University of
Stavanger

Faculty of Science and Technology

MASTER'S THESIS

Study program/Specialization:

Cybernetics and Signal Processing

Spring semester, 2023

Open / ~~Restricted access~~

Writers:
Marie Bø-Sande

Edvin Benjaminsen

. . . .

. . . .
(Writer's signatures)

Faculty supervisor(s): **Kjersti Engan**

Thesis title: **Diagnosis, Localization, and Prognosis of Melanoma in WSIs with a Complete Pipeline by Digital Pathology**

Credits: **30p**

Key words:
Melanoma,
Machine learning,
Computational pathology,
Diagnosis,
Prognosis

Pages:

+ enclosure:

Stavanger, 2023-06-15

Contents

Abstract

The most dangerous and aggressive form of skin cancer is melanoma, responsible for 90% of skin cancer mortality. Early detection of melanoma plays a crucial role in the prognostic outcome. The diagnostic has to be performed by a pathologist, which is time-consuming. The recent increase in melanoma incidents indicates the growing demand for a more efficient diagnostic process.

This thesis's main objective is to develop a pipeline utilizing two independent pre-trained models built on the VGG16 architecture. This pipeline consists of a diagnostic and a prognostic model. The diagnosis model is responsible for localizing malignant patches in WSIs and giving a patient-level diagnosis. The prognosis model uses the output from the diagnosis model to provide a patient-level prognosis. The complete pipeline provides both a prognostic and a diagnostic tool, which can be used by a pathologist when evaluating Whole Slide Images (WSIs). A total of 243 WSIs were provided by Stavanger University Hospital for this thesis. All have been provided a patient-level label. 203 of the WSIs were used for parameter tuning and 40 were used for testing.

The diagnosis model performed with a 100% accuracy when evaluated on the original test, which was provided together with the training set. The prognosis model also performed well on the original dataset, with an accuracy of 0.7885. The model's capability to predict diagnosis and prognosis decreases significantly when being introduced to the new dataset.

In addition to developing the pipeline, some parameters for the diagosis model was found using a ROC cuve. By using the new parameters for the diagnosis model on the validation set, the performance of the diagnosis model increased when using the test set. The prognosis model performed relatively equally in all experiments. A correlation between the number of patches in a WSI and the number of patches predicted malignant was discovered and counteracted by altering the patient-level threshold calculation method (MT_{rate}).

Acknowledgments

This thesis marks the end of our master's degree in Robotics and Signal Processing at the University of Stavanger. First and foremost we would like to express our sincere gratitude towards our supervisor, Kjersti Engan. We are appreciative of for her guidance and support during this thesis.

Additionally, we want to thank PhD student Saul Fuster Navarro and PhD student Neel Kanwal for helping us whenever we needed them.

Lastly, we want to thank our co-supervisor Emiel Janssen for giving us an informative tour of the pathology section at SUH (Stavanger University Hospital), and for providing us with the data we needed for this thesis.

Chapter 1

Introduction

Globally, skin cancers are the most prevalent type of cancer, with an estimated 1.5 million new cases reported in 2020 [34]. Among these, melanoma was diagnosed in approximately 325,000 cases, leading to 57,000 fatalities. The incidence rates of melanoma exhibit significant geographical disparities, with certain regions reporting higher rates than others. Norway especially saw a significant increase in skin cancer between the years 1953 and 2022, with 2,911 people new cases only in 2022. That is an increase of 468 from the previous year, which is an increase of approximately 4.2% [20]. The growing number of recorded skin cancer incidents makes this one of the fastest-growing cancer types in Norway.

There are two main types of skin cancer: melanoma and non-melanoma. Melanoma is the most aggressive type of skin cancer with the ability to spread, which makes early diagnosis important. According to World Cancer Research Foundation, Norway ranks as the 5th highest-rated country for melanoma cancer incidents [28]. The melanoma cancer incidents in Norway are depicted in Figure 1.1, and melanoma incidents for men and women in Norway are shown individually in Figure 1.2. The graphs show a concerning increase in melanoma cases, especially in the last few years. Apart from the overall rise, differences in cancer incidence based on geography have also been documented. Rogaland has been found to have the highest age-adjusted cancer rates [8]. This highlights the importance of melanoma diagnosis and treatment in the exposed areas, as early detection is the most critical predictor of melanoma survival [9][17].

The rise in melanoma cases has placed considerable stress on pathology departments, often leading to extended waiting periods for patients awaiting a diagnosis. Research published in the Journal of the American Academy of Dermatology indicates that early treatment could increase the patient's chances of survival, particularly in the early stage of melanoma [30]. As a countermeasure, the fusion of artificial intelligence (AI) and ma-

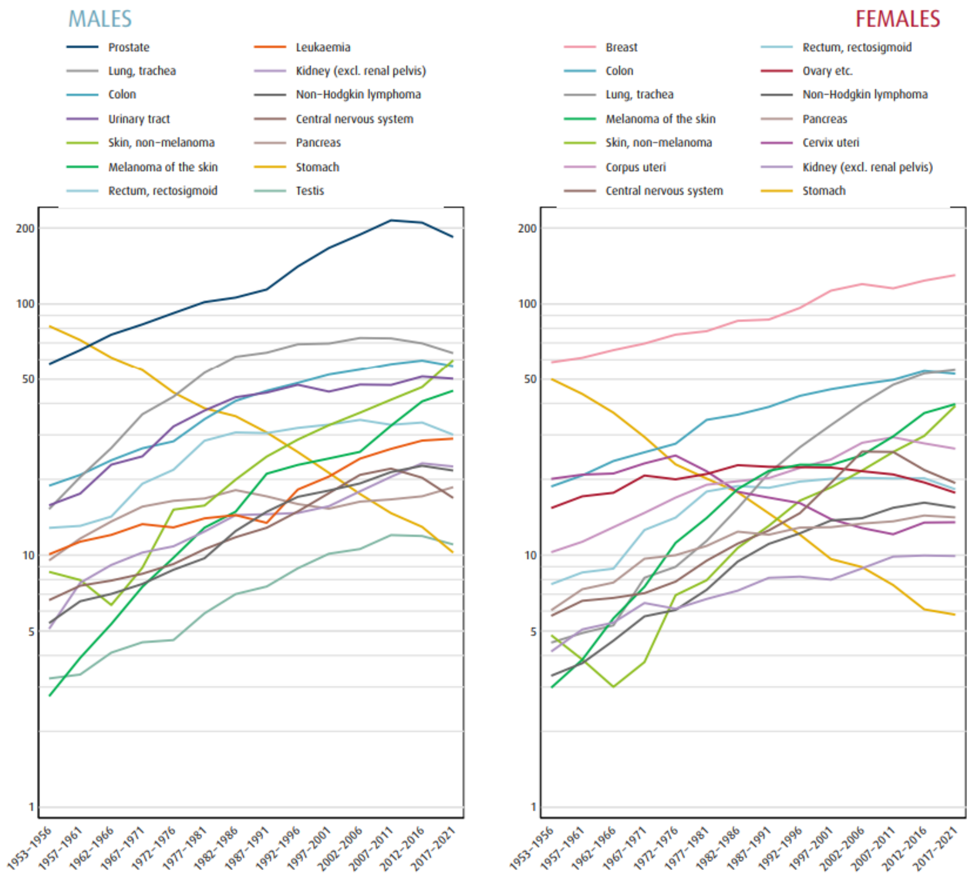


Figure 1.1: Graph displaying incidence rates for selected cancers from 1953–2021. The graph displays a considerable rise in melanoma (dark green) and non-melanoma (light green) skin cancer cases since 1953. //textitFigure is sourced unaltered from the *Cancer Registry of Norway (Cancer in Norway 2021)* [8].

chine learning methodologies in the field of pathology has emerged as a promising strategy to assist pathologists. Specifically, advanced techniques such as deep learning (DL) and convolutional neural networks (CNNs) have shown significant potential in improving the accuracy and efficiency of diagnoses [6].

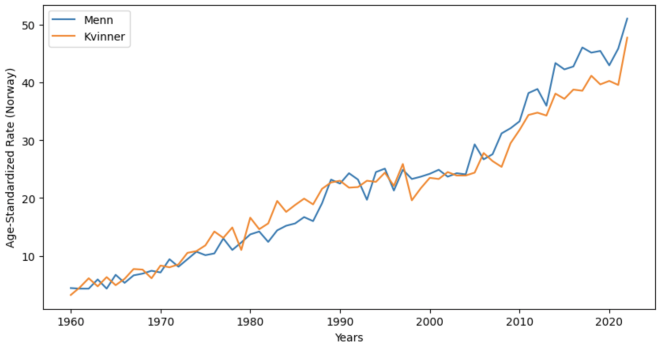


Figure 1.2: Time series data for melanoma cancer in age-standardized (as per Norwegian standards) incidence rates during the period 1960–2022. The graph shows a significant rise in melanoma rates for skin cancer cases since 1960. The figure was created using data acquired from kreftregisteret.no [20].

1.1 Motivation and Objective

The escalating incidence of melanoma worldwide necessitates innovative solutions to support pathologists in their diagnostic and prognostic tasks. This thesis is motivated by the need to expedite diagnosis, enhance accuracy, alleviate the workload on pathology departments, reduce patient waiting time, and potentially improve survival rates through early detection and treatment [30].

To address these challenges, this work aims to develop a pipeline that integrates two pre-trained Deep Neural Network (DNN) models. The first model focuses on the detection and localization of melanoma, while the second provides a prognosis for identified melanoma areas. By leveraging machine learning methodologies, we aim to provide a complete pipeline to predict and locate melanoma in Whole-Slice Images (WSI) and give a patient-level diagnosis and prognosis.

Each DNN model will be individually evaluated before integrating them into a comprehensive pipeline. The performance of this integrated system will then be assessed using WSI, thereby ensuring its effectiveness in real-world scenarios.

The increasing number of melanoma incidents highlights the need for assisting pathologists in their work. The primary motivation of this thesis is to design a pipeline that integrates machine learning methodologies to expedite and enhance the accuracy of melanoma diagnosis and prognosis. This pipeline aims to alleviate the workload of pathologists, reduce patient waiting time, and potentially improve survival rates through early detection and treatment [30]. By leveraging the capabilities of pre-trained DNN models, this pipeline is expected to provide valuable insights into the detection and localization of melanoma, as well as offer a prognosis on the identified areas. The ultimate goal is to assist pathologists in making more effective and accurate decisions, thereby improving patient outcomes.

Chapter 2

Background Theory

This thesis combines the field of machine learning and pathology. To provide a comprehensive understanding of the nature of this work, the following chapter will introduce the main background theory from both fields. The medical background will first be presented, followed by the technical background.

2.1 Medical Background

This section will provide a medical context for this thesis. Firstly the skin layers will be explained, followed by some main concepts in pathology. Lastly, Whole-Slide Images and melanoma are presented.

2.1.1 Composition of the Skin

A fundamental understanding of skin composition and structure is crucial for grasping the development and progression of melanoma skin cancer. Hence, before delving into the intricacies of melanoma, we will first examine the layers of the skin.

The skin is composed of three primary layers: the epidermis, dermis, and hypodermis [16]. Figure 2.1 depicts the skin composition, including the epidermis, dermis, and hypodermis layers. Each layer has a unique structure and function, which will be discussed briefly in the following sections.

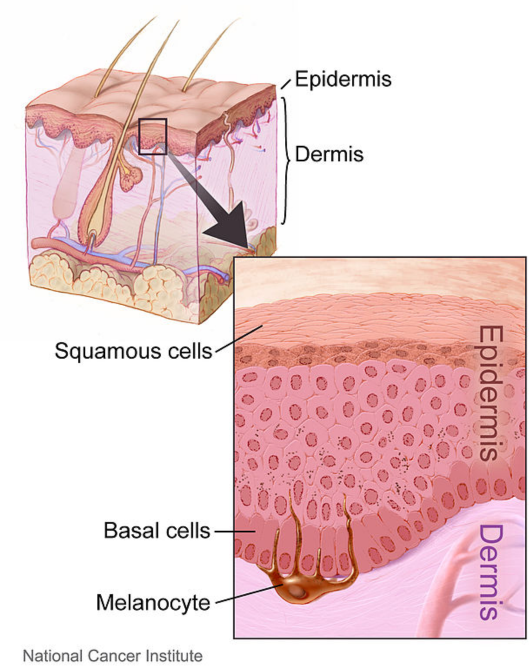


Figure 2.1: This figure illustrates the different layers of the skin, with a particular focus on the location of melanocytes. The illustration demonstrates how melanocytes are situated within the stratum basale of the epidermis. Melanocytes produce melanin, which contributes to skin pigmentation. *This figure was obtained from Wikimedia Commons. The original illustration was created by Don Bliss and is in the public domain.* Source: Wikimedia Commons.

Epidermis

The epidermis is the outermost layer of the skin and acts as a protective barrier against external factors. It mainly consists of keratinocytes, which are cells that produce keratin. Melanocytes are found in the stratum basale, the deepest layer of the epidermis. Figure 2.1 demonstrates how melanocytes are situated within the stratum basale of the epidermis. Melanocytes' main function is to produce melanin, the pigment responsible for the darkening of the skin. When the skin is exposed to UVB (ultraviolet B) radiation, melanocytes increase their production of melanin as a response[35]. Even though melanocytes play an important role in protecting us from radiation, it also plays a critical role in the development of melanoma [22], as discussed later.

Dermis

The dermis, situated beneath the epidermis, is a robust fibrous layer providing strength and elasticity to the skin. It supports the epidermis and houses sensory receptors for touch, temperature, and pain. The dermis comprises collagen, elastin fibers, follicles, glands, and nerves, among others [7]. Melanoma originates from melanocytes in the stratum layer (Figure 2.1), located above the dermis. If melanoma invades the dermal layer containing blood and lymph vessels, it can potentially facilitate cancer spread [19].

Hypodermis

The hypodermis is the inner layer of the skin. It mainly consists of adipose tissue (fat cells), which provides insulation, cushioning, and energy storage. The hypodermis anchors the skin to the underlying muscles and bones. Melanoma can potentially invade the hypodermis tissue layer but is more likely to grow and spread in the dermis and other layers [33].

2.1.2 Pathology

Pathology is the study of diseases, their causes, mechanisms, and effects on the body [27]. The pathology field plays a crucial role in evaluating melanoma. Tissue samples in the form of WSIs are examined to identify cancerous cells and determine the stage of the disease. The process of diagnosing a patient involves several steps, from consulting the mole at the doctor's office to digitizing the stained tissue sample for further analysis. The process of handling a tissue sample and creating a WSI are illustrated in Figure 2.2. Every step involved in this process will be briefly explained in the subsequent sections.

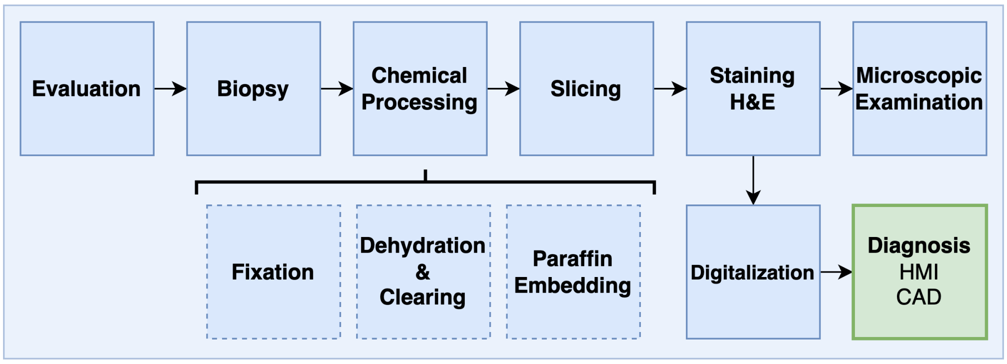


Figure 2.2: Overview of the steps involved in the management of a tissue sample, from initial biopsy to final diagnosis. The green box indicates the stage of the process in which this thesis takes place.

Evaluation

The first step in this process is visually inspecting the mole. When visiting a healthcare provider with concerns about a mole or lesion, the medical professional will examine its visual characteristics and assess whether it is or may become cancerous (melanoma). A common way to assess the mole is to use an acronym called ABCDE, which serves as a standard tool for doctors to determine if the mole or nevus should be removed. ABCDE stands for Asymmetry, Border irregularity, Color variation, Diameter, and Evolving appearance. A more deliberate explanation of each word is presented in Table 2.1. If any of these features are present and raise suspicion, the doctor may recommend surgically removing the mole for further examination. It is worth mentioning that other factors may also play a crucial role in the evaluation. These factors are presented in the following list [14][22]:

- Patient history: A family history of melanoma or other skin cancers can increase the risk of developing melanoma.

- Sun exposure: Excessive sun exposure, particularly during childhood, can increase the risk of developing skin cancer.
- Skin type: People with fair skin, light hair, and light eyes are at a higher risk of developing skin cancer.
- Number of moles: A large number of moles (more than 50) can increase the risk of melanoma.
- Location of moles: Moles in areas frequently exposed to the sun, such as the face, neck, and arms, may be more likely to become cancerous.
- Immune status: Individuals with weakened immune systems, such as HIV/AIDS or organ transplant recipients, are at a higher risk of developing skin cancer.
- Age: The risk of developing skin cancer increases with age.

Letter	**Explanation**
A: Asymmetry	One half of the mole does not match the other in size, shape, color, or thickness. If you were to draw a line through the mole, the two halves would not mirror each other.
B: Border	The mole's edges are irregular, ragged, notched, or blurred. The pigment may also spread into the surrounding skin.
C: Color	The color is inconsistent and may include shades of black, brown, and tan. There may also be white, gray, red, pink, or blue areas.
D: Diameter	The diameter of the mole is larger than 6mm.
E: Evolution	The mole changes in size, shape, or color, particularly if it changes quickly or drastically. Other changes to watch for include new symptoms like bleeding, itching, or crusting.

Table 2.1: The ABCDE acronym for visual melanoma detection was developed to provide a straightforward method for both doctors and patients to assess the atypical features of nevi. The ABCD variant of this acronym was first introduced by R. Friedman et al. in 1985 [11] and later modified to include E: Evolving [1].

Biopsy

If the clinician determines that a mole needs further examination, they will perform a biopsy. This involves surgically removing the mole and neighboring tissue and sending it to a pathology laboratory for analysis [15].

Chemical Processing

Once the mole is surgically removed, it is sent to a pathology laboratory for further analysis. The tissue sample undergoes a series of preprocessing steps, including fixation (to preserve its structure), dehydration, clearing (removal of water and solvents), and embedding in paraffin wax [15].

Slicing

The embedded tissue sample is sliced into thin sections, typically around 4-5 micrometers thick, using a microtome. These thin sections allow for better visualization of cellular structures and facilitate the staining process [15].

Staining

To visualize cellular structures and identify cancerous cells, the tissue sections are stained with Hematoxylin and Eosin (H&E) stains. Hematoxylin stains cell nuclei blue, while Eosin stains cytoplasm and extracellular matrix pink. This staining technique allows pathologists to differentiate between various cell types and assess the presence of cancer cells in the tissue sample [15].

Digitization

After staining, the tissue sections are mounted on glass slides and coverslipped. The tissue sections can then either be examined under a microscope or digitized into an image of the whole slide, hence Whole-Slide Image (WSI). For digitization of the glass slides, they are imaged by a whole slide scanner, such as the Hamamatsu NanoZoomer S60, at 40x magnification. The scanner captures high-resolution images of the entire tissue section, creating a WSI which can be viewed and analyzed on a computer using an appropriate display [15].

Whole-Slide Images (WSI)

WSI technology has become increasingly important in digital pathology. It offers various benefits, including the potential to improve diagnostic accuracy and enhance collaboration among pathologists. Additionally, it allows for the integration of computational tools for image analysis, streamlines workflow, reduces physical storage space requirements, and facilitates accessibility and education by providing easier access to a wide range of cases. An example of a WSI is shown in Figure 2.3. The size of a WSI can vary depending on the scanner's resolution, magnification level, and the size of the tissue section. However, it is common for WSIs to be several gigabytes in size due to their high resolution.

