaches for H&E stained histopathological images have been developed as CNNs [Pang et al., 2010; Salvi and Molinari, 2018], which would be interesting to include in a future solution to the problem regarding detection of lung cancer stage.

The subimages were created to enable the investigation performance of the models based on different patch sizes as input for the CNN. Therefore, the size of the subimages was chosen to be 512×512, as this was possible to square into different patch sizes. It could have been interesting to investigate which affect the size of the input patches would have on the classification accuracy.

The network did not manage to differ the four stages completely. Therefore, another approach could have been to train a classifier for each of the four stages, and then perform a fusion of these classifiers into one ensemble. Then it is anticipated, that these four classifiers will learn more specific representations of the data from the individual stage which they have been trained on, and thereby may facilitate a more accurate classification when exposed to new data. [Zhao and Liu, 2019]

5.2 Conclusion

In this project, a model of the AlexNet was fine-tuned in the last two layers to classify the four stages of lung cancer, for which data from the TCGA was used in this project. Prior to the fine-tuning of a CNN, data selection was performed to ensure homogeneity within the WSIs used. Furthermore, data preprocessing was performed to make the data manageable for the CNN. Throughout the data selection, WSIs with ink stains, WSIs with errors during the download, and WSIs which did not have a reported cancer stage were excluded, which resulted in a reduction from a total number of 3,220 WSIs available in the TCGA to 549 WSIs used in this project. All these WSIs were used in the large dataset, which included all the data, while the number of WSIs was further reduced to 107 in the small dataset as only WSIs with a tumor nuclei percentage of 70% or more were included.

The AlexNet was fine-tuned during four different experiments. During the training and validation, all these experiments showed a steep decrease in the loss during the initial epochs after which a tendency of a flattering curve was observed. Furthermore, during all the experiments, great oscillations were observed in the accuracies. In the experiments, in which each epoch was comprised of two WSIs with 2,000 patches extracted from each, no tendency of a descending or increasing accuracy was observed during the training and validation. On the contrary, in the experiments, in which each epoch was comprised of 18 WSIs with 250 patches extracted from each, a slight tendency of an increasing accuracy was observed during the training and validation.

During the tests, three approaches were tested along with a general patch-based performance. The best performance of the patch-based classification, was achieved with an accuracy of 0.56. Additionally, WSI-classification approach 2 and 3, in which patches classified with a probability above 0.3 and all classified patches respectively, showed the best overall performance of classifying WSIs, with a mean accuracy of 0.56±0.01 for the four experiments. Since the patch-based classification had this low an accuracy, it was not possible to perform postprocessing, which classified the WSI labels with a high accuracy.

This project confirms the difficulty of working with noisy labels and a skewed dataset. Therefore, it is suggested that methods to equalize the distribution of the dataset should be implemented in future investigations. Additionally, it is suggested that further methods to address noisy labels should be implemented in future investigations.

5.3 Acknowledgment

The data for this project was acquired from the bronchus and lung subset of The Cancer Genome Atlas (TCGA), which is a program by the National Cancer Institute and the National Human Genome Research Institute. We would like to acknowledge the TCGA Research Network for their contribution to the data collection and analysis. Furthermore, we would like to mark our appreciation for the specimen donors from the bronchus and lung subset and research groups.
The content of this report is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Human Genome Research Institute.

Bibliography

Bándi et al., 2019. Péter Bándi, Oscar Geessink, Quirine Manson et al. *From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge*. IEEE Transactions on Medical Imaging. doi: 10.1109/TMI.2018.2867350.

Bengio et al., 1994. Y Bengio, P Simard and P Frasconi. *Learning Long-term Dependencies with Gradient Descent is Difficult*. IEEE Transactions on Neural Networks. doi: 10.1109/72.279181.

Brody et al., 2014. Herb Brody, Michelle Grayson, Alisdair Macdonald et al. *Lung Cancer*. Nature Outlook. doi: https://doi.org/10.1038/513S1a.

Büttner et al., 2017. Juliane Büttner, Annika Lehmann, Frederick Klauschen et al. *Influence of mucinous and necrotic tissue in colorectal cancer samples on KRAS mutation analysis*. Pathology Research and Practice. doi: 10.1016/j.prp.2017.04.028.

Chollet, 2018. Francois Chollet. *Deep Learning With Python*. ISBN 9781617294433.

Coudray et al., 2018. Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos et al. *Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning*. Nature Medicine. doi: 10.1038/s41591-018-0177-5.