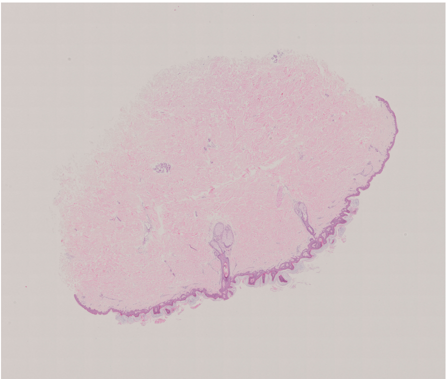


Figure 2.3: Example of a WSI showcasing a high-resolution digital scan of a tissue sample. This WSI was scanned using Hamamatsu NanoZoomer S60 at 40x magnification and stained using H&E stain. This WSI is from a dataset provided by Stavanger University Hospital.

Artifacts in WSI

Throughout the tissue processing and digitization steps, various artifacts may be introduced. Common artifacts include tissue folds, air bubbles, dust particles, and uneven staining. These artifacts can potentially hinder the accurate diagnosis and prognosis of melanoma using WSIs. This challenge is particularly pronounced when employing AI models, as they may misinterpret the artifacts as features from other tissue types [18].

Digital Pathology

The traditional method of melanoma diagnosis involves a pathologist examining stained tissue sections under a microscope. In recent years, more clinics have transitioned to digital pathology [25]. This new method primarily employs two approaches:

- **Human-Machine Interface (HMI)**: Pathologists utilize specialized software tools to inspect digital images of tissue samples, replacing traditional microscope examination.

- **Computer-Aided Pathology**: Digital tools can assist pathologists by analyzing digital images, effectively aiding in identifying and categorizing cancerous cells. This can enhance the diagnostic process, combining algorithms with human expertise [25]. This thesis's objective is to develop such a tool by using machine learning.

2.1.3 Melanoma

Recall that melanoma is a type of skin cancer originating from melanocytes [22]. It develops when the DNA (Deoxyribonucleic Acid) of melanocytes becomes damaged, often due to excessive exposure to UV radiation. This damage can cause the melanocytes to grow uncontrollably and form a malignant tumor.

Stages of Melanoma

The progression of melanoma can be divided into several stages, often starting from benign nevi, developing into dysplastic nevi, followed by the two growth states (Radial-Growth Phase and Vertical-Growth Phase), and finally, the metastatic state [22]. Figure 2.4 illustrates the following stages, which will be explained subsequently:

1. **Benign Nevus**: The first stage is called a benign nevus, better known as a mole, and is a limited growth of melanocytes in the epidermis. Moles are benign, and most do not progress to cancer (melanoma).

2. **Dysplastic Nevus**: The next stage is called dysplastic nevus. It can develop within a preexisting benign nevus or emerge in a new location. It is not cancerous but carries a higher risk of progressing into melanoma. A dysplastic nevus is characterized by atypical features such as irregular borders, diverse colors, or increased size. Medical professionals examine these visual characteristics to assess whether a mole is or may become cancerous (melanoma) [14].

3. **Radial-Growth Phase**: In this phase, malignant cells multiply within the epidermis in a phase known as radial growth. At this point, the tumor is limited to the epidermis and has not yet invaded the dermis. This stage is characterized by unrestrained growth and division of melanoma cells, resulting in exponential cell increase. Even as the tumor expands laterally along the epidermal layer, it does not penetrate deeper skin layers. Melanoma cells in this phase exhibit decreased differentiation, losing their unique roles and functions as cells, which paves the way for quicker growth and division. However, they can't yet grow in soft agar, limiting their growth to solid surfaces. This phase's tumor originates from a single mutated melanocyte that has undergone clonal proliferation, resulting in a mass of cells genetically identical to the mutated cell [10][14].

4. **Vertical-Growth Phase**: Eventually, most melanomas progress to the vertical-growth phase, where the malignant cells penetrate the basal layer and invade the dermis, growing vertically and deeper into the skin. The dermis contains blood and lymph vessels, which can facilitate the spread of malignant cells. At this stage,

melanoma cells can grow in soft agar, which increases the aggressiveness of the cancer. The cancer cells continue to proliferate rapidly and may gain additional genetic mutations that further drive tumor growth and invasion ability.

5. **Metastatic Melanoma**: Following the infiltration of blood and lymph vessels during the vertical-growth phase, malignant cells can detach from the primary tumor, travel through the blood or lymphatic system, and form new metastatic tumors in other tissues, most commonly the bone, brain, liver, lung, skin, and muscle [21]. This process is referred to as metastatic melanoma. Metastatic tumors originate from the primary site and consist of the same type of cancer cells as the primary tumor, indicating that cancer cells have disseminated from their initial location to distant regions. This stage is the most challenging to treat and has the poorest prognosis.[10][14].

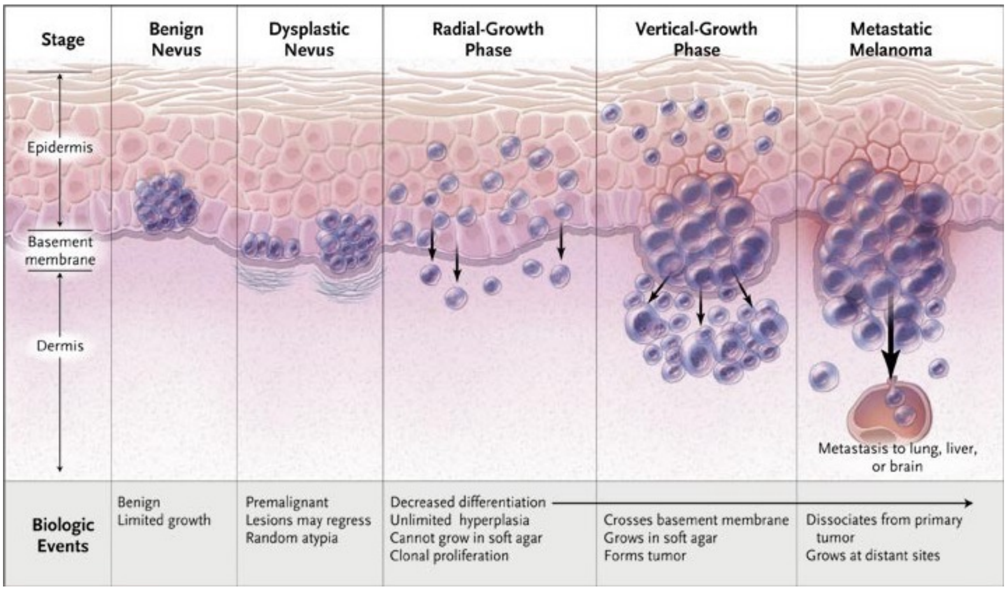


Figure 2.4: Visual representation of melanoma progression starting from benign nevus to metastatic melanoma. The diagram, moving from left to right, depicts increasing severity. *(Melanoma progression diagram, (reproduced with permission from Miller and Mihm, "Melanoma," 355: 51–65. Copyright 2020 Massachusetts Medical Society)*

Prognosis

In addition to giving a diagnosis, the likely course or outcome of a disease can also be given. This is referred to as the prognosis. It is a prediction made by medical professionals based on current knowledge and understanding of the disease, as well as the patient's specific condition and overall health status. In the context of melanoma, a bad prognosis usually refers to situations where cancerous cells have advanced or spread beyond their original location in the skin. This is referred to as metastasis [29], and it is the main contributor to the high mortality rate associated with melanoma [31].

2.2 Technical Background

This section provides a brief explanation of the main concepts and techniques related to digital images, image processing, and artificial intelligence, with a particular focus on deep learning methods. This knowledge is crucial for understanding the development and implementation of the proposed pipeline for melanoma prediction and prognosis using pre-trained convolutional neural networks (CNNs).

2.2.1 Digital Images

Digital images are a representation of visual information in the form of discrete numerical values called pixels. Each pixel corresponds to a specific location in the image. An associated intensity value determines its color or brightness. Digital images can be categorized into grayscale, binary, and color images based on the number of color channels.

Grayscale images represent the intensity of gray levels, where each pixel value ranges from 0 (black) to 255 (white). Binary images contain only two possible pixel values: 0 (black) and 1 (white). These images are often used for representing shapes or outlines of objects. Color images store more complex visual information by combining multiple color channels. RGB (Red, Green, Blue) color space is the most frequently common color space in digital images. Each pixel contains three values, red, green, and blue.

HSV is also a common color format, representing colors in terms of their HSV components (Hue, Saturation, Value). The term "hue" refers to the color type and has a range from 0-360 degrees. Saturation is the intensity of the color type and ranges from 1-100%. The last component, value, refers to the brightness of the color and has a range from 0-100%. This color representation is more closely related to human perception of color than the RGB color space. The HSV foreman is also promising for color thresholding in many scenarios [12].

2.2.2 Image Processing

When working with digital images, such as WSIs, some image modifications are often wanted or even necessary. This section will present the image processing technics used in this thesis.

Segmentation

In the context of WSI, segmentation plays a crucial role in separating different components, such as background regions, from the tissue. Due to the vast number of pixels in WSIs, segmentation is particularly important for identifying regions containing relevant information. One effective approach to achieve this in H&E stained WSIs is through color thresholding. In H&E stained WSIs, various tissue components exhibit distinct hues and saturations due to the staining process. By converting the RGB image to HSV and applying appropriate thresholds on hue and saturation channels, it becomes easier to differentiate between tissue components and separate them from the background, leading to more accurate segmentation and improved analysis of WSIs in digital pathology applications.

Patch extraction

As discussed earlier, a WSI is typically a gigapixel image. It is unrealistic to expect a DNN model to handle this many pixels as input. To address this issue, a standard technique involves first identifying the foreground and then dividing the ROI into smaller parts called patches. This technic is commonly known as patch extraction.

Data Augmentation in WSI

Data augmentation is a technique employed to enhance the diversity and size of training datasets by applying various transformations to the original images. In the context of WSI, augmentation processes such as rotation, flipping, scaling, and color adjustments can be applied to create new variations of the original images. These augmented images contribute to improving the generalization capabilities of AI models by exposing them to a broader range of image variations during training. Specifically, for WSI, data augmentation can be performed on individual patches, introducing variance within the training set and further strengthening the model's ability to recognize diverse features and patterns.

Morphological Operations

Morphological operations are image processing techniques that modify the shape and structure of objects in an image. Two common morphological operations used in WSI are erosion and dilation.
Erosion is a process that removes pixels from the boundaries of objects, effectively

shrinking them. This operation can help separate closely connected objects or remove small artifacts from the image [13].
Dilation adds pixels to the boundaries of objects, causing them to expand. Dilation can fill small gaps or holes within objects and connect disjointed components [13].

Both erosion and dilation can be applied sequentially in various combinations to refine the segmentation results and enhance the quality of WSI images before feeding them into AI models.

Opening and Closing is used in image processing to manipulate shapes and contours of objects and is a two-step process of combining erosion and dilation. The opening involves a process of erosion followed by dilation. The goal is primarily to smooth the contour of an object. The closing, on the other hand, is the reverse of the opening and involves dilation followed by erosion. Closing is used to close gaps, connect fragmented regions, and create a more complete and continuous contour [13].

2.2.3 Introduction to Artificial Intelligence

Artificial intelligence (AI) describes a system that can function intelligently and independently. It aims to create machines and systems capable of performing tasks that typically require human intelligence. Machine learning (ML) and deep learning (DL) are subsets of AI that focus on enabling machines to learn from data and improve their performance over time [31]. Figure 2.5 illustrates the hierarchical relationship between AI, ML, and DL. This chapter will delve into the artificial intelligence aspect, focusing on machine learning and deep learning techniques that can be employed to analyze the processed whole-slide images for melanoma detection and prognosis prediction.

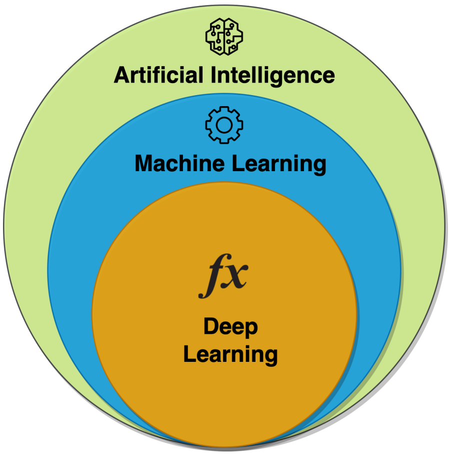


Figure 2.5: The figure illustrates the hierarchical relationship between Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL). AI, represented as the largest circle, encompasses all facets of mimicking human intelligence. Within AI is ML, a subset dedicated to statistical techniques for enabling machines to improve tasks with experience. Nested within ML is DL, a specialized subset of ML that uses artificial neural networks with multiple layers (or 'depth') to model and understand complex patterns. Figure inspired by [31].

Machine learning is a subset of AI that focuses on developing algorithms and statistical models that enable machines to learn from data and improve their performance over time. In this context, learning refers to the ability of a machine to recognize patterns in data, make predictions, and adapt its behavior based on experience. Machine learning can be broadly categorized into supervised learning, unsupervised learning, and reinforcement learning [31]. Each category will be subsequently explained in the subsequent paragraphs.

Supervised learning involves training a model using labeled data, where each input example is associated with a corresponding output label. The model learns to map inputs to outputs by minimizing the difference between its predictions and the true labels. In the context of this thesis, this could be using annotated WSIs to train a model to recognize features from malignant patches.[31].

Unsupervised learning deals with unlabeled data. The goal is to uncover underlying patterns or structures in the data without any prior knowledge of the labels. [31]. A common example of unsupervised learning is clustering.

Weakly-supervised learning falls somewhere between supervised and unsupervised learning. The model is trained on weakly labeled data, meaning the labels are limited. In the context of this thesis, an example could be having labels only for patients with melanoma without any information about which patches from the WSI that contain melanoma.

Reinforcement learning is a method that trains algorithms to make optimal decisions through interaction with an environment. The algorithm learns to select actions that give the highest cumulative reward [31]. For instance, in the context of diagnosing skin cancer using WSIs and machine learning, reinforcement learning could be used to train an AI model to identify melanoma by analyzing image data. The model would learn from each decision it makes - correct identifications would increase its 'reward', while incorrect ones would decrease it.

2.2.4 Deep Learning

Deep learning (DL) is a category of machine learning that uses artificial neural networks with multiple layers to model and understand complex patterns in the data. These networks consist of connected nodes or neurons organized into layers. DL has been particularly successful in tasks involving large-scale, high-dimensional data, such as image recognition, natural language processing, and speech recognition [24].

Neural Networks

Artificial neural networks (ANNs) are computational models inspired by the structure and function of biological neural networks. ANNs are made up of layers of interconnected nodes or neurons. Each neuron receives input from one or more neurons in the previous layer, processes the input using an activation function, and passes the result to neurons in the next layer. The connections between neurons have associated weights, which determine the strength of the influence between connected neurons. During training, these weights are adjusted to minimize the difference between the network's predictions and the true labels.

Convolutional Neural Networks

Deep learning models known as Convolutional neural networks (CNNs) are created specifically for processing grid-like data, such as images. CNNs are made up of one or more convolutional layers followed by fully connected layers. The input data is subjected to a set of filters by convolutional layers, which helps the network learn local features and patterns in the data. Pooling layers are often used between convolutional layers to reduce the spatial dimensions of the feature maps, which lowers computational complexity and improves translation invariance. CNNs have been highly successful in various computer vision tasks, including image classification, object detection, and semantic segmentation [2].

VGG16 is a deep convolutional neural network developed and used for illustration in Figure 2.6. It was trained on the ImageNet dataset, which contains over 14 million images and 1000 object categories. The VGG16 model consists of 16 layers, including 13 convolutional layers and 3 fully connected layers. It takes an input size of 224x224 pixels. When using a different input size, larger images are often cropped to fit the weights trained on ImageNet. Due to its depth and a large number of parameters, the VGG16 model has been highly successful in various image recognition tasks and is often used as a starting point for transfer learning in computer vision applications [36].

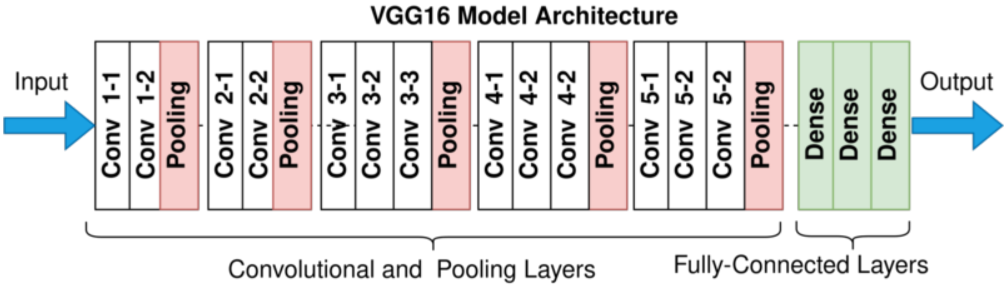


Figure 2.6: Overview of the architecture of VGG16. The network has 16 layers and has an image input size of 224-by-224. The network consists of 16 layers, 13 convolutional and three fully-connected layers. *This figure was obtained from Wikimedia Commons. The original illustration was created by Gorlapraveen and is in the public domain.* Source: Wikimedia Commons.

Transfer Learning

Transfer learning is a technique in which a pre-trained neural network is fine-tuned for a new task or domain [23]. The idea is to leverage the knowledge learned by the network during its initial training on a large dataset to improve its performance on a related but smaller dataset. Transfer learning is particularly useful when the target task has limited labeled data available, as it allows the model to benefit from the more extensive dataset's learned features and representations.

2.2.5 Supervised Learning

In the context of whole-slide image (WSI) analysis for diagnosis and prognosis, supervised learning plays a crucial role in training ML models to recognize patterns and make predictions based on labeled data. This section discusses various aspects of supervised learning relevant to computer-aided pathology.

Training and Validation

Training and validation are essential steps in the process of developing a supervised learning model. During the training phase, the model learns from a labeled dataset, known as the training set. By adjusting its parameters, the model aims to minimize the difference between the model's predictions and the true labels. During this phase, the model learns to recognize patterns and relationships in the data.

Once the training phase is complete, the model's performance is evaluated on a separate labeled dataset, the validation set. This set is not used during training and serves to assess the model's ability to generalize to new, unseen data. The validation phase helps identify potential issues such as overfitting, where the model performs well on the training data but poorly on new data.

After training and validation of the model, the performance needs to be measured on a completely independent and unseen data set, known as the test set. The test set helps to determine how well the model can make accurate predictions and can be used as a checkpoint to ensure that the model has not been overfitted to the training data. This will help to validate the model's generalization ability.

Cross-Validation

Cross-validation is a technique used to assess the performance of a supervised learning model. It is useful in situations where there is limited data, as it effectively utilizes the available data by using it both for training and validation.

A type of cross-validation is k-fold, as displayed in Figure 2.7. It involves dividing the available labeled data into multiple folds. The model is then trained and validated multiple times, with each fold serving as the validation set once while the remaining folds form the training set. The average performance across all iterations provides a more robust estimate of the model's generalization capabilities.

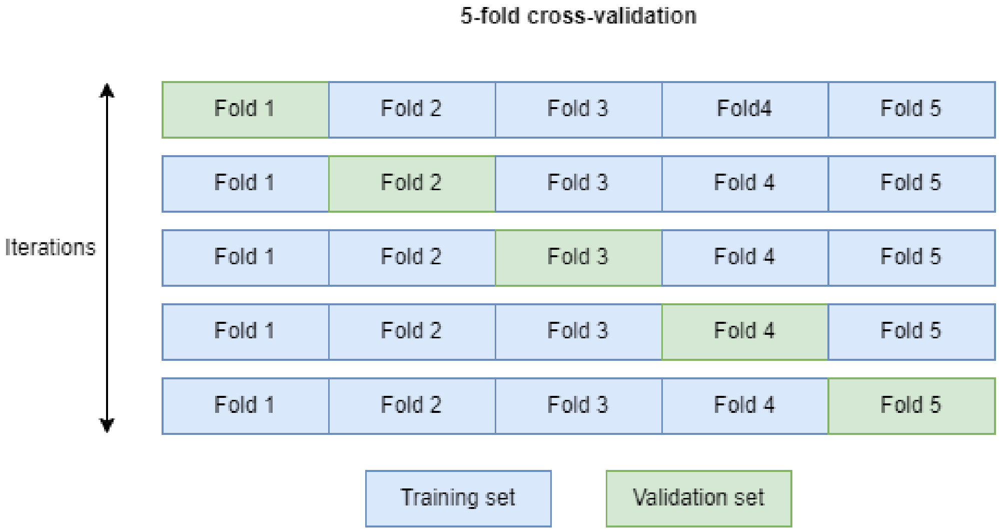


Figure 2.7: Illustration of a 5-fold cross-validation. The dataset is divided into five folds. In each iteration, one of the 5 folds is used as the validation set.