Deniz et al., 2018. Erkan Deniz, Abdulkadir Şengür, Zehra Kadiroğlu et al. *Transfer learning based histopathologic image classification for breast cancer detection*. Health Information Science and Systems. doi: 10.1007/s13755-018-0057-x.

Duraiyan et al., 2012. Jeyapradha Duraiyan, Rajeshwar Govindarajan, Karunakaran Kaliyappan and Murugesan Palanisamy. *Applications of immunohistochemistry*. Journal of Pharmacy and Bioallied Sciences. doi: 10.4103/0975-7406.100281.

Ejersbo et al., 2014. Dorthe Ejersbo, Preben Sandahl and Marianne Schou Martiny. *Den cytologiske undersøgelse - fra prøvetagning til arkivering*.

Feldman and Wolfe, 2014. Ada T. Feldman and Delia Wolfe. Tissue Processing and Hematoxylin and Eosin Staining. In *Histopathology*. 2014. ISBN 9781493910496.

Godin et al., 2018. Fréderic Godin, Jonas Degrave, Joni Dambre et al. *Dual Rectified Linear Units (DReLUs): A replacement for tanh activation functions in Quasi-Recurrent Neural Networks*. Pattern Recognition Letters. doi: 10.1016/j.patrec.2018.09.006.

Hinton et al., 2012. Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky et al. *Improving neural networks by preventing co-adaptation of feature detectors*.

Hou et al. Le Hou, Dimitris Samaras, Tahsin M Kurc et al. *Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification.* Proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016. doi: 10.1109/CVPR.2016.266.

Houlihan and Tyson, **2012**. Nancy G. Houlihan and Leslie B. Tyson. *Lung Cancer.* ISBN 9781935864103.

Ioffe and Szegedy, **2015**. Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.* doi: 10.1016/j.molstruc.2016.12.061.

Kalyuzhny, **2016**. Alexander E Kalyuzhny. *Immunohistochemistry - Essential Elements and Beyond.* ISBN 9783319308913.

Khosravi et al., **2017**. Pegah Khosravi, Ehsan Kazemi, Marcin Imielinski et al. *Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images.* EBioMedicine. doi: 10.1016/j.ebiom.2017.12.026.

Krizhevsky et al., **2012**. Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. *1 ImageNet Classification with Deep Convolutional Neural Networks.* Advances In Neural Information Processing Systems.

Lecun et al., **2015**. Yann Lecun, Yoshua Bengio and Geoffrey Hinton. *Deep learning.* Nature. doi: 10.1038/nature14539.

Llewellyn, **2009**. B. D. Llewellyn. *Nuclear staining with alum hematoxylin.* Biotechnic and Histochemistry. doi: 10.1080/10520290903052899.

Lyons, **2019**. Stephen Lyons. *A cancer researcher reflects on the evolution of lung cancer therapies.*

Marinelli et al., **2007**. Robert J. Marinelli, Kelli Montgomery, Chih Long Liu et al. *The Stanford Tissue Microarray Database.* Nucleic Acids Research. doi: 10.1093/nar/gkm861.

Martini et al., **2012**. Frederic H. Martini, Judi L. Nath and Edwin F. Bartholomew. *Fundamentals of Anatomy and Physiology.* Pearson, 9th edition. ISBN 978-0-321-70933-2.

McCance et al., **2010**. Kathryn L. McCance, Sue E. Huether, Valentina L. Brashers et al. *Pathophysiology - The Biologic Basis for Disease in Adults and Children.* Elsevier. ISBN 9780323065849.

Mishra et al., **2017**. Rashika Mishra, Ovidiu Daescu, Patrick Leavey et al. *Bioinformatics Research and Applications.* doi: 10.1007/978-3-319-59575-7.

Moreira and Saqi, **2015**. Andre Luis Moreira and Anjali Saqi. *Diagnosing Non-small Cell Carcinoma in Small Biopsy and Cytology.* ISBN 9781493916061.

NCI, **2019a**. NCI. *The NCI's Genomic Data Commons - About TCGA.* URL `https://cancergenome.nih.gov/abouttcga`.

NCI, **2019b**. NCI. *The NCI's Genomic Data Commons - About the Data.* URL `https://gdc.cancer.gov/about-data`.

Ngo et al., 2016. Tuan Anh Ngo, Zhi Lu and Gustavo Carneiro. *Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance*. Medical Image Analysis. doi: 10.1016/j.media.2016.05.009.

Pang et al., 2010. Baochuan Pang, Yi Zhang, Qianqing Chen et al. *Cell nucleus segmentation in color histopathological imagery using convolutional networks*. 2010 Chinese Conference on Pattern Recognition, CCPR 2010 - Proceedings. doi: 10.1109/CCPR.2010.5659313.

Pecorino, 2012. Lauren Pecorino. *Molecular Biology of Cancer*. Oxford University Press, 3rd edition. ISBN 978–0–19–957717–0.

Philipp et al., 2018. George Philipp, Dawn Song and Jaime G Carbonell. *The exploding gradient problem demystifi ed - definition, prevalence, impact, origin, tradeoffs, and solutions*.

Qidwai et al., 2014. Kiran Qidwai, Michelle Afkhami and Christina E. Day. The Pathologist's Guide to Fixatives. In *Histopathology*. 2014. ISBN 9781493910496.

Ramentol et al., 2011. Enislay Ramentol, Yailé Caballero, Rafael Bello et al. *SMOTE-RSB *: A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory*. Knowledge and Information Systems. doi: 10.1007/s10115-011-0465-6.

Roser and Ritchie, 2019. Max Roser and Hannah Ritchie. *Cancer*. Our World in Data.

Salvi and Molinari, 2018. Massimo Salvi and Filippo Molinari. *Multi-tissue and multi-scale approach for nuclei segmentation in H&E stained images*. BioMedical Engineering Online. doi: 10.1186/s12938-018-0518-0.

Soria et al., 2009. Emilio Soria, Jose David Martin-Guerrero, Marcelino Martinez et al. *Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques*. Information Science Reference. ISBN 1605667668.

Statistik, 2016. Danmarks Statistik. *No Title*. URL `http://statistikbanken.dk/statbank5a/selectvarval/define.asp?PLanguage=0{&}subword=tabsel{&}MainTable=DOD1{&}PXSId=213775{&}tablestyle={&}ST=SD{&}buttons=0`.

Statistik, 2018. Danmarks Statistik. *danmarksStatistik2018.pdf*. URL `https://www.dst.dk/da/Statistik/bagtal/2018/2018-10-24-Danskere-doer-oftere-af-kraeft-end-vores-naboer`.

Yu et al., 2016. Kun Hsing Yu, Ce Zhang, Gerald J. Berry et al. *Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features*. Nature Communications. doi: 10.1038/ncomms12474.

Yu et al., 2018. X. Yu, H. Zheng, C. Liu et al. *Classify epithelium-stroma in histopathological images based on deep transferable network*. Journal of Microscopy. doi: 10.1111/jmi.12705.

Zhao and Liu, 2019. Hui-huang Zhao and Han Liu. *Multiple classifiers fusion and CNN feature extraction for handwritten digits recognition*. Granular Computing. doi: 10.1007/s41066-019-00158-6.

Magnification of the Acquired Data A

This appendix depicts the distribution of data recorded with magnification levels of x20 or x40 in regards to the lung cancer stage and the two different datasets the data would be included in.

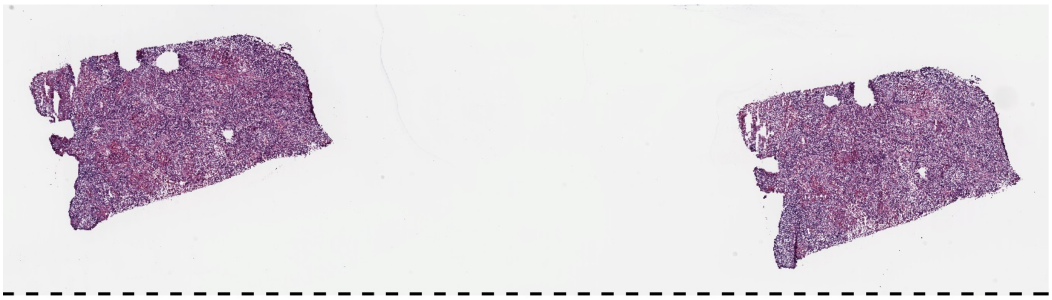
	Magification x40		**Magnification x20**	
	All data	Tumor nuclei ≥70%	All data	Tumor nuclei ≥70%
Stage I	302 (55.0%)	59 (55.1%)	566 (52.7%)	72 (48.0%)
Stage II	157 (28.6%)	28 (26.2%)	259 (24.1%)	35 (23.3%)
Stage III	74 (13.5%)	17 (15.9%)	210 (19.6%)	34 (22.7%)
Stage IV	16 (2.9%)	3 (2.8%)	39 (3.6%)	9 (6.0%)
Total	549	107	1074	150