Batch Normalization

Batch normalization is a technique used in deep learning models to reduce internal covariate shifts and enhance training stability. Internal covariate shift refers to the change in the distribution of inputs to a layer during training, which can slow down convergence and make it challenging to choose an appropriate learning rate.

Batch normalization addresses this issue by normalizing the inputs to each layer during training. It computes the mean and standard deviation of the input batch and applies a linear transformation to ensure that the inputs have zero mean and unit variance. This normalization process helps maintain a consistent distribution of inputs across layers, allowing for faster convergence and better generalization.

In addition to its regularization effect, batch normalization also allows for higher learning rates and reduces the sensitivity to weight initialization. This makes it easier to train deep learning models with many layers and achieve better performance on various tasks.

2.2.6 Weakly Supervised Learning Strategies

Weakly supervised learning strategies are particularly relevant in the context of WSI analysis for diagnosis and prognosis, as they allow AI models to learn from data with limited or noisy annotations. In many cases, WSIs have an assigned annotation for the entire image rather than precise labels for individual regions or cells. This type of annotation is considered a weak label, as it provides limited information about the specific locations and characteristics of the objects of interest within the image.

2.2.7 Evaluation Metrics

When it comes to the evaluation of deep learning models, it is crucial to have a standard for measuring the models' performance. Evaluation metrics offer a quantitative method to evaluate the model's capacity to make precise predictions. They can be utilized to compare various models or identify areas for improvement. The ensuing sections present some commonly used evaluation metrics in machine learning, which include the confusion matrix, precision, recall, F1-score, accuracy, specificity, and receiver operating characteristic (ROC) curve.

Confusion Matrix

The confusion matrix is a table that presents a thorough performance summary of a classification model by comparing its predicted labels with the true values (annotated labels) . It offers a detailed distribution of the model's predictions across different categories, thereby making it easier to spot particular types of errors and imbalances in the model's performance. Figure 2.8 presents the four elements forming the confusion matrix. Each element will be subsequently explained and expressed mathematically:

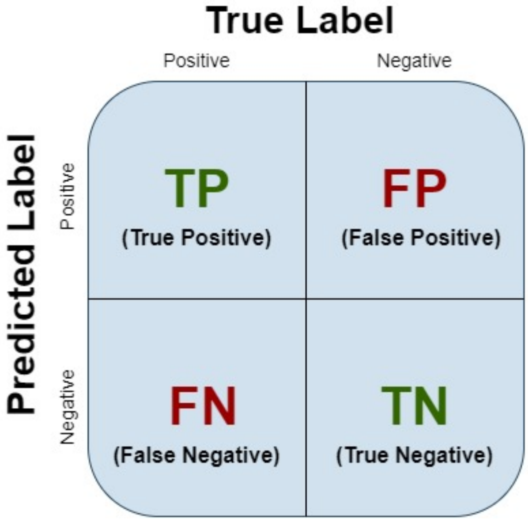


Figure 2.8: Illustration of a confusion matrix showing the components: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN)

Let's first denote:
y_i: The actual class of the i-th instance
$\hat{y}_i$: The predicted class of the i-th instance

We can define the positive class as 1 and the negative class as 0. Then, the four types of outcomes in a binary classification problem can be represented mathematically as follows:

- **True Positives (TP):** The number of instances that the model correctly identified as positive. For instance, correctly identifying malignant skin lesions in a set of WSIs.
$$\mathrm{TP} = \sum_i I(y_i = 1, \text{and}, \hat{y}_i = 1) \tag{2.1}$$
- **False Positives (FP)**: The number of instances that the model wrongly marked as positive. For example, incorrectly identifying benign skin lesions as malignant

in a set of WSIs.

$$\text{FP} = \sum_i I(y_i = 0, \text{and}, \hat{y}_i = 1) \tag{2.2}$$

- **True Negatives (TN)**: The number of instances that the model correctly identified as negative. This could mean correctly identifying benign skin lesions in a set of WSIs.

$$\text{TN} = \sum_i I(y_i = 0, \text{and}, \hat{y}_i = 0) \tag{2.3}$$

- **False Negatives (FN)**: The number of instances that the model wrongly identified as negative. An example of this would be misclassifying malignant skin lesions as benign in a set of WSIs.

$$\text{FN} = \sum_i I(y_i = 1, \text{and}, \hat{y}_i = 0) \tag{2.4}$$

In the above equations, I denote the indicator function, which is defined as:

$$I(\text{condition}) = \begin{cases} 1, & \text{if condition is true} \\ 0, & \text{if condition is false} \end{cases}$$

This function gives a sum of 1 each time the condition inside is true and 0 otherwise. In this way, we can count the number of times each condition (TP, TN, FP, FN) occurs in the dataset.

Precision

Precision is a key assessment parameter that measures the proportion of instances that the model correctly identified as positive out of all instances predicted as positive. The mathematical representation is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.5}$$

In the medical field, precision is of great importance when the implications of FP are severe. For instance, when screening for skin cancer, high precision implies that the model correctly identifies a majority of real cancer cases while reducing the number of false alarms. This can help avert unneeded subsequent tests and treatments for patients who do not have cancer.

Recall

Recall, alternatively known as sensitivity, is an assessment parameter that measures the proportion of instances that the model correctly identified as positive out of all actual positive instances. The mathematical representation is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.6}$$

In the medical field, recall is critically important when the implications of FN are severe. For example, when it comes to the early detection of diseases, a high recall signifies that the model identifies most of the real cases while minimizing the number of missed cases. This is crucial as it ensures that patients with the disease receive timely treatment, potentially improving their prognosis.

F1-score

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance in terms of both FP and FN. It is calculated as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.7}$$

In the medical field, the F1-score becomes particularly useful when both FP and FN have serious consequences, and a balance between precision and recall is desired. For instance, when diagnosing life-threatening conditions, a high F1-score implies that the model effectively identifies true cases while minimizing both false alarms and missed cases. This balance ensures that patients receive appropriate diagnosis and treatment while also preventing unnecessary procedures for patients without the condition.

Accuracy

Accuracy is a basic assessment parameter that measures the proportion of instances correctly classified by the model. It is calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.8}$$

In medical diagnosis, accuracy can be used to gauge the overall performance of a model in correctly identifying both positive and negative cases. However, it may not be the most appropriate metric when dealing with imbalanced datasets, where one class is significantly more prevalent than the other. In such scenarios, a high accuracy might be misleading as the model could achieve a high score by simply predicting the majority class.

Specificity

Specificity, also known as the TN-Rate, measures the proportion of negatives that are correctly identified as such. It is mathematically represented as:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.9}$$

In the medical field, specificity is vital when the cost of FP is high. For instance, in disease screenings, a high specificity indicates that the model correctly identifies a large proportion of healthy cases while minimizing the number of false alarms. This can help reduce the stress and unnecessary medical procedures for healthy patients.

Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC)

The Receiver Operating Characteristic (ROC) curve is a graphical illustration of the performance of a classification model across different decision thresholds. It plots the TP-Rate (recall) against the FP-Rate (1 - specificity) for various threshold values. The ROC curve provides a visual way to assess the trade-off between sensitivity and specificity, thereby helping identify the optimal decision threshold for a given application, see figure 2.9.

The area under the ROC curve (AUC) is a scalar value that summarizes the overall performance of the classification model. An AUC of 1 indicates perfect classification, while an AUC of 0.5 corresponds to random chance. A higher AUC value denotes better classification performance, making it a valuable metric for comparing different models or assessing the effectiveness of feature extraction and learning strategies in WSI analysis.

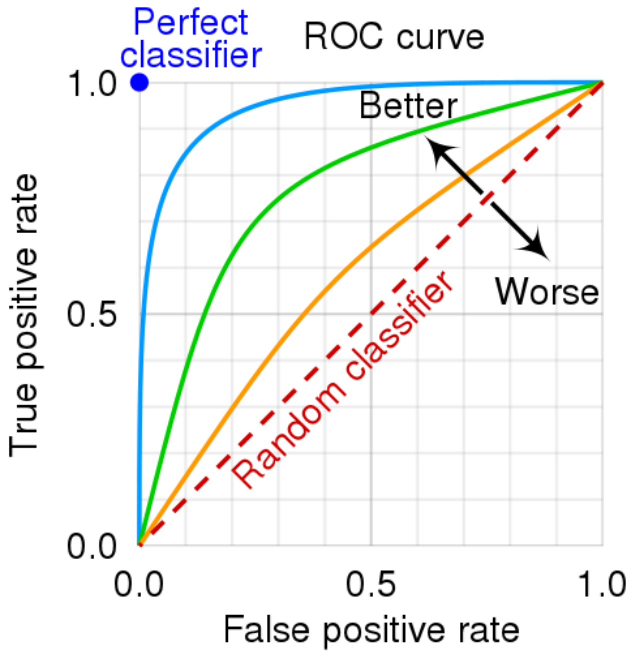


Figure 2.9: ROC-curve showing where the "perfect classifier" would be. The goal is for the classifier to have a low false positive rate and a high true positive rate. *This figure was obtained from Wikimedia Commons. The original illustration was created by cmglee, MartinThoma, and is in the public domain.* Source: Wikimedia Commons.

Chapter 3

Data Material

This chapter introduces the datasets provided by Stavanger University Hospital from 2022 and a new dataset from 2023. Additionally, it covers the provided annotations and the utilization of the data in this study.

3.1 D_D: Dataset used in Diagnosis Prediction and Localization

The dataset used to train the model for diagnosis prediction and localization (D_D - Data Diagnosis and Localization) was, provided by Stavanger University Hospital in 2022. It consists of 90 Whole Slide Images (WSIs) scanned with a Hamamatsu Nanozoomer s60. Each WSI has an associated diagnosis label (benign nevus, lentigo benign or malignant melanoma) provided by a pathologist working at the hospital. The pathologist also annotated each slide with different features, such as the location of benign, malignant melanocytic lesions and/or normal tissue.

The dataset was divided into three subsets: a training set (D_{DTrain}), a validation set (D_{DVal}), and a test set (D_{DTest}). The labels benign nevus and lentigo benigna were merged into a single category, "benign nevus." The distribution of WSIs across the three sets based on labels can be found in Table 3.1.

Set	D_{DTrain}	D_{DVal}	D_{DTest}	D_D
Benign Nevus	38	4	5	47
Melanoma	35	4	4	43
Total	73	8	9	90

Table 3.1: The number of WSIs provided to the diagnosis thises [3] is shown in this table. Each set is based on an image-based diagnosis.

3.2 D_P: Dataset used in Prognosis Prediction

The dataset used for prognosis, denoted as D_P, was provided by the Stavanger University Hospital in 2022. It consists of 52 WSIs obtained from different patients. All of the WSIs represent patients diagnosed with malignant melanoma who underwent a follow-up examination after a 5 years-period. The WSIs in the dataset are labeled with weak labels indicating the prognosis results of the follow-up. The prognosis was determined by the presence of metastasis.

The dataset is equally distributed of 26 patients with good prognoses and 26 with bad prognoses. Alongside the WSIs, included annotations provided by a pathologist, are stored in an XML file.

To create the dataset, a stratified 5-fold cross-validation approach was employed. The data was divided into five iterations of training and validation using up to 250 patches per WSI.

3.3 Annotation

Both D_D and D_P datasets were in addition to the weak labels, provided annotations around ROIs or features by a pathologist at Stavanger University Hospital. These annotations were created using the University of Stavanger's in-house developed web-based annotation tool for histological images (annotation tool). The annotations were exported as XML files containing polygon coordinates and tags corresponding to the features.

An annotation protocol was established upfront through collaboration between expert pathologists and experts in image processing and AI development at the University of Stavanger. The protocol aimed to provide some annotations on ROI and lesions for all patients while minimizing time spent on delineating borders accurately, as displayed in fig 3.1.

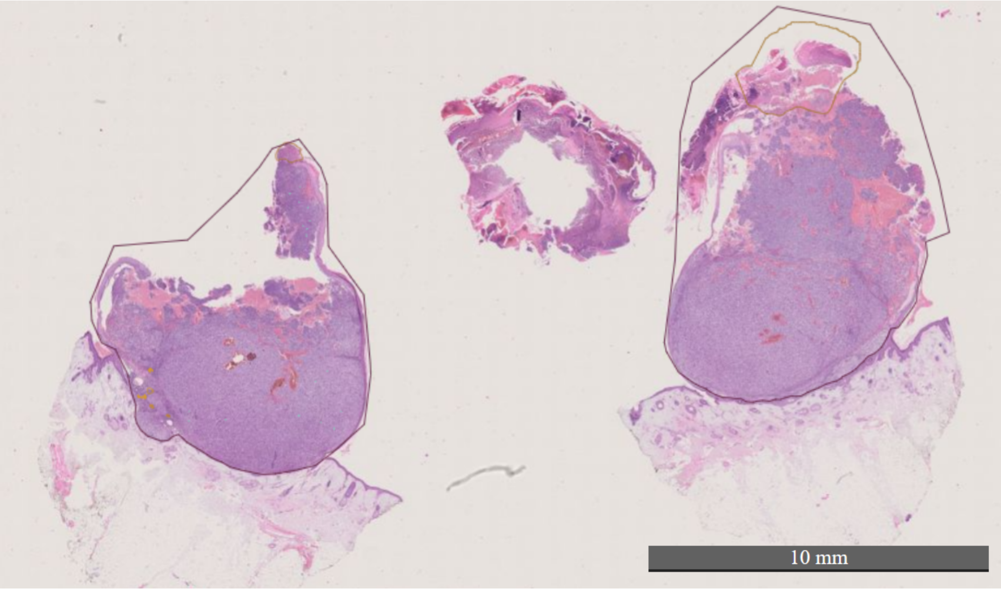


Figure 3.1: Example on an annotated WSI consisting of three slices of the same nevi. Each annotated region is provided with a tag telling if the area is malign, benign, or normal tissue. Each WSI is also provided with a patient-based diagnosis.

3.4 D_{New}: Dataset from 2023

The dataset used for this thesis, denoted as D_{New}, consists of a total of 243 WSIs obtained from different patients. The data was collected from Stavanger University Hospital and consists of 110 patients diagnosed with a benign nevus and 133 patients diagnosed with malignant melanoma, see Tabel 3.2. Among the patients with melanoma, 10 patients were diagnosed with a metastasis prognosis, while 101 patients had no metastasis prognosis based on a 5-year check-up. Each WSI in the dataset is given a weak label, indicating the diagnosis and/or prognosis of each corresponding patient.

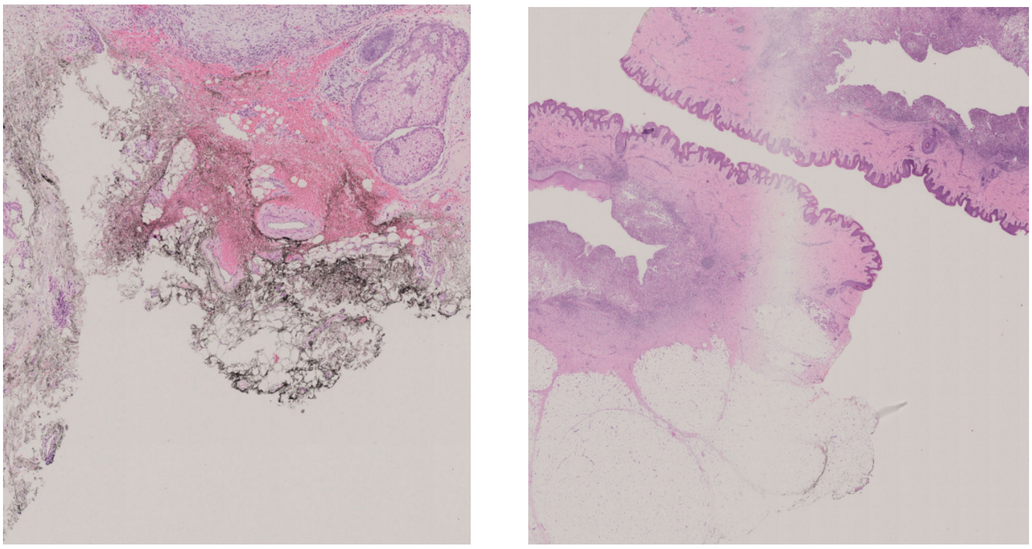
The data set D_{New} is divided into two subsets: a validation set (D_{NewVal}) and a test set ($D_{NewTest}$). The test contains 40 WSIs, where 18 of them are benign and 22 malignant. Of the malignant 20 are labeled with metastasis and 2 with no metastasis.

Set	D_{NewVal}	$D_{NewTest}$	D_{New}
Benign Nevus	92	18	110
Melanoma	111	22	133
Total	203	40	243
Metastasis	10	2	12
No Metastasis	101	20	121

Table 3.2: The dataset D_{New} provided for this thesis. Divided up into a validation D_{NewVal} set and a test set $D_{NewTest}$.

3.4.1 Artifacts

Some WSIs contain artifacts in the form of black stains from pen markings, Figure 3.2 (a). Other artifacts could be light areas affected by the scanning, Figure 3.2 (b). These artifacts should be taken into consideration when analyzing the model for diagnosis prediction and localization [18].



(a) Artifacts caused by black stains.

(b) Artifacts caused by scanner.

Figure 3.2: Artifacts caused by the black stains and scanning of the WSIs.

Chapter 4

Methods

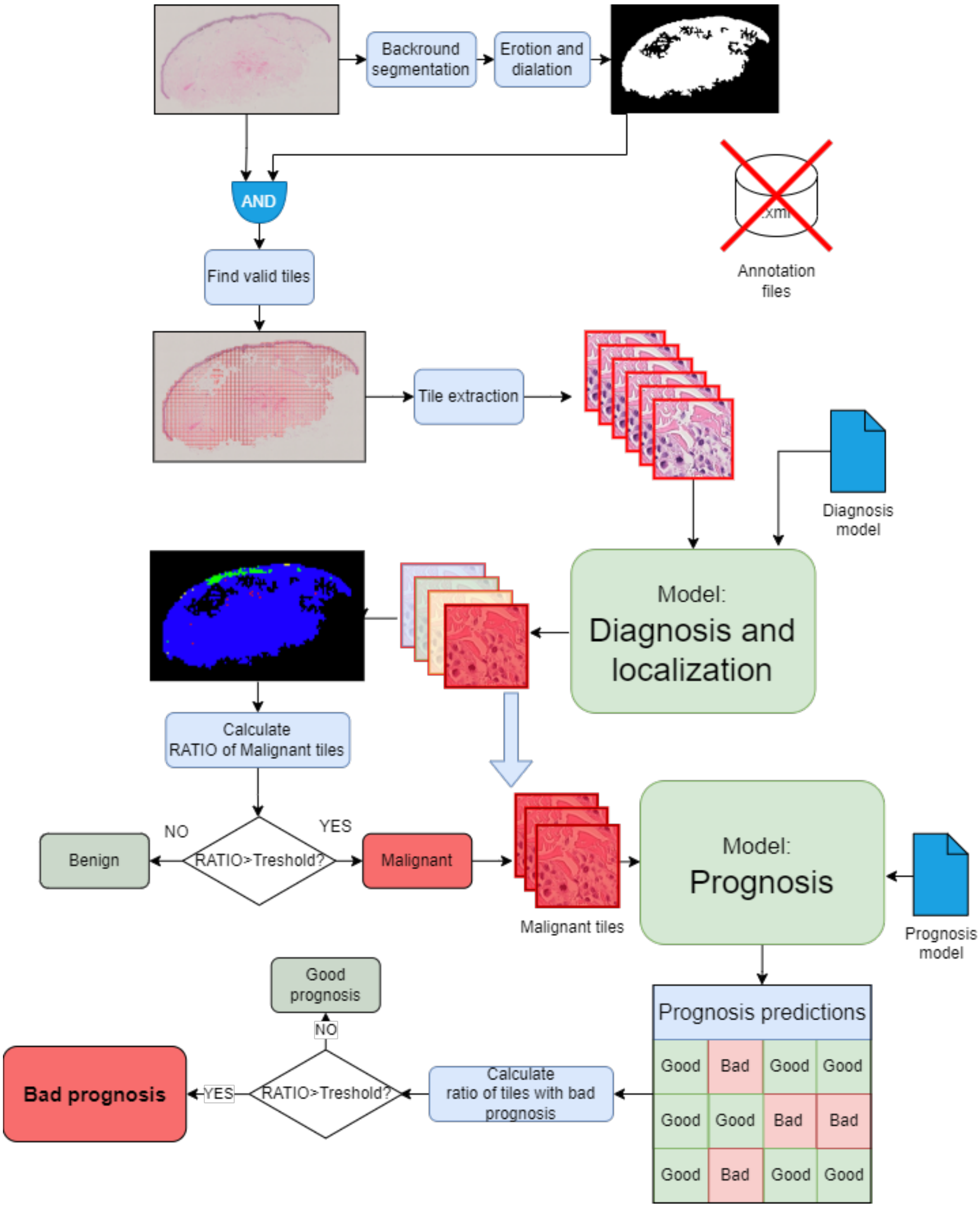
This chapter presents a detailed overview of the methodologies employed in this thesis. The proposed pipeline comprises three main steps: preprocessing, diagnosis prediction with lesion localization, and prognosis prediction. Each of these stages is thoroughly described to ensure a comprehensive understanding of the methodology.