Table A.1: *Number of WSIs which would be included in the two different dataset with magnification levels of x20 and x40 respectively.*

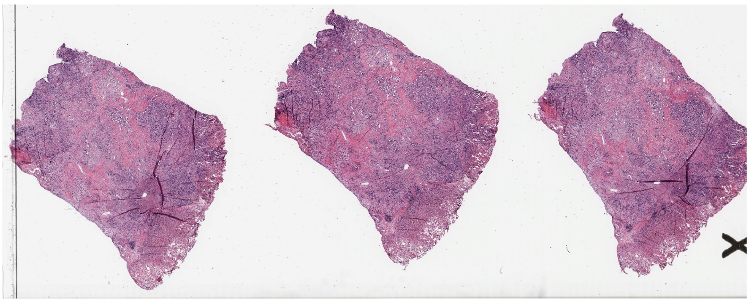
The WSIs acquired from TCGA were recorded with magnification levels of x20 or x40. To ensure homogeneity in the input images, only WSIs acquired with one of these magnification levels used in the project. It was assumed that shape and contrast features in WSIs with a higher magnification level would be more distinct than with a lower magnification level.
If data with a magnification level of x20 were used, the dataset containing all the data would be almost twice as large as the dataset with all data with a magnification level of x40. However, for the dataset containing WSIs with a tumor nuclei percentage of 70% or more, the difference between the available data recorded with a magnification level of x20 and x40 was less distinct.

Whole Slide Images with Ink Stains B

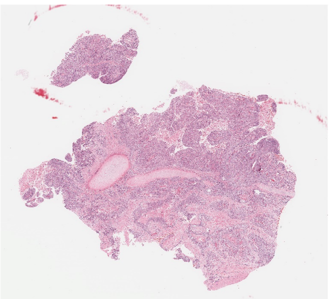
This appendix illustrates the WSIs which have been excluded due to ink stains. All of these WSIs have been recorded with a magnification level of x40. The images are illustrated in the figures below.



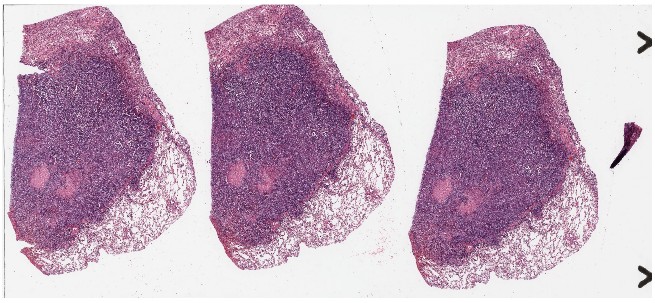
(a)



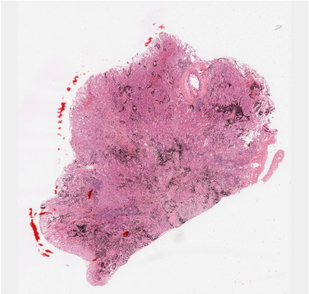
(b)



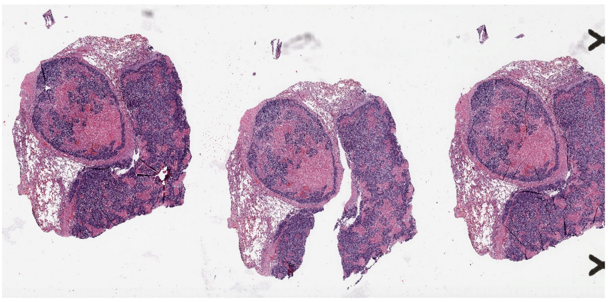
(c)



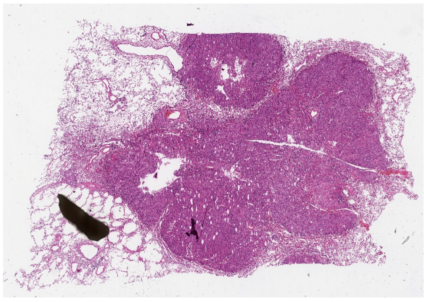
(d)



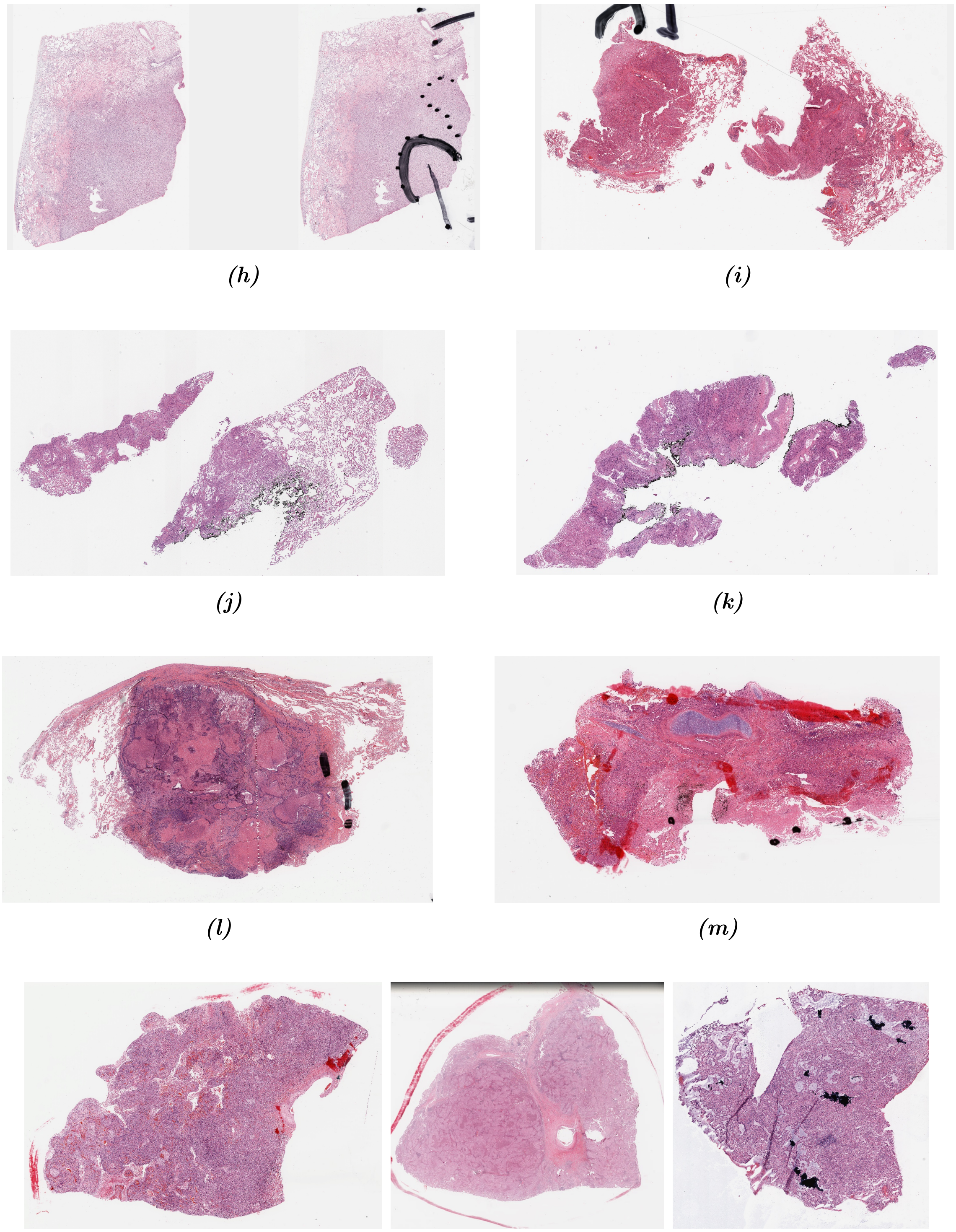
(e)



(f)



(g)



(h)

(i)

(j)

(k)