4.1 Overview

The proposed method is illustrated in Figure 4.1. The initial step of the pipeline involves preprocessing, which is essential due to the large size of the Whole Slide Images (WSI). The preprocessing phase primarily focuses on background segmentation and extracting patches from the tissue regions within the WSI.

The subsequent step involves using the extracted patches as input for the diagnosis and localization model. This model utilizes pre-trained weights from previous work [3]. The output from the model contains a multi-class prediction for each patch: malignant, benign, and normal tissue. To ascertain the diagnosis prediction for the entire image, the ratio of malignant patches is examined. If the ratio of malignant patches exceeds a given threshold (t_r), the WSI is predicted as malignant. Conversely, if the ratio falls below t_r, the WSI is predicted as benign. If the WSI is predicted as malignant, the malignant tiles are used as input for the next stage: prognosis prediction.

The prognosis prediction model is responsible for providing a binary prediction for a patient's prognostic outcome after a five-year period. This model uses pre-trained weights from earlier works [4]. It was trained on a limited dataset of WSIs containing only malignant diagnoses. Each patch prediction is stored, and the image is then predicted based on the ratio between patches predicted as good and bad prognoses. If the ratio is greater than a threshold, the image is predicted as having a poor prognosis.



Prognosis predictions			
Good	Bad	Good	Good
Good	Good	Bad	Bad
Good	Bad	Good	Good

Figure 4.1: Overview of the pipeline for this project. The pipeline contains three main parts: Preprocessing, diagnosis with localization, and prognosis. This overview is simplified and is created to give a comprehensible overview of the workflow in this pipeline.

4.2 Previous work

This project is built upon two master theses from 2022 [3][4]. The first project was assigned to Roger Amundsen [3], with the objective of training a model to detect and localize lesions and predict a diagnosis (benign lesion or melanoma) using a dataset D_D provided by the Stavanger University Hospital (3.1). The second project was assigned to Christopher Andreassen [4], who trained a model to predict the prognosis of malignant melanoma using a different dataset D_P (3.2). For each WSI in both datasets, a weak label was provided, indicating the known diagnosis or/and prognosis, and an annotation showing the lesion's localization. The diagnosis label specifies if the patient had cancer or not, and the prognosis label is based on whether a patient with cancer experienced metastasis within five years or not. This section will explain the work done by each of them.

4.2.1 Diagnosis and Localization Model

The diagnosis model presents a method to predict patches as either normal tissue, benign nevus, or melanoma and provides an overview of these predictions. Additionally, the model calculates an image-based prediction of the WSI based on the ratio of malignant pixels. The model utilizes the VGG16 architecture with pre-trained weights and is fine-tuned on WSIs provided and annotated by Stavanger University Hospital. The D_D dataset consisted of 93 WSIs, which were divided into D_{DTrain}, D_{DVal}, and D_{DTest} as explained in Chapter (3.1). All images were assessed by a pathologist and given a patient-based diagnosis label of either benign or malignant. In addition, the pathologist provided a rough annotation of the lesion in all WSI.

Numerous models were trained, and the most promising one was selected for incorporation into this project. The multi-class model was specifically designed to predict three classes "Lesion Benign" (LB), "Lesion Malignant" (LM), and "Normal Tissue" (NT), while the binary-class model was trained to predict only LB and LM.

Table 4.1 presents an overview of the performance of the two multi-class models (16 and 17), along with a binary-class model for the purpose of comparison. The table presents F1 scores and accuracy for the three classes (LB, LM, and NT) as predicted by the two multi-class models. Additionally, LB and LM for the binary-class model. The F1 score reflects the model's ability to correctly recognize the different tissue types compared to the labels for the annotated regions. Examining the table, model 17 performs less accurately when predicting LB and LM, but outperformed the other model in detecting NT. Model 17 was chosen for this thesis because it performed the best when introduced with unseen tissue.

$Model$	$Accuracy$	F_{1LB}	F_{1LM}	F_{1NT}	Epochs
Model 10	0.9813	0.9702	0.9864	-	8
Model 16	0.9706	0.9572	0.9865	0.5399	9
Model 17	0.9663	0.955	0.9799	0.5814	8

Table 4.1: Tabel shows one binary-class model and the two multiclass models from R.Amundsen thesis [3]. F1 scores for the three classes are shown: Lesion Benign (LB), Lesion Malignant (LM), and Normal Tissue (NT). The models was evaluated on D_{DVal}.
Values are collected and unaltered from [3].

Model 17 is a multi-class model designed to predict three classes: benign, malignant, and normal tissue. The model proved to be the best overall model. During its development, experiments were performed to find the optimal thresholds for making predictions at the patch level (for assigning a class to a patch) and on the image level (to determine patients with melanoma).

The model's output layer uses a softmax function that produces a probability vector, $\hat{p}$, indicating the likelihood of the patch belonging to a specific class LB, LM, and NT. If the maximum probability in the vector is greater than or equal to the threshold t_p, the class with the highest probability is chosen for that specific patch. If the maximum is below the threshold t_p, the patch is classified as "tissue". ROC and precision-recall curves were examined to find the optimal threshold on patch level t_p for each model, displayed in Figure 4.2. Based on the results obtained by R.Amundsen [3], a patch threshold value of t_p=0.999 was found to be the best threshold for this model.

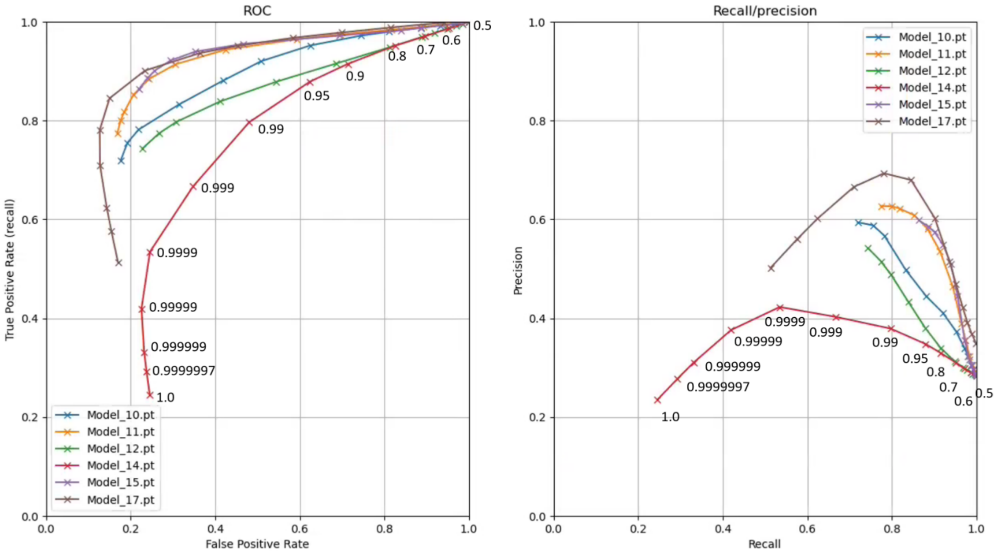


Figure 4.2: ROC and Recall/Precision from R. Amundsen's experiments to find the best patch-level threshold for the different models. model 17 with t_p= 0.999 was used.
Values are collected and unaltered from [3].

The patient-level classification $\hat{y}_s$ was found by using the best model and corresponding t_p to find the best ratio threshold t_r. To classify a WSI, the ratio of malignant patches (R_{mp}) was computed using equation 4.1. If the MB_{rate} ratio exceeds the threshold t_r the WSI is classified as malignant, and if the ratio is below the threshold it is classified as benign, as shown in 4.1.

$$MB_{rate} = \frac{\text{Malignant Area}}{\text{Lesion Area}} = \frac{\sum_p \hat{y}_{p\text{M}}}{\sum_p(\hat{y}_{p\text{M}} + \hat{y}_{p\text{B}})}$$

$$\hat{y}_s = \begin{cases} \text{Malignant}, & \text{if } MB_{rate} \geq t_r \\ \text{Benign}, & \text{else} \end{cases} \tag{4.1}$$

Dataset D_{DVal} was used to find classification results $\hat{y}_s$ with a selection of thresholds t_r to decide the optimal threshold to use. To avoid false positives (malignant WSI predicted as benign), the ratio threshold (t_r) was determined based on the point where no more false positives (benign WSI predicted malignant) were observed, at t_r=0.04. This threshold provided 100% correct patient-level classification on the validation set D_{DVal}, effectively ensuring accurate predictions and minimizing the risk of misclassification.

By using the best model obtained from the training data and the best threshold from the validation phase, R. Amundsen was able to achieve 100% accuracy on patient-level classification on the test set D_{DTest} consisting of nine WSIs. This result indicates that the model accurately classifies them as either benign nevus or melanoma.

Moreover, for localization of the lesion, four WSIs from D_{DTest} were used. The evaluate the precision of the model, the entire tissue within each WSI was fed into the model, generating predictions for each patch in the WSI. The model discovered both benign and malignant classifications with great precision when compared to the provided annotation, as shown in Table 4.2. This shows the model's ability to accurately locate and differentiate between different lesion types within the WSIs.

Metrics	Recall	Precision	F1-score
Lesion Benign	0.8378	0.9999	0.9117
Lesion Malignant	0.9354	0.9999	0.9666

Table 4.2: Results on the diagnostics and localization model when it comes to localization of the lesions. Performance metrics for the two classes of lesions, benign lesions, and malignant lesions based on the associated annotations [3].

4.2.2 Prognosis Model

The prognostics model presents a methodology for predicting whether a patient diagnosed with malignant melanoma will experience metastasis within five years, i.e. a spread of the cancer to a different part of the body, based on follow-up information. Patients who develop metastatic melanoma within this timeframe are classified as having a poor prognosis, while those who do not are categorized as having a good prognosis.

The prognostics method utilized the VGG16 architecture with pre-trained weights on ImageNet, augmented with fully connected layers, including two dense layers of 4096 neurons each and a final layer for the binary prognosis output [3].

The network was trained, validated, and tested on a dataset D_P of a total of 51 WSIs from different patients with known melanoma prognoses from Stavanger University Hospital. Weak labels of prognosis and annotation of the lesion were also provided for this thesis, as discussed in Chapter 3.2. The output of this classification was subsequently used to predict the overall prognosis for each WSI. Patch-based predictions, in the case of predicting the prognosis of patches, are challenging as there is only truth data at a patient level available. All extracted patches within a WSI will then inherit the patient truth label.

There was done some experiments involving training on different magnification levels of the patches to find the optimal model, using cross-validation. The best performing model turned out to be the model trained on only one magnification level (mono-scale), using 20x, see results in Tabel 4.3.

Iteration	Recall	1-Specificity	F1-score	Accuracy	AUC
Total:	0.8846	0.4400	0.7667	0.7255	0.81

Table 4.3: Results from the prognostics model predicting prognosis of WSIs, only showing the total iterations of the dataset. Using magnification level 20x, and threshold set to 0.3720 [4].

An overview of the model trained on multiple magnifications (multi-scale) is displayed in Figure 4.3. The best performing model was a mono-scale, trained only on one magnification level (20x), with a single backbone, and the output of the feature embedding is directly fed into the classifier. Each patch going through the network generates a prediction of either 0 or 1, indicating a bad or good prognosis, respectively.

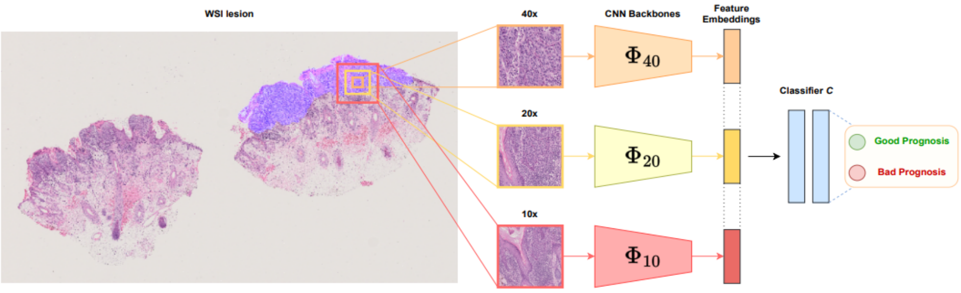


Figure 4.3: Overview over the multiscale model. The output feature embedding from mono scale models, as used in this thesis, is fed directly into C using a single backbone..
Figure are reprinted and unaltered from[5] with permission from the author.

The threshold for classifying WSIs based on image-level predictions was decided by using a ROC curve, displayed in Figure 4.4. Aiming to minimize false negative predictions, the threshold was chosen based on relatively high sensitivity and a high AUC. The selected threshold was established at 0.3720. WSIs are predicted to have a bad prognosis if the proportion of patches predicted as "bad" exceeds that threshold, while those below the threshold were classified as having a good prognosis.

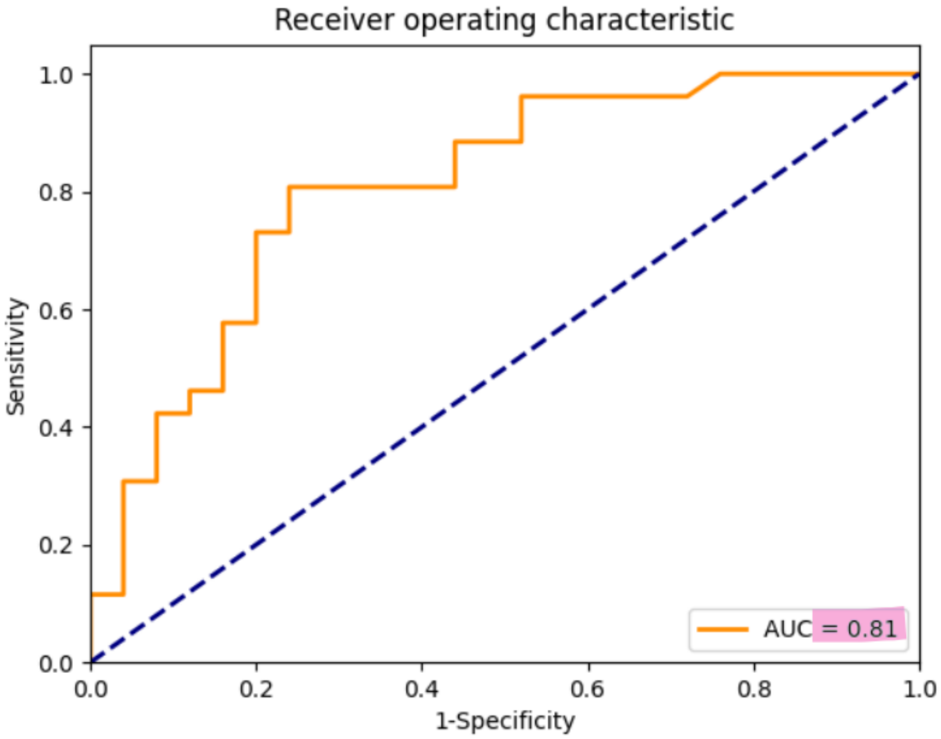


Figure 4.4: ROC and AUC from the results on mono-scale using all iterations from the dataset D_P.
Image are collected and unaltered from C. Andreassens' thesis [4].

4.3 Pipeline - Diagnosis, Localization, and Prognosis

This section presents the proposed pipeline and provides a comprehensive explanation of each step involved in preprocessing, diagnostics, localization, and prognosis. The primary objective of this thesis is to integrate the diagnosis and prognosis tasks into a unified pipeline, specifically focusing on WSIs without annotations provided by a pathologist, as depicted in Figure 4.1.

The main goal is to evaluate the two models using both old and new data sets, generate a diagnosis output that aligns with the input format used to predict prognosis, and finally improve the result using the new dataset by adjusting thresholds. To facilitate this process, a separate directory was created exclusively for storing masks generated for predicted malignant melanoma making it possible to unify the models.

4.3.1 Preprocessing

Preprocessing is a crucial step, especially for high-resolution images such as Whole Slide Images (WSIs), which encompass a vast amount of pixel data and demand time-consuming processing. By utilizing preprocessing techniques like color thresholding and morphological operations, the input region into the model can be minimized. This reduction leads to a smaller number of patches being fed to the model, ensuring that it receives patches containing more pertinent information for distinguishing between malignant tissue and other tissue types. Figure 4.5 provides an overview of the preprocessing steps. The segmented region is depicted as an image in the top right corner of the figure. The region of interest is displayed in white, while the background appears black. This is referred to as the tissue mask throughout this chapter. Tabel 4.4 shows the different methods used along with authors and whether or not they are modified.

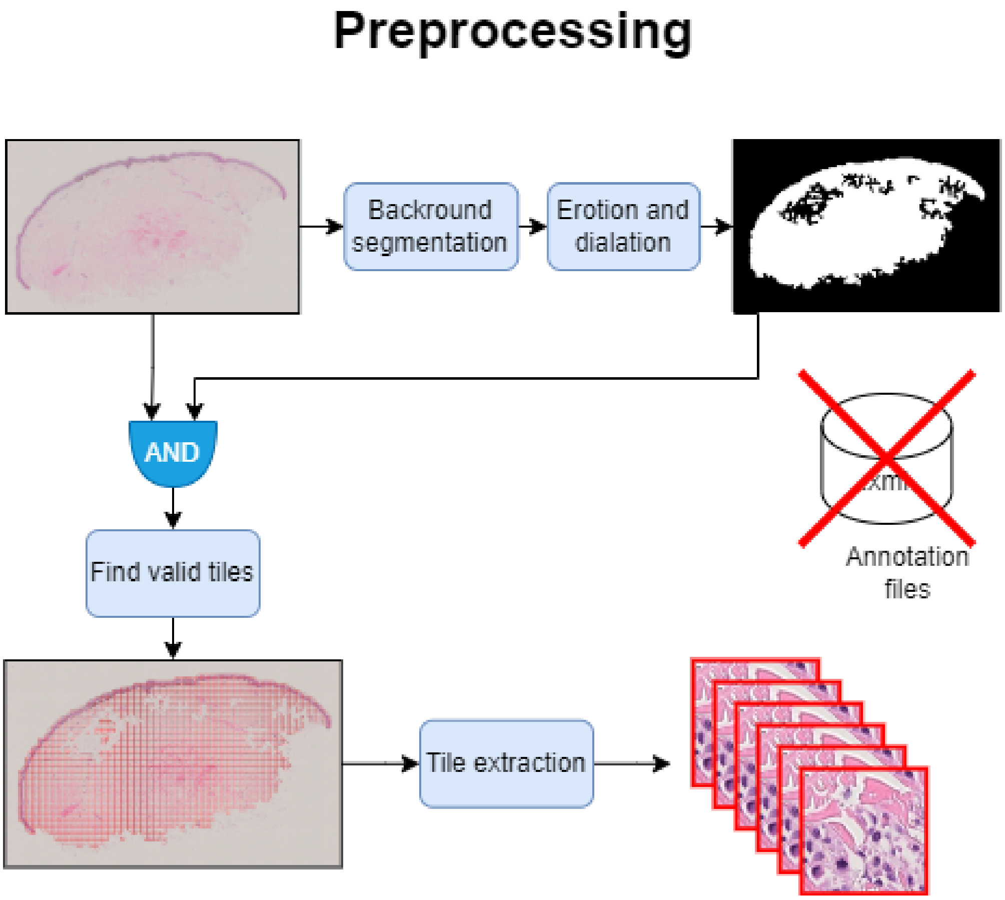


Figure 4.5: Overview of the preprocessing of the WSIs in this work. The WSIs are loaded, then a color threshold is used in the HSV color channels for background segmentation. The foreground mask is then eroded and dilated. Erosion ensures that small separated areas of the foreground mask are removed. The dilation fills in small gaps in the mask, resulting in a more coherent foreground mask. The foreground mask is then used in combination with the original WSI to create a WSI image only containing the region of interest. Lastly, the ROI is divided into patches of 256x256 pixels. The upper left coordinate for each patch is stored and used by the inference function to extract the patches.

The proposed preprocessing method generates a binary tissue mask of the background and tissue by converting the color representation from RGB to HSV format and then applying a threshold on the Hue channels. Compared to the RGB model, the HSV color model offers a more reliable approach for manipulating and representing colors, as explained in Section 2.2.2. Pixels with hue values outside the range of 100-180 are identified as belonging to the background. This threshold effectively segments out the majority of the area that is not affected by the H&E stain. Once the tissue has been separated from the background, erosion, and dilation techniques are applied. Erosion operations are responsible for removing small regions, while dilation unifies areas, resulting in a more consistent tissue mask, explained in background 2.2.2.

Both the diagnosis and prognosis model used in this thesis uses VGG16 architecture. As mentioned in chapter 2.2.4, the VGG16 architecture uses 224x224 pixels as input. To convert the patch size to the correct input format, each patch is cropped from 256x256 to 224x224. This is illustrated in Figure 4.6.

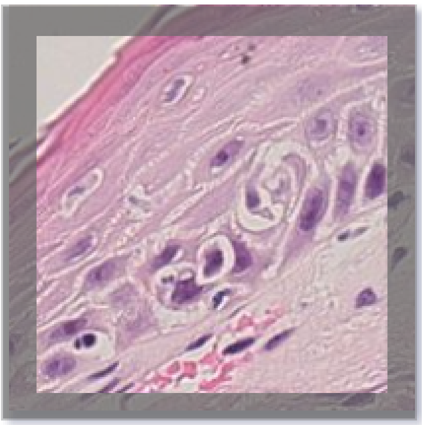


Figure 4.6: Depicting the cropping of a patch from 256x256 to 224x224. The VGG16 architecture uses 224x224 pixels as input. To convert the patch size to the correct format, each patch was cropped from 256x256 to 224x224.

As mentioned earlier in Section 3.4, the dataset D_{New} only contains image-level labels. The objective of creating the tissue mask is to isolate the tissue, hence removing areas that do not provide any relevant information for the model.

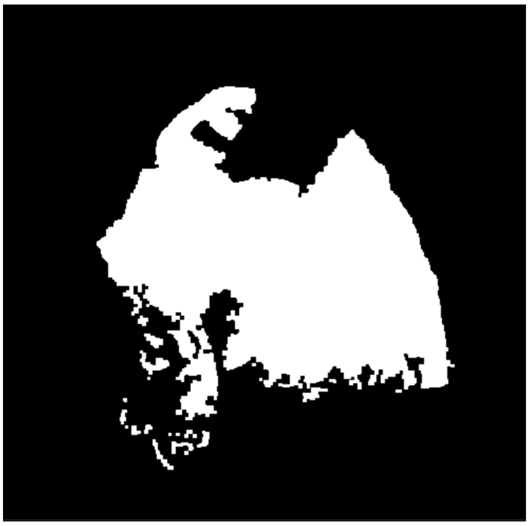


Figure 4.7: Tissue mask before morphological operations.

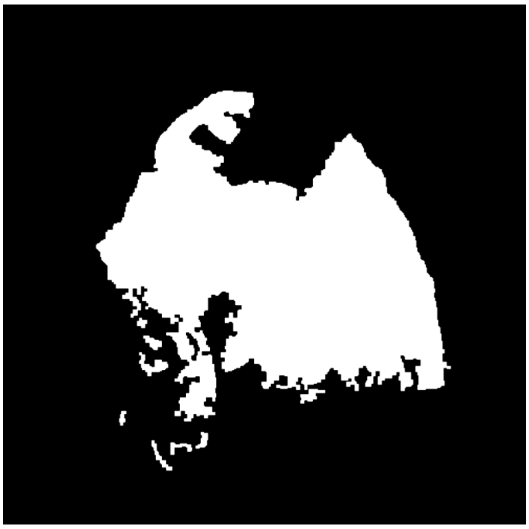


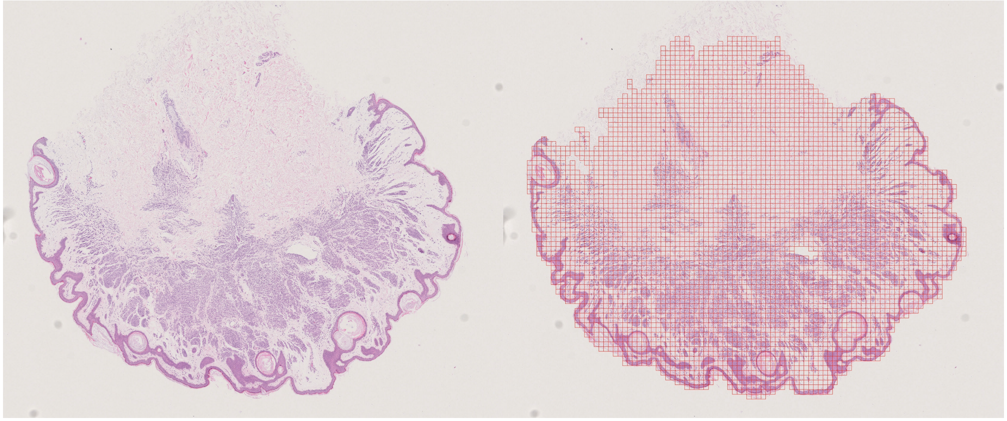
Figure 4.8: Tissue mask after dilation and erosion.

After the tissue mask is generated, the ROI has to be processed by the model. This is done by patch extraction. Extracting patches from Whole Slide Images (WSIs) is essential due to the challenges associated with processing an entire gigapixel image at its full resolution in a single step. The patch extraction method employed in this work, initially developed by Rune Wetteland [32], was modified by R.Amundsen to address specific challenges. These modifications include using only the areas inside the tissue mask to scan for valid patches, using dynamic τ, and invalidating patches containing a considerable amount of white backgrounds.

The preprocessing method extracts patches from tissue masks. After the valid patches are found, the upper left coordinate for each valid patch enables multi-level patch extraction at varying magnification levels (2.5x, 10x, 20x, and 40x). This approach provides the possibility of a wider field of view and a more comprehensive understanding of the surrounding tissue context. By using a high magnification level for the patch extraction, more patches can be extracted.

All patches have a consistent size of 256x256 pixels. Due to their square shape, some patches inevitably extend beyond the boundaries of the tissue masks. To determine if a patch is valid, an area threshold ϕ is employed, specifying the percentage of overlap with the tissue mask required for a patch to be considered valid. In this case, the threshold is set to 0.7. This means if less than 70% of the tissue mask is inside the patch, it will not be considered valid.

Finally, valid patches are extracted, and their coordinates are stored as a dictionary in a pickle file for efficient storage and retrieval during further analysis and processing in this study. Figure 4.9 shows an example of a WSI and its valid tiles.



(a) WSI before preprocessing and tile extraction **(b)** WSI after foreground segmentation and divided into valid patches.

Figure 4.9: Showing valid patches after extracting patches from a WSI. Image (a) is showing the WSI before preprocessing, and image (b) is after valid patches are determined. Leaving out background patches without important information will make the computation easier and more effective.

The preprocessing utilizes code from previous projects. Authors of the code and a list of the files used and modified during preprocessing are presented in Table 4.4.

Method	Author	Comment
Background segmentation	C.Andreassen	Modified
Patch extraction	Wetteland, modified by R.Amundsen	Unaltered
Remove background patches	R.Amundsen	Unaltered
Patch normalization	R.Amundsen	Unaltered

Table 4.4: This table shows an overview of the methods used in developing the model responsible for preprocessing. The original author and modification status are also shown.

4.3.2 Step 1: Melanoma Detection

Upon the successful completion of the preprocessing phase, the valid patches are inputted into the diagnosis and localization phase. Figure 4.10 illustrates the components of the pipeline responsible for localizing malignant patches and providing an image-level prediction.

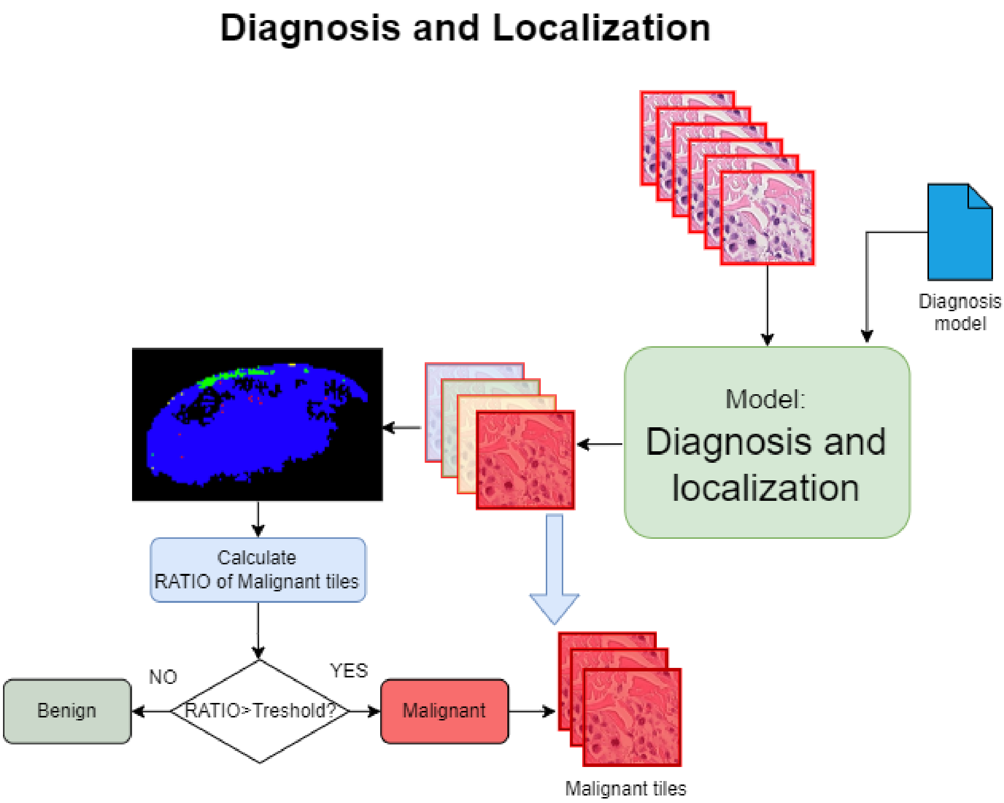

Diagnosis and Localization
Diagnosis model
Model:
Diagnosis and localization
Calculate RATIO of Malignant tiles
NO
YES
Benign
RATIO>Treshold?
Malignant
Malignant tiles

Figure 4.10: Overview of the pipeline's diagnosis and localization process. The model receives input from all of the valid patches discovered during preprocessing. The output consists of a prediction of each patch, and a mask displaying the location of the predictions. To predict the patient-level diagnosis, the ratio of malignant patches is compared with a threshold. The malignant patches for the WSIs that were predicted cancerous will be saved for further examination.

The diagnostic model employs the VGG16 architecture and pre-trained weights from R. Amundsen's master thesis [3], as detailed in section 4.2.1. model 17 demonstrated the best performance (F_1-score 0.7535), as displayed in table 4.5.

Model	10	11	12	14	15	17
F1	0.6613	0.7102	0.6270	0.5139	0.7069	0.7535
tp	0.9999997	0.99999	1.0	0.99	1.0	0.999

Table 4.5: Each models' best F1-score and corresponding patch-based threshold t_p. The results are patch-level prediction ($\hat{y}_p$) using all patches of the validation set. Results from R.Amundsens thesis [3].

Initially, an instance of model 17 is initialized by loading the pre-trained weights. The valid patches from preprocessing are retrieved using the coordinates for each patch in conjunction with the tissue mask.

Inference is then executed on each valid patch and stored in a dictionary. As Model 17 is capable of multiclass prediction, the output ($\hat{y}_{ip}$) consists of an array of three prediction values for each patch, where benign nevi (B), malignant nevi (M), and normal tissue (NT) represent the outputs as depicted in Equation 4.2.

$$\mathbf{\hat{y}_{ip}} = \begin{bmatrix} \hat{y}_{p\text{B}} \\ \hat{y}_{p\text{M}} \\ \hat{y}_{p\text{NT}} \end{bmatrix} \tag{4.2}$$

The patch-level prediction is denoted as $\hat{y}_{ip}$, where p denotes the patch, and i denotes the current image. The sum of all elements should equal 1, as demonstrated in Equation 4.3.

$$\sum_{i} \mathbf{\hat{y}_{ip}} = 1 \tag{4.3}$$

To decide whether a patch is malignant, a threshold t_p is employed. This threshold determines which prediction values of $\hat{y}_{p\text{M}}$ and $\hat{y}_{p\text{B}}$ should be considered melanoma and benign, respectively. The third prediction in $\hat{y}_{ip}$ is excluded and grouped with other tissue because normal tissue predictions are irrelevant in this case where the true value of each patch is unknown. In this thesis, the malignant and benign predictions are in focus. The patch-level thresholding is illustrated in Equations 4.4, 4.5, and 4.6.

$$\hat{y}_{p\text{M}} = \begin{cases} 1, & \text{if } \hat{y}_{p\text{M}} \geq t_p \\ 0, & else \end{cases} \tag{4.4}$$

$$\hat{y}_{p\text{B}} = \begin{cases} 1, & \text{if } \hat{y}_{p\text{B}} \geq t_p \\ 0, & else \end{cases} \tag{4.5}$$

$$\hat{y}_{p\text{NT}} = \begin{cases} 1, & \text{if } \hat{y}_{p\text{M}} \wedge \hat{y}_{p\text{B}} \leq t_p \\ 0, & else \end{cases} \tag{4.6}$$

Thresholding t_p is utilized to decide if the current patch is predicted as benign or malignant. If neither of these prediction values satisfies the t_p threshold, the tissue is assigned as normal tissue.

Although the model can predict normal tissue, it does so with an F_1 score of only 0.5814 (Table 4.1). For the purpose of this thesis, differentiating between normal tissue and other tissue is unnecessary and may lead to the exclusion of numerous patches. Consequently, a patch is assigned the label NT if it does not predict B or M.

The image-level prediction $\hat{y}$ is determined by first calculating the ratio MB_{rate} of patches predicted as malignant and benign. A threshold value is then used to ascertain $\hat{y}$. Equation 4.7 demonstrates how the MB_{rate} was calculated in [3].

$$MB_{\text{rate}} = \frac{\sum_p \hat{y}_{p\text{M}}}{\sum_p (\hat{y}_{p\text{M}} + \hat{y}_{p\text{B}})} = \frac{\text{Malignant Area}}{\text{Lesion Area}} \tag{4.7}$$

In this thesis, an alternative method for calculating the MB_{rate} is presented. Instead of only considering the lesion when calculating the malignant rate present in a WSI, all tissue regions can be used in the denominator to obtain a new rate, referred to as the Melanoma-Tissue rate (MT_{rate}). The new method considers the rate of malignant patches relative to all valid patches in the image, as shown in Equation 4.8.

$$MT_{\text{rate}} = \frac{\sum_p \hat{y}_{p\text{M}}}{\sum_p (\hat{y}_{p\text{M}} + \hat{y}_{p\text{B}} + \hat{y}_{p\text{NT}})} = \frac{\text{Malignant Area}}{\text{All Tissue Areas}} \tag{4.8}$$

Table 4.1 show that the F_{1LB}-score (benign patches) was 0.9550 and F_{1LM}-score (malignant patches) was 0.9799, while F_{1NT}-score (normal tissue patches) was 0.5814. The model performed well in detecting both malignant and benign patches on the D_{DVal} set. It is suggested here that basing the malignant ratio only on the lesion can introduce an extra layer of error into the equation. If the model is not able to detect the benign patches correctly, the MB_{rate} will consequently be largely impacted. MT_{rate} is suggested as a more reliable option. Since there was no annotation provided for the precise area encompassing all of the tissue, it is not possible to calculate performance metrics that would evaluate the accuracy of the tissue segmentation and patch extraction. However, a visual examination suggests that background segmentation and finding valid patches containing tissue are reliable. For a visual representation of the patch extraction, see Figure 4.9.

For simplicity, the MB_{rate} and MN_{rate} are going to be referred to as M_{rate} when using common equations for both methods. The value used for thresholding the M_{rate} is denoted as t_r (threshold rate). This value specifies the minimum percentage of the lesion that must be predicted as malignant in order to classify the Whole Slide Images (WSIs) as malignant, as illustrated in Equation 4.9.

$$\hat{y} = \begin{cases} 1, & \text{if } M_{rate} \geq t_r \\ 0, & else \end{cases} \tag{4.9}$$

In R. Amundsen's thesis [3], t_r of 0.04 in combination with a t_p of 0.999 was found to perform well on model 17. The dataset D_{DVal} used for determining the threshold contained nine WSIs, which may have resulted in too few data points to determine the general optimal thresholds.

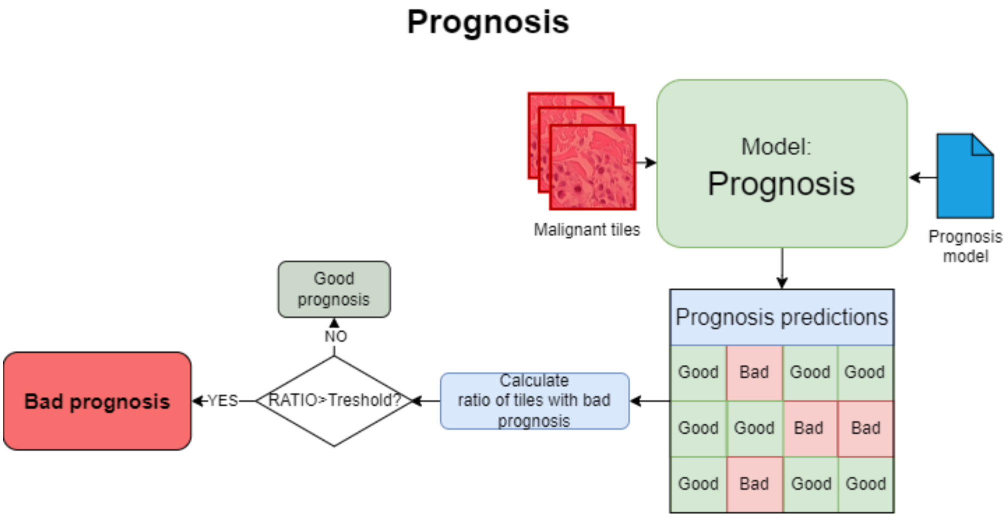
The portion of this pipeline that is responsible for diagnosis prediction and localization employs code from R. Amundsen's thesis [3] with modifications. An overview of the functions used in this part of the pipeline is presented in Table 4.6.

Method	Author	Comment
Model initialization	R. Amundsen	Unaltered
Running inference	R. Amundsen	Modified
Patch-level classification	R. Amundsen	Modified
Image-level classification	R. Amundsen	Modified
Prediction storage	R. Amundsen	Modified

Table 4.6: This table provides an overview of the methods used in developing the model responsible for diagnosis, along with the original author and modification status.

4.3.3 Step 2: Melanoma prognosis

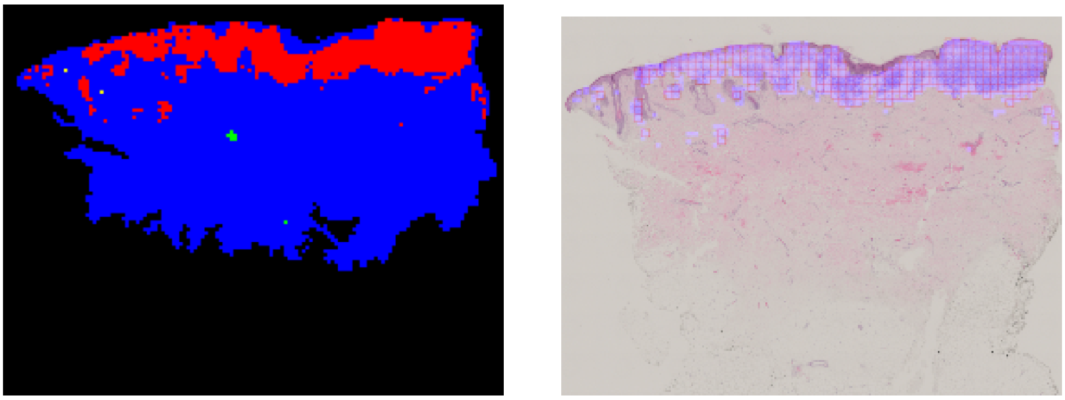
Once all of the images have been predicted and the malignant patches have been saved for their corresponding WSIs, the prognosis model uses these patches for further analysis. This process involves merging the predicted mask with the tissue mask and generating a mask that identifies the specific region(s) the prognosis model will make predictions. See Figure 4.11 for an illustration of this process.



Prognosis predictions			
Good	Bad	Good	Good
Good	Good	Bad	Bad
Good	Bad	Good	Good

Figure 4.11: Overview of the pipeline's prognosis process. The model receives all the malignant patches extracted from a predicted malignant WSI and employs pre-trained weights to provide a prognosis for each valid patch. The overall patient-level prediction is determined based on the ratio of predicted "bad" patches within an image, compared with a predefined threshold.

The first step involves using the mask derived from the malignant pixels to extract new patches. In order to ensure compatibility with the prognostic model's expected input format, the mask is adjusted accordingly. However, some of the malignant patches will be discarded during the process of extracting valid tiles. Malignant patches from the mask will be discarded in cases where 70% or more of the patch lies outside of the extracted patches. By removing patches where the majority of the area falls outside the region, the model can prioritize analyzing patches that provide meaningful insights for prognosis assessment. To illustrate this process, Figure 4.12. provides an example of this approach.



(a) Predicted image from the diagnosis. The red area indicates predicted melanoma and the green area is predicted benign.

(b) Valid tiles from the malignant mask combined with the WSI.

Figure 4.12: Figure (a) depicts an illustration of a predicted malignant WSI, while (b) demonstrates the valid patches extracted from the predicted malignant mask. These patches are used as input for the prognosis model.

For each patch prediction, the classifier assigns a prognosis label using a binary output. A label of "0" indicates a bad prognosis, while a label of "1" indicates a good prognosis.

In this thesis, the same threshold value is used to predict a patient-level prognosis as the optimal threshold selected in the Method described in 4.2.2. The specific threshold value t was set at 0.3720 and is used for comparison with the ratio of metastasis, denoted as R_{meta}, to determine the prognosis of a WSI. The output provides the predicted prognosis of a WSI, denoted as $\hat{y}_p$. If the ratio R_{meta} exceeds the threshold t, the WSI will be predicted to have a "bad" prognosis. Conversely, if the ratio falls below the threshold t, the prognosis is considered as being "Good", as shown in Equation 4.11.

$$R_{meta} = \frac{\#\text{predicted metastasis patches}}{\#\text{All valid patches}} \tag{4.10}$$

$$\hat{y}_p = \begin{cases} 0, & \text{if } R_{meta} \geq t \\ 1, & else \end{cases} \tag{4.11}$$

Method	Author	Comment
Extract patches	C.Andreassen	Modified
Running inference	C.Andrassen	Unaltered
Predict inference	C.Andreassen	Unaltered

Table 4.7: This table provides an overview of the methods used in developing the model responsible for prognosis, along with the original author and modification status [4].

Chapter 5

Experiments and Results

This chapter presents the results of the experiments performed in this thesis. The diagnosis and prognosis model was first evaluated on the old dataset. The diagnosis model was then evaluated on the new dataset, providing a starting point for subsequent experiments. Based on the result, it was concluded that new parameters for the diagnosis model were necessary to create the pipeline. The parameters were found using the validation set D_{NewVal}. Lastly, the new parameters were used for evaluating the entire pipeline.

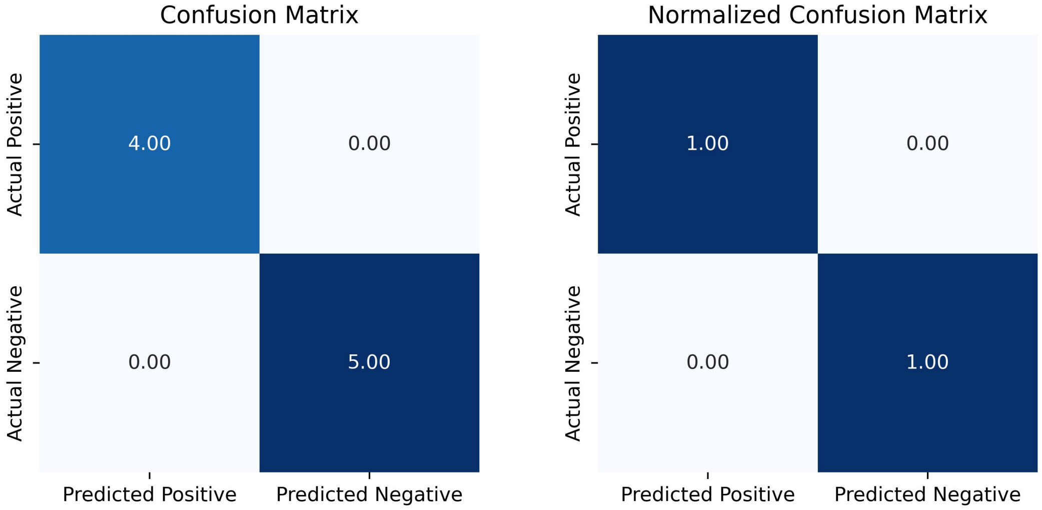
5.1 Validation of the Diagnosis and Prognosis Model on old data set

It is crucial to evaluate the performance of each model separately before developing a pipeline. The evaluation process begins with a separate assessment for each model.

5.1.1 Validation of the Diagnosis Model on D_{DTest}

As mentioned in Section 3.1, the D_{DTest} dataset consists of 9 WSIs; 5 labeled benign and 4 melanoma. Earlier tests on this dataset using parameters $t_p = 0.999$ and $t_r = 0.04$ provided an accuracy of 100% [3]. Using identical parameters on the same dataset should result in the same accuracy as it did. This emphasizes that the model is the same as used in R. Amundsen's thises [3]. The result from the run is displayed in Table 5.1, and the confusion matrix in Figure 5.1 shows the distributions of predictions against the labels.

It is worth noting that the confusion matrixes used in this thesis have predictions along the x-axis and labels at the y-axis. This is emphasized to avoid confusion.



(a) Displaying the true value on the y-axis and the number of predicted WSIs on the x-axis.

(b) Normalized. Showing percentage in decimal format.

Figure 5.1: CM of the diagnosis model on dataset D_{DTest}. Thresholds $t_p = 0.999$ and $t_r = 0.04$ were used, which were the most promising thresholds for the diagnosis model, on D_{DVal}.

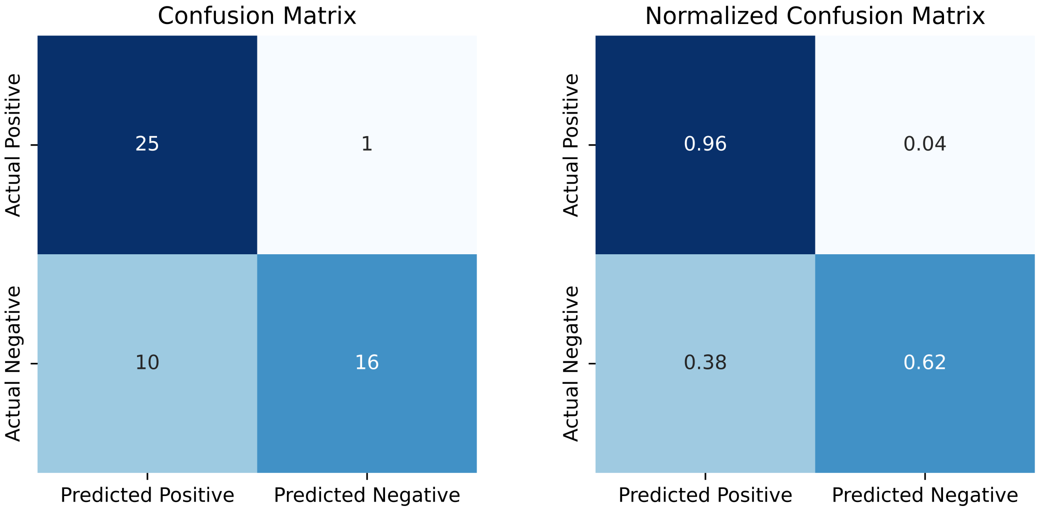
Metric	Specificity	Precision	Recall	Accuracy	F1-Score
Value	1.0000	1.0000	1.0000	1.0000	1.0000

Table 5.1: Performance metrics from diagnosis model on the D_{DTtest}. Thresholds used were $t_p = 0.999$ and $t_r = 0.04$.

5.1.2 Validating the Prognosis Model on D_P

In the next experiment, the prognosis model is evaluated on the same dataset as used for training (D_P). The dataset contains annotations for all but one WSI. This experiment was performed by C. Andreassen in his thesis [4]. The result from his run was 100&, which will be tried to reproduce. The annotations for the images are used as a mask when finding valid patches for this run.

The result from the experiment is shown in Table 5.2, which gave as expected a 100% accuracy. Figure 5.2 shows the distribution of predictions against labels. The prognosis model performed as expected. This validates that both models work and perform according to what was documented in previous work [3][4].



(a) Displaying the true value on the y-axis and the number of predicted WSIs on the x-axis.

(b) Normalized. Showing percentage in decimal format

Figure 5.2: CM of the prognosis model on dataset D_P. The annotation masks provided were used as input masks for finding valid patches to run inference on.

.

Metric	Specificity	Precision	Recall	Accuracy	F1-Score
Value	0.6154	0.7143	0.9615	0.7885	0.8197

Table 5.2: Performance metrics from prognosis model on the D_P using the annotated masks as input for finding valid patches.

5.1.3 Evaluating the Pipeline on $\hat{D}_{\mathbf{P}}$

As both models are evaluated, they will be tested on a dataset containing the training set for the prognosis model. This experiment tests both models' initial ability to be integrated into a pipeline without finetuning. Moreover, if the diagnosis produces correct patch predictions for the WSIs, the prognosis model should be provided with the same ROI. Conclusively, the accuracy should be 100%.

D_P contains some of the images used for training and validating the diagnosis model. To give a more fair result, it was decided to exclude these images and define a new dataset, $\hat{D}_P$.

Defining the test set for D_P: $\hat{D_P}$

D_D is the dataset used for the diagnosis model. The dataset used for training and evaluating the model is denoted D_{DTrain} and D_{DVal}, respectively. The dataset D_{DTest} is the dataset used for testing the prognosis model. A definition of D_D is shown in equation 5.1.

$$D_D = D_{DTrain} \cup D_{DVal} \cup D_{DTest} \tag{5.1}$$

In equation 5.2 $hatD_P$ is defined.

$$\hat{D}_P = D_P \setminus (D_{DTrain} \cup D_{DVal}) \tag{5.2}$$

The modified dataset $\hat{D}_P$ is obtained by removing any data from D_P that is present in either D_{DTrain} or D_{DVal}, as depicted in Figure 5.3.

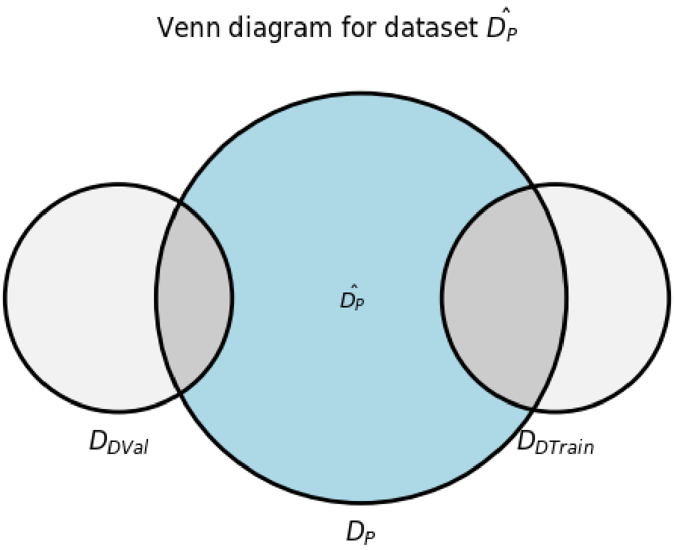


Figure 5.3: A Venn diagram demonstrating $\hat{D}_p$, illustrated in blue, with the execution of the dataset D_{DVal} and D_{DTrain}

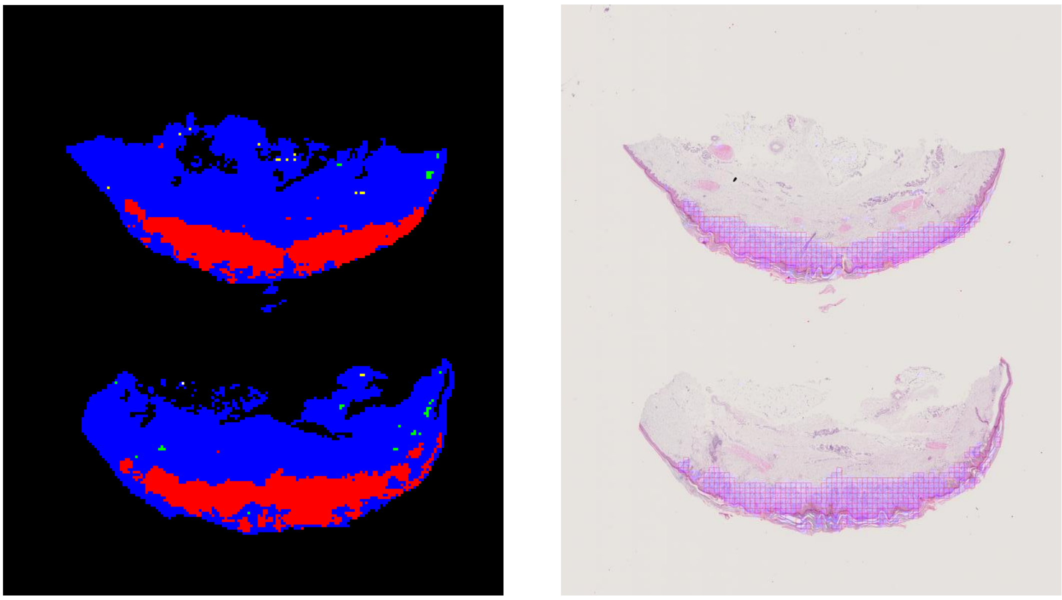
It was decided by C. Andreassen that the limited number of WSIs provided (D_P) was going to be used for training. This resulted in all WSIs from the dataset being introduced to the prognosis model [4].

Diagnosis and Prognosis Inference on $\hat{D}_P$

In the following experiment, the inference was performed on $\hat{D}_P$, containing 38 WSIs with melanoma; 17 of the images labeled bad prognosis (metastasis) and 21 labeled good prognosis. The diagnosis model was first used for generating the malignant patches. The malignant patches were then fed into the prognosis model. The diagnosis model used thresholds found to be optimal by R. Amundsen [3] in his thesis, which was $t_p = 0.999$ and $t_r = 0.04$. As mentioned previously, if the diagnosis model generates identical patches from the annotated masks, the result should be the same (accuracy=100%).

All 38 WSIs were predicted to be malignant by the diagnosis model, resulting in all WSIs being presented to the prognosis model. This means that the diagnosis model predicted all images correctly, resulting in an accuracy of 100%.

Figure 5.4 illustrates the integration between the diagnosis output and the prognosis in the pipeline. On the figure to the left side, a prediction mask generated by the diagnosis model is shown. The red pixels within the mask represent malignant patches identified by the diagnosis model. These malignant patches are then extracted and combined to create the malignant mask, which is then used as input for the prognosis model. The figure on the right depicts the resulting patch extraction after preprocessing for the prognosis model. These patches are selected based on the malignant mask, ensuring that only relevant regions are considered for further analysis by the prognosis model.

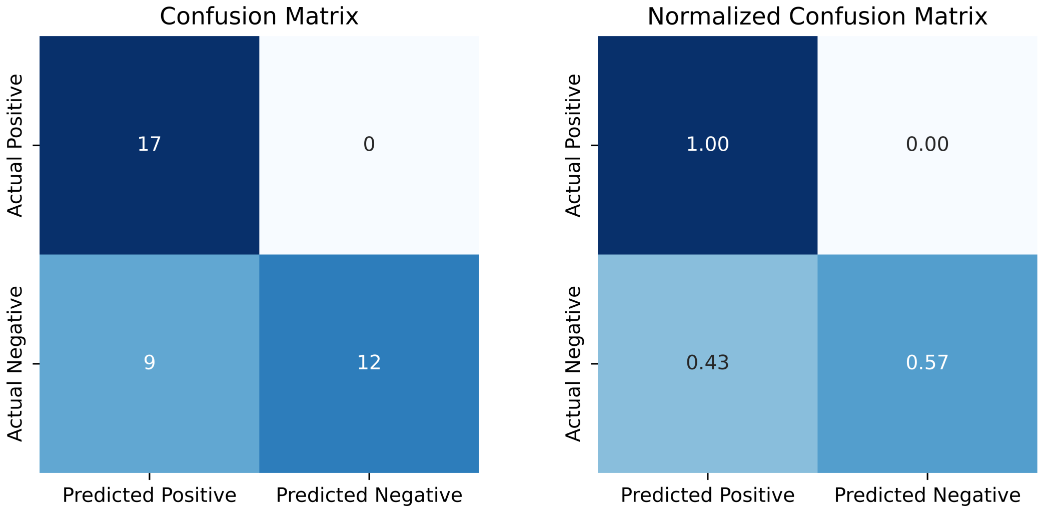


(a) A WSI is provided to the diagnosis model, and patch predictions are generated. The red pixels represent malignant predictions.

(b) Patch extraction for prognosis model, using diagnosis prediction as input. Each pink square represents a patch predicted malignant by the prognosis model. These patches are then used as input for the prognosis model.

Figure 5.4: Integration between the diagnosis output and the prognosis. Figure (a) represents the predicted patches from the diagnosis model, which collectively form the malignant mask. Figure (b) illustrates how the prognosis model uses the malignant mask to extract valid patches.

The confusion matrix presented in Figure 5.5 shows the results from the prognosis model on the dataset $\hat{D}_P$. The results are different from the initial prognosis test using annotated masks, indicating some discrepancies from the masks presented to the prognosis model. The model still predicted all WSI labeled metastasis correctly but misclassified nine.



(a) Displaying the true value on the y-axis and the number of predicted WSIs on the x-axis. Dataset $\hat{D}_P$.

(b) Normalized confusion matrix, showing percentage in decimal format.

Figure 5.5: Confusion matrix of results from the prognostic model on dataset $\hat{D}_P$ using predicted masks from diagnosis model. Thresholds $t_p = 0.999$ and $t_r = 0.04$ were used, which were the most promising thresholds for the model, on D_{DVal}.

Table 5.3 displays the evaluation metrics after running the pipeline on dataset $\hat{D}_P$.

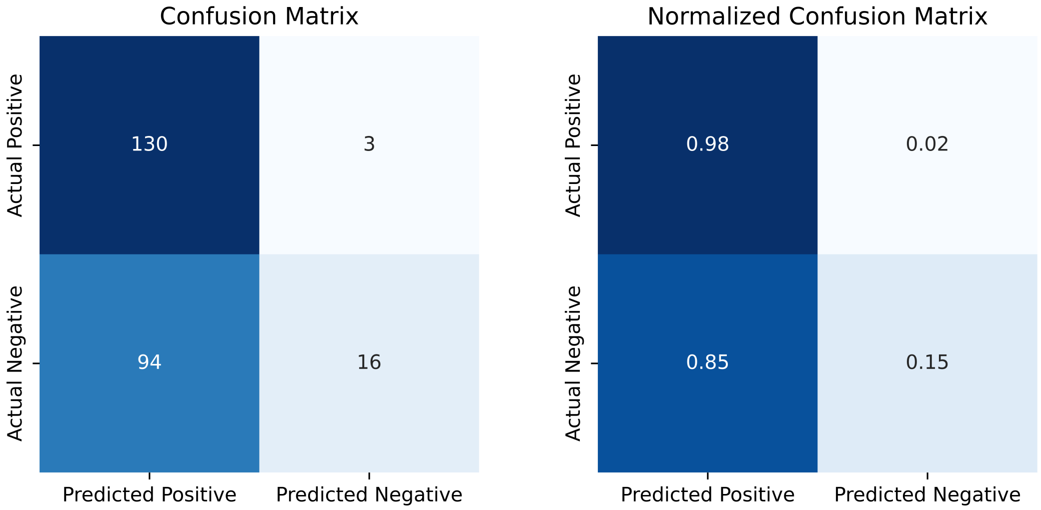
Metric	**Specificity**	**Precision**	**Recall**	**Accuracy**	**F1-Score**
Score	0.5714	0.6538	1.0000	0.7632	0.7907

Table 5.3: Evaluation Metrics after running the pipeline on dataset $\hat{D}_P$.

5.2 Diagnosis Inference on New Dataset with Old Thresholds

In this experiment, the diagnosis model's performance on new data will be evaluated. The dataset used is D_{New}, which contains WSIs. The dataset contains 133 malignant images and 110 benign. The initial thresholds for patch-level classification (t_p) and patient-level classification (t_r) are set at 0.999 and 0.04, respectively. The thresholds used are the same as in the previous experiment, where the model predicted all images correctly (accuracy=100%).

The model got an accuracy of 0.6008, which corresponds to the model correctly classifying only 60% of the WSIs. With a recall of 0.9774, the model predicts almost all melanoma cases correctly but struggles to predict benign cases. The results of the experiment are presented as a confusion matrix in Figure 5.6.



(a) Displaying the true value on the y-axis and the number of predicted WSIs on the x-axis.

(b) Normalized. Showing percentage in decimal format.

Figure 5.6: Confusion matrix of results from diagnosis model on dataset D_{New}. Thresholds $t_p = 0.999$ and $t_r = 0.04$ were used, which were the most promising thresholds for the model on dataset D_{DVal}.

Eval Metric	Specificity	Precision	Recall	Accuracy	F1-Score
Score	0.1455	0.5804	0.9774	0.6008	0.7283

Table 5.4: Results from running inference with diagnosis model on D_{new} with thresholds $t_p = 0.999$ and $t_r = 0.04$.

From these results, it is evident that the diagnosis model is generalizable for new data with the current settings. This highlights the need for further optimization of the model. This will be done by tuning the parameters for the model.

5.3 Optimizing Thresholds for Diagnosis Model

The previous result showcased the evident drop in performance when using the new dataset. This could be a result of poorly chosen thresholds. The threshold for the diagnostic model was determined using the D_{DVal} dataset, which only contains eight images. To enhance the model's performance, new thresholds can be identified using dataset D_{new}. The following section will research threshold values in addition to looking at the data distribution for the new dataset D_{NewVal}.

5.3.1 Evaluating the Patch-Level Threshold (t_p)

The first threshold to be evaluated is t_p, which is used to threshold the patch-level predictions from the diagnosis model

Values between 0.99 and 1.00 were tested. The new dataset does not contain annotation, so a visual inspection was used to select reasonable values. The prediction masks are highly dependent on the t_p value, which is used as input by the prognosis model. This has to be taken into account when deciding the threshold value. Figure 5.7 displays a representative collection of WSIs from dataset D_{New} and the respective prediction masks with a range of thresholds.

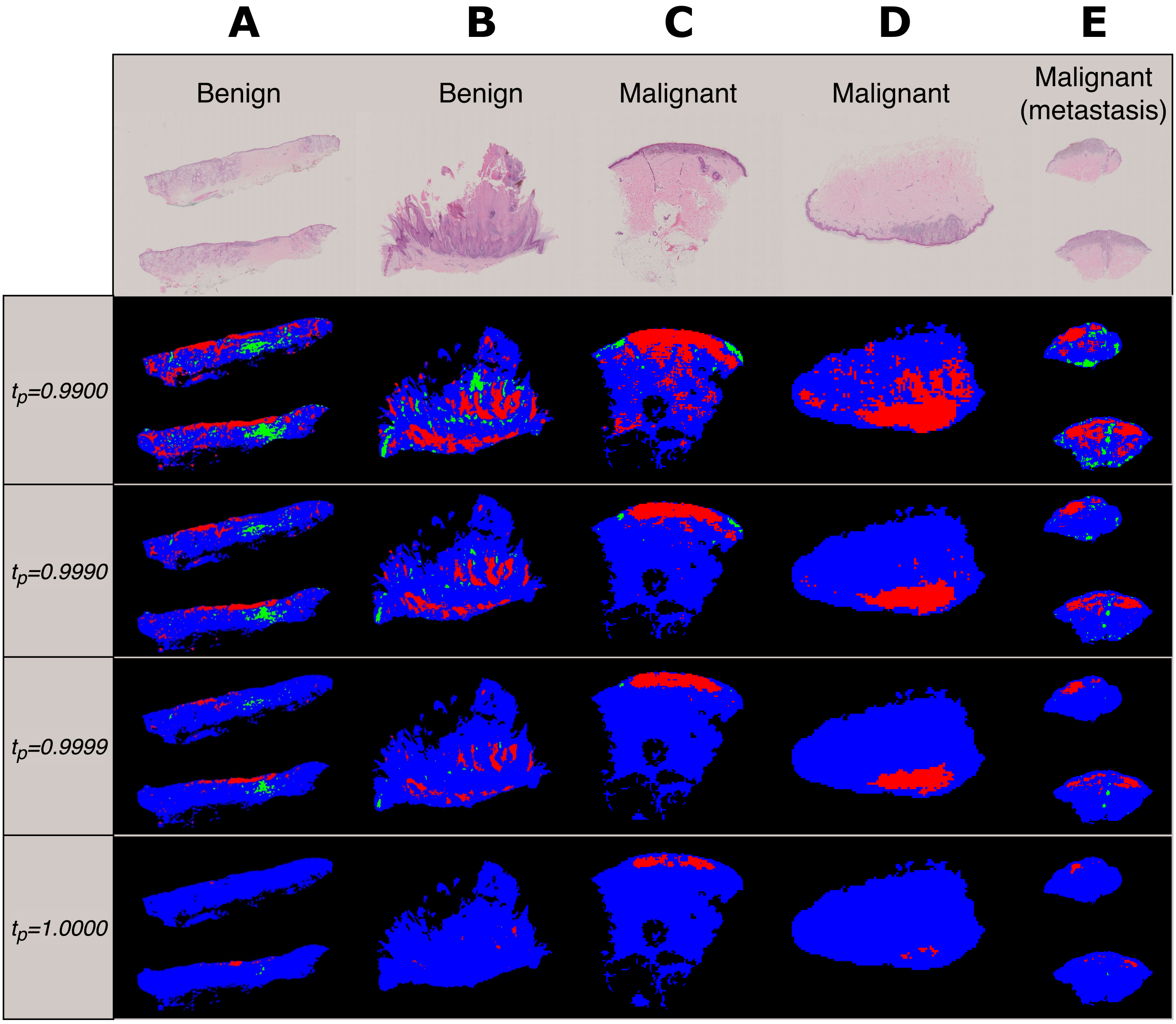


Figure 5.7: Prediction masks table for the diagnosis model, representing dataset D_{New}. Each column displays a WSI with its corresponding letter and diagnosis label at the top. Each row presents the prediction masks for each WSI with the given t_p on the left side. Red pixels represent patches predicted malignant, green represents patches predicted benign, and blue are patches predicted as other tissue.

- At threshold value $t_p = 0.9900$, all WSIs (A-E) exhibit scattered prediction masks.
- At threshold value $t_p = 0.9990$, a relatively accurate prediction mask is observed with some islands, particularly for A and B. The prediction mask displays a good

region of malignant predictions for C, D, and E, which are labeled malignant.

- At threshold value $t_p = 0.9999$, smaller malignant prediction regions are seen, but islands are reduced. The prediction mask still provides a reasonable number of patches for the prognosis model. All WSIs have a small ratio of malignant patches compared to other tissue patches (blue). Malignant WSIs (C-E) exhibit more isolated neighborhoods of malignant regions, while malignant patches A and B are more isolated.
- At threshold value $t_p = 1.0000$, the prediction mask contains few patches predicted as malignant, resulting in a limited number of patches sent to the prognosis model.

From the figure, it was concluded that using $t_p = 0.999$ and $t_p = 0.9999$ for the diagnosis model results in the most suitable prediction mask for the prognosis model.

5.3.2 Best Method for Calculating M_{rate}

Valuable insight can often be obtained by examining the dataset. In this section, the melanoma-rate calculation will be evaluated based on the data distribution. The distribution will be looked at as a collection with respect to each label and for each WSI.

Firstly, the data distribution for patch predictions is examined to identify an appropriate method for patient-level prediction. Subsequently, the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) are employed with $t_p = 0.999$ and 0.9999 to find the optimal patient-level threshold (t_r).

In order to achieve a good image-level classification, it is important to establish a suitable method for deciding what rate of malignant patches have to be present in a WSI for it to be classified as malignant. The following equation is taken from section 4.3.2 and depicts how the M_{rate} was calculated in R. Amundsens' thesis [3]:

$$MB_{\text{rate}} = \frac{\text{Malignant Area}}{\text{Lesion Area}} = \frac{\sum_p \hat{y}_{p\text{M}}}{\sum_p(\hat{y}_{p\text{M}} + \hat{y}_{p\text{B}})}$$

This thesis proposes considering all tissue areas within the WSI for analysis. As discussed in Section 4.3.2, the calculation of M_{rate} can involve examining the malignant matches in relation to all tissue areas (MT_{rate}), as shown in the Equation below:

$$MT_{\text{rate}} = \frac{\text{Malignant Area}}{\text{All Tissue Areas}} = \frac{\sum_p \hat{y}_{p\text{M}}}{\sum_p(\hat{y}_{p\text{M}} + \hat{y}_{p\text{B}} + \hat{y}_{p\text{NT}})}$$

Analyzing the Patch Distribution for D_{New}

To gain a better understanding of the importance of selecting an appropriate M_{rate}, the distribution of patches for each prediction class in the diagnosis model with $t_p = 0.999$ and $t_p = 0.9999$ is presented in Figure 5.8 and Figure 5.9, respectively.

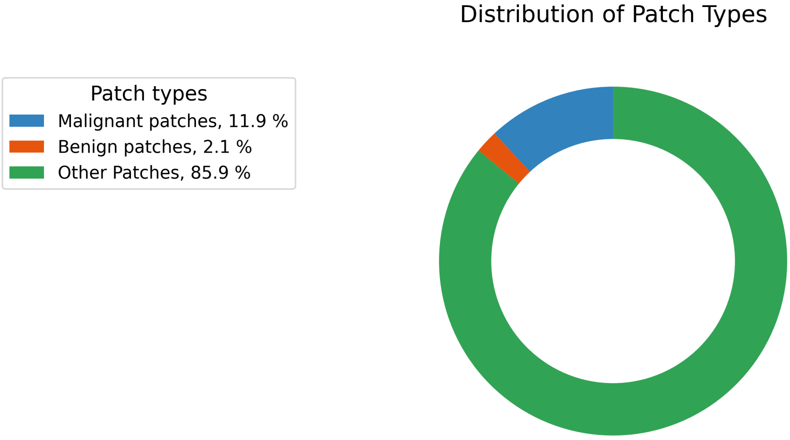


Figure 5.8: Distribution of all patches in the D_{New} dataset with $t_p = 0.999$. The patches are categorized based on the predictions made by the diagnosis model and are presented as percentages.

For $t_p = 0.999$, malignant patches constitute nearly 12% of all valid patches, while benign patches account for about 2%. Other patches make up 85.9%, resulting in the following M_{rate}:

$$MB_{rate} = \frac{11.9}{11.9 + 2.1} = 0.850$$
$$MT_{rate} = \frac{11.9}{100} = 0.119$$

This indicates a relatively high MB_{ratio} but a low MT_{rate}.

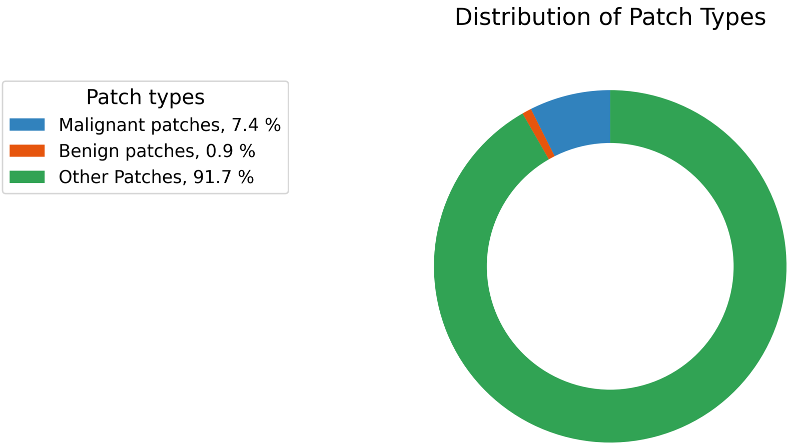


Figure 5.9: Distribution of all patches in D_{New} with $t_p = 0.9999$. The patches are categorized based on the predictions made by the diagnosis model and are presented as percentages.

When considering a threshold of $t_p = 0.9999$, we obtain a mean $MB_{rate} = 0.892$ and a mean $MT_{rate} = 0.074$. The mean MB_{rate} increases, while the MT_{rate} decreases compared to $t_p = 0.999$. This can be attributed to a 37.8% decrease in malignant patch predictions and a 57.1% decrease in benign patch predictions, suggesting that the model predicts malignant patches with higher confidence than benign ones.

To further explore the relationship between malignant and benign patch predictions, Figure 5.10 illustrates the correlation between the total number of predicted patches and proportions classified as malignant or benign, with respective trendlines.

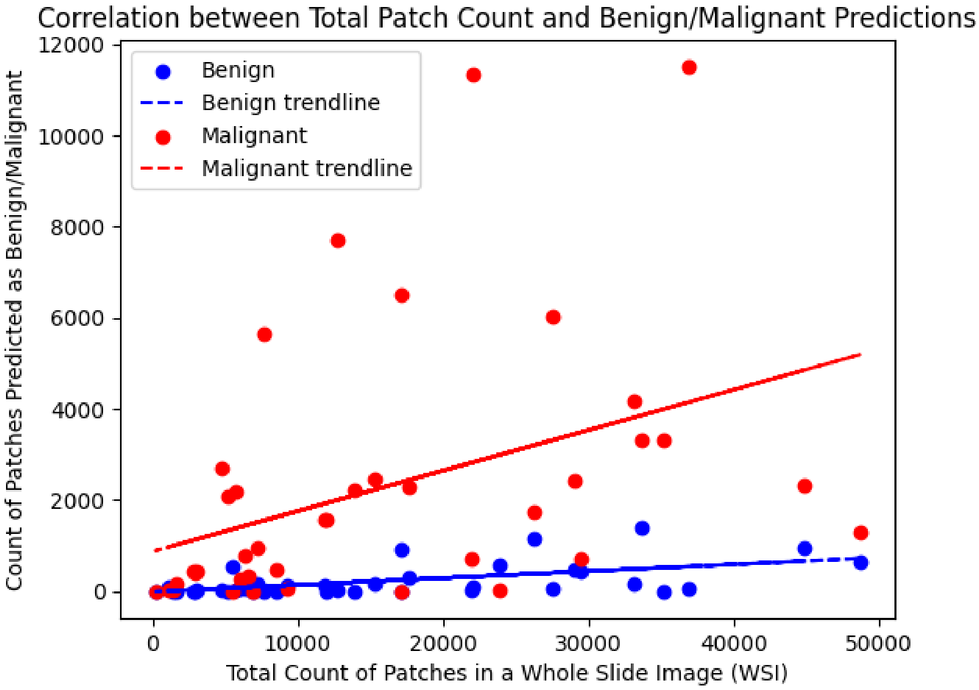


Figure 5.10: Plot showing the number of valid patches in a single WSI on the x-axis and the y-axis representing the number of patches predicted malignant and benign. There are two corresponding points for each WSI (blue and red). The blue represents the number of patches predicted benign, while the number of malignant patches is represented by red. Each plot pair (red/blue) is a single WSI from the dataset $D_{NewTest}$. Correlation between the total number of valid patches in a WSI and the number of patches predicted malignant/benign are shown in trendlines. The value used for selecting the threshold for the patch predictions t_p was 0.999.

The trendlines provide valuable insights into the model's behavior regarding the prediction of malignant and benign patches. The observed trend reveals that the model tends to predict more malignant patches than benign ones, indicating higher prediction confidence for malignant patches. However, various unknown factors may influence this correlation, necessitating further experiments using both M_{rate} methods before determining the most appropriate one.

5.3.3 Finding Optimal Patient-Level Thresholds on New Dataset

Based on the previous experiments, threshold values t_p of 0.999 and 0.9999 yielded promising results when generating prediction masks, as displayed in Figure 5.7. The results from evaluating the M_{rate} methods were not conclusive, resulting in both methods being used in further experiments. In the following sections, experiments using fixed t_p and varying t_r will be used for generating ROC curves and calculating the corresponding area under the curve (AUC) scores. The AUC score serves as an aggregate measure of performance, summarizing the model's ability to distinguish between malignant and benign patches across all possible t_r thresholds. A higher AUC score indicates better overall performance, with values ranging from 0 to 1. The experiments of finding optimal thresholds will be performed on the validation set D_{NewVal} and will, in the next section, be tested D_{New} to evaluate the performance of the entire pipeline.

ROC AUC for Diagnosis Model with tp = 0.999

In the following experiments, the t_p value is set to 0.999, while exploring a range of possible classification thresholds for t_r. The goal is to generate two ROC plots, one for each M_{rate} method. The t_r, True Positive Rate (TPR), and False Positive Rate (FPR) values for the ROC plot were autogenerated by the 'roc_curve()' and 'auc()' functions from the *sklearn.metrics* library [26].

Figure 5.11 displays the ROC plot for the MB_{rate} method, while figure 5.12 shows the ROC plot for the MT_{rate} method. Both for dataset D_{NewVal}.

In each ROC plot, data points representing every fifth t_r threshold are depicted. These points represent the TPR and FPR for the corresponding t_r thresholds. The AUC score is provided for each ROC plot, providing an evaluation metric for the chosen threshold t_p and M_{rate} method.

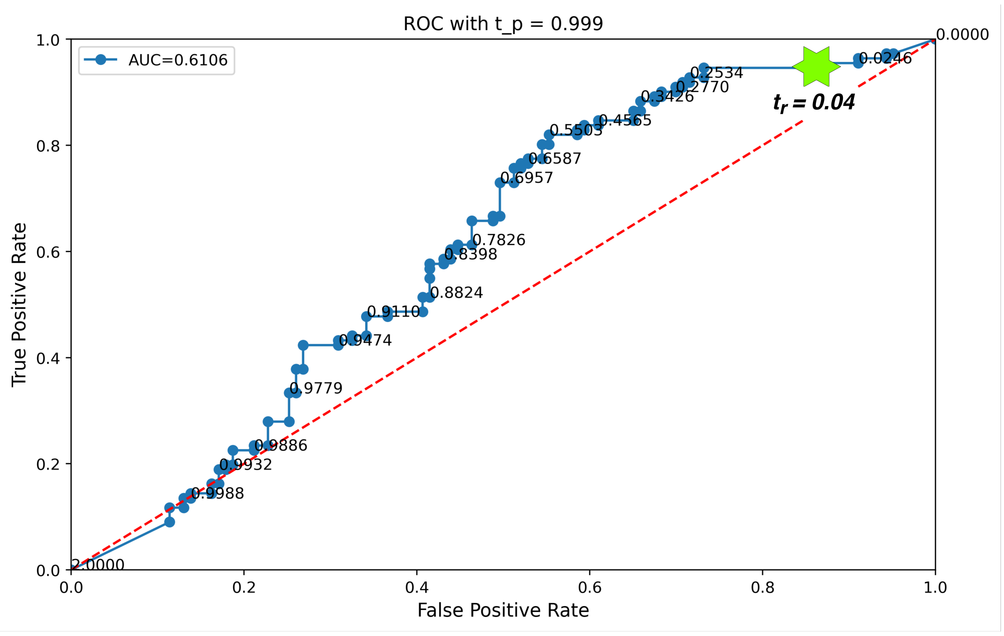


Figure 5.11: ROC plot diagnosis for the model with $t_p = 0.999$ and MB_{rate} method on dataset D_{NewVal}. The threshold t_r denoted by a green star was used in the previous diagnosis experiments.

The diagnosis model's performance using the MB_{rate} method was relatively poor across all tested t_r values. An AUC value of 0.6106 indicates that the t_p threshold or MB_{rate} method might not be suitable for achieving desirable results. It is important to consider that R. Amundsen's thesis [3] employed the MB_{ratio} method with a t_r ratio of 0.04, which resulted in a remarkable accuracy of 100% on the dataset D_{DTest}. Comparing this to the previous experiment's results in Section 5.2, where the same thresholds were used, provides valuable insights. The threshold value t_r used by R. Amundsen's thesis is denoted by a green star in Figure 5.11.

Next, a new ROC curve with corresponding AUC was generated using the MT_{rate} method instead of the MB_{rate}, as displayed in Figure 5.12. Using the same threshold t_p of 0.999 and dataset D_{NewVal}.

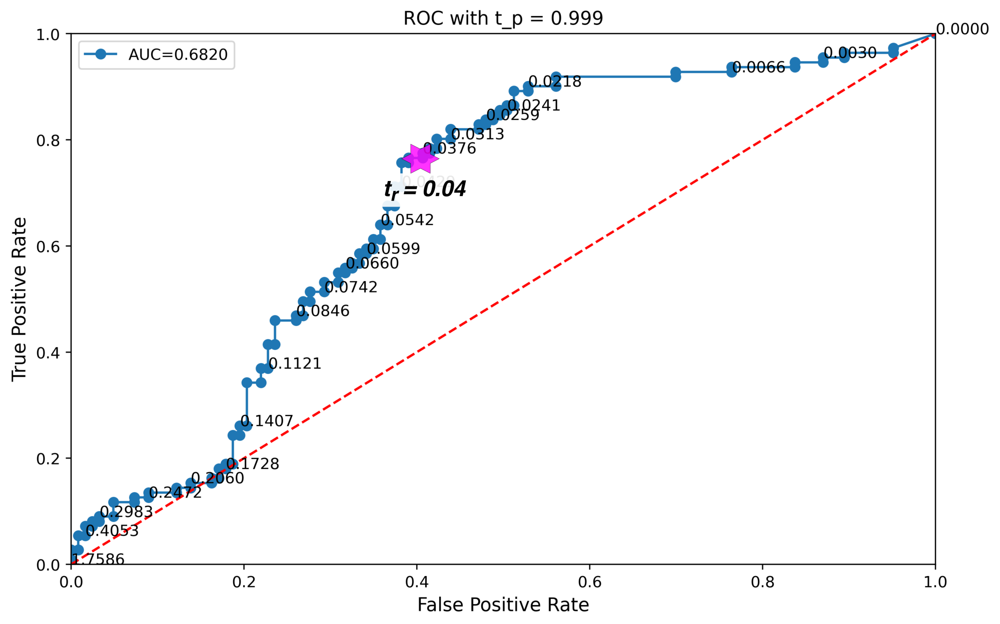


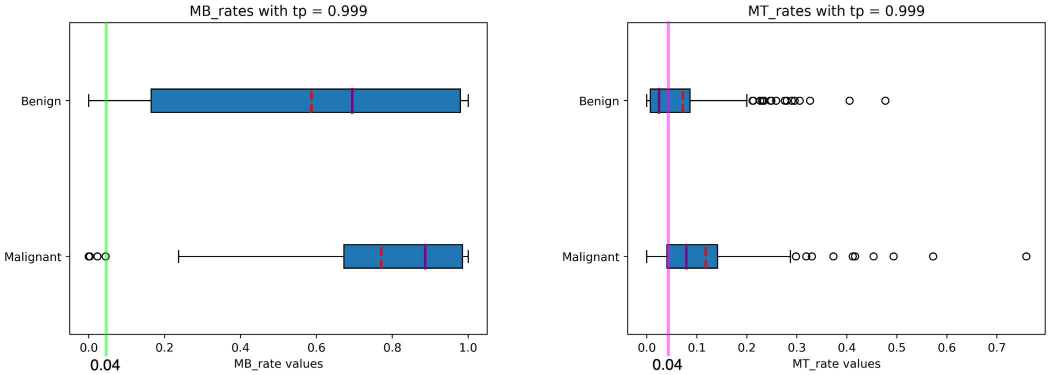
Figure 5.12: ROC plot for diagnosis model with $t_p = 0.999$ and MT_{rate} method on dataset D_{newVal}. The purple star denotes a promising t_r threshold used in subsequent experiments.

The model demonstrated improved performance with an AUC value of 0.6820. This experiment was conducted on an identical dataset and utilized the same t_p threshold. The increase in the AUC score suggests that the MT_{rate} method is more effective for calculating M_{rate} when using a threshold $t_p = 0.999$. Internal discussions with pathologists at SUH revealed a preference for avoiding false negative diagnoses, despite the potential for false positives. A t_r value of 0.04 was determined to be a reasonable balance between TPR and FPR, particularly in assisting pathologists.

Table 5.5 shows the mean value and the standard deviation of the M_{rate} for all WSIs in the D_{NewVal} dataset with a threshold $t_p = 0.999$. There is an overlap for both M_{rate} methods when taking into account the standard deviation. A visual representation of the M_{rate} distribution is depicted in 5.13, using the MB_{rate} to the left and MT_{rate} to the right. The green and pink lines in the figure denote the threshold value t_r 0.04.

	$\mathbf{MB_{rate}}$	$\mathbf{MT_{rate}}$
benign WSIs	0.5866(±0.3879)	0.0717(±0.0982)
malignant WSIs	0.7706(±0.2843)	0.1176(±0.1265)

Table 5.5: Mean and standard deviation values for MB_{rate} and MT_{rate} methods for benign and malignant WSIs. t_p threshold used was 0.999 on D_{NewVal}



(a) Boxplot of MB_{rate} with $t_p = 0.999$. The green line denotes the threshold $t_r = 0.04$ for MB_{rate} method, which was used in the previous experiment.

(b) Boxplot of MT_{rate} with $t_p = 0.999$. The purple line denotes the threshold $t_r = 0.04$ for MT_{rate} method, which was found to be promising from the ROC curve in the previous experiment.

Figure 5.13: Box plots of MB_{rate} and MT_{rate} for benign and malignant WSIs with $t_p = 0.999$. The box (blue area) represents the interquartile range (IQR), which contains 50% of the data, with the lower edge at the first quartile (25th percentile) and the upper edge at the third quartile (75th percentile). The horizontal purple line inside the box marks the median value, while the red line represents the mean value. The lines extend from the box to show the range of data within 1.5 times IQR, and any data points outside this range are plotted as individual points.

The box plots provide visual representations of the different variations and mean values for both benign and malignant WSIs for each M_{rate} method. By analyzing the boxplot, it becomes clear that the majority of the malignant patches are correctly predicted as malignant when using a $t_p = 0.999$, $t_r = .04$, and the MT_{rate} method on the validation set D_{NewVal}. As seen from the ROC Figure 5.11, the threshold values $t_p = 0.999$ and $t_r = 0.04$ will predict most WSIs as malignant when using the MB_{rate} method.

ROC AUC Analysis for Diagnosis Model with $t_p = 0.9999$

This section presents an analysis of the ROC curve with the corresponding AUC score for the diagnosis model with a patch-level threshold, t_p, of 0.9999. The experiment involves varying the patient-level threshold (t_r) and generating two ROC plots for different M_{rate} methods. The True Positive Rate (TPR) and False Positive Rate (FPR) values for the ROC plot are computed using the 'roc_curve()' and 'auc()' functions from the *sklearn.metrics* library [26].

Figures 5.14 and 5.15 displays the ROC plots for MB_{rate} and MT_{rate} methods, respectively. Each graph represents the TPR and FPR for different t_r thresholds, with points plotted at every fifth threshold. The AUC score, included in each graph, serves as an evaluation metric for the overall performance of the model.

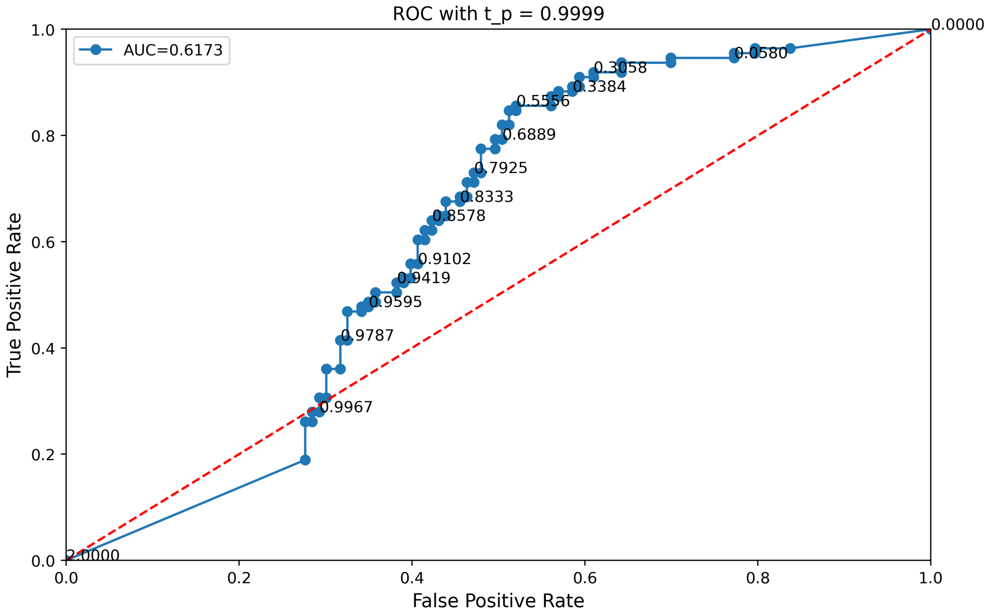


Figure 5.14: ROC plot for diagnosis model with $t_p = 0.9999$ and MB_{rate} method on dataset D_{newVal}. The *line of no discrimination* is represented as a stippled red line, with all t_r over 0.999 above it, indicating that the model can differentiate between malignant and benign patches beyond randomness at this threshold level.

The AUC score obtained from Figure 5.14 is 0.6173, which is comparable to the previous experiment's score of 0.6106 at $t_r = 0.999$. This indicates that when using the MB_{rate} method with $t_p = 0.9999$, there is a slight improvement in TPR as t_r decreases compared to when using $t_p = 0.999$.

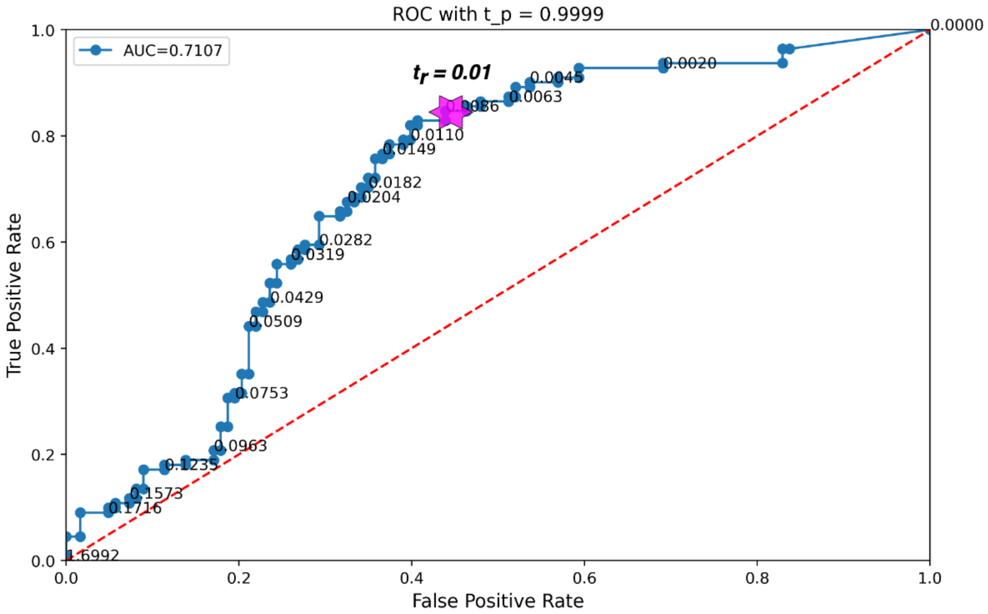


Figure 5.15: ROC plot for diagnosis model with $t_p = 0.9999$ and MT_{rate} method on dataset D_{newVal}. All values of t_r are above the *line of no discrimination*, indicating promising results at this threshold level, particularly at $t_r = 0.01$, denoted by a purple star, which will be used in subsequent experiments.

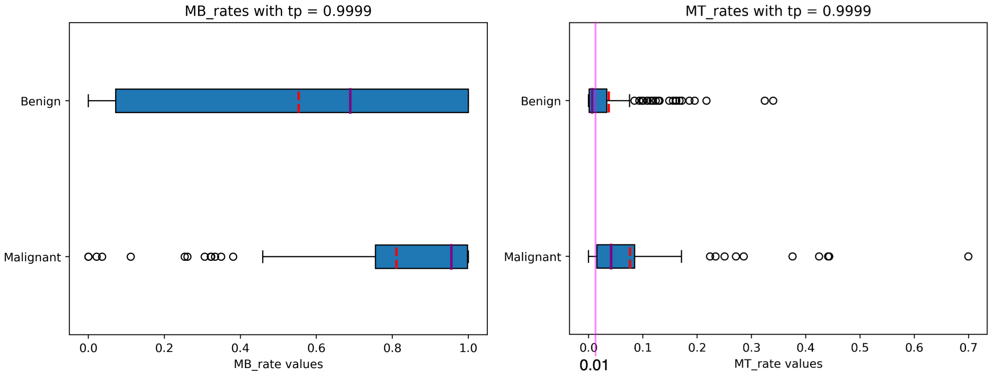
Figure 5.15 shows that the combination of $t_p = 0.9999$ and MT_{rate} yields an AUC score of 0.7107, suggesting promising results at various t_r values, particularly at $t_r = 0.01$. Given that a high TPR is more critical than a low FPR in this context, this threshold will be used in future experiments.

Table 5.6 provides the mean and standard deviation values for the MB_{rate} and MT_{rate} methods applied to benign and malignant WSIs. The patch-level threshold (t_p) used in this experiment was 0.9999 on the dataset D_{NewVal}. For benign WSIs, the mean MB_{rate} is 0.5534 with a standard deviation of 0.4261, while the mean MT_{rate} is significantly lower at 0.0371 with a standard deviation of 0.0644. For malignant WSIs, both methods yield higher rates than benign WSIs. The mean MB_{rate} is 0.8107 with a standard deviation of 0.2797, and the mean MT_{rate} is 0.0764 with a standard deviation of 0.1082.

	$\mathbf{MB_{rate}}$	$\mathbf{MT_{rate}}$
benign WSIs	0.5534(±0.4261)	0.0371(±0.0644)
malignant WSIs	0.8107(±0.2797)	0.0764(±0.1082)

Table 5.6: Mean and standard deviation values for MB_{rate} and MT_{rate} methods on benign and malignant WSIs with a patient-level threshold (t_p) of 0.9999 on dataset D_{NewVal}.

Figure 5.16 presents box plots of the distribution of MB_{rate} and MT_{rate} for both benign and malignant WSIs at a patch-level threshold (t_p) of 0.9999.



(a) Boxplot of MB_{rate} with $t_p = 0.9999$

(b) Boxplot of MT_{rate} with $t_p = 0.9999$. The purple line denotes the threshold $t_r = 0.01$ for MT_{rate} method, which was found to be promising from the ROC curve in the previous experiment.

Figure 5.16: Box plots of MB_{rate} and MT_{rate} for benign and malignant WSIs with $t_p = 0.9999$. The box (blue area) represents the interquartile range (IQR), which contains 50% of the data, with the lower edge at the first quartile (25th percentile) and the upper edge at the third quartile (75th percentile). The horizontal purple line inside the box marks the median value, while the red line represents the mean value. The lines extend from the box to show the range of data within 1.5 times IQR, and any data points outside this range are plotted as individual points.

These visualizations provide further insights into the performance of the diagnosis model at different thresholds, aiding in understanding its effectiveness in distinguishing between benign and malignant patches.

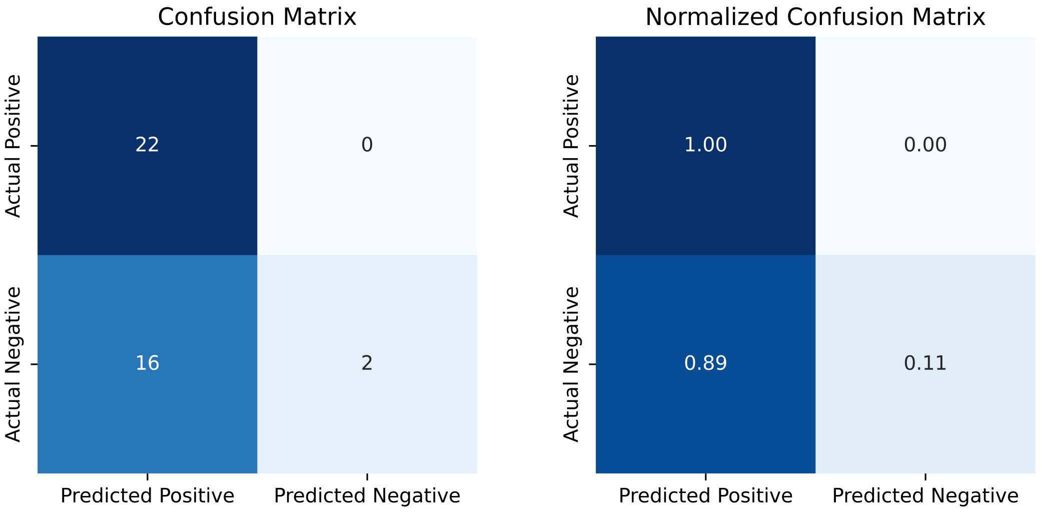
5.4 Testing Complete Pipeline on New Test Dataset ($D_{NewTest}$)

This section discusses the performance assessment of the integrated pipeline, comprising both diagnosis and prognosis models, on a distinct test set ($D_{NewTest}$). The evaluation first considers the threshold and M_{rate} settings from R. Amundsens' thesis[3], followed