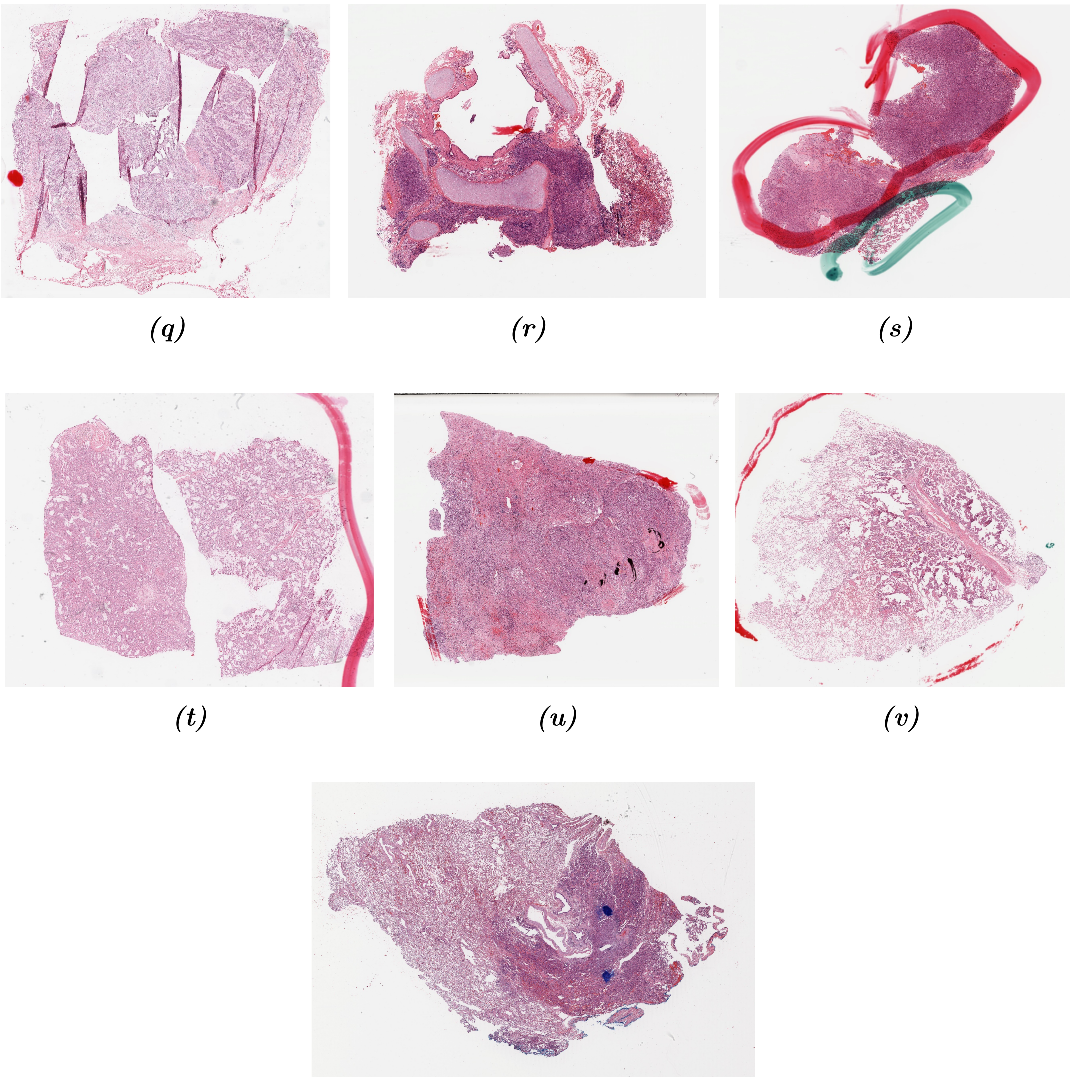
(l)

(m)

(n)

(o)

(p)



(q)

(r)

(s)

(t)

(u)

(v)

(w)

Furthermore, the lung cancer type and stage of the individual images with ink stains is depicted in table B.1.

Image number	Cancer stage	cancer type
a	I	scc
b	I	Adenocarcinoma
c	II	SCC
d	I	Adenocarcinoma
e	I	Adenocarcinoma
f	II	SCC
g	IV	SCC
h	III	Adenocarcinoma
i	not reported	Adenocarcinoma
j	I	Adenocarcinoma
k	I	SCC
l	I	SCC
m	II	SCC
n	II	SCC
o	II	SCC
p	II	Adenocarcinoma
q	I	SCC
r	I	SCC
s	III	Adenocarcinoma
t	III	Adenocarcinoma
u	II	Adenocarcinoma
v	I	Adenocarcinoma
w	I	Adenocarcinoma

Table B.1: *Image number corresponding to the images illustrated above with the associated lung cancer stage and type.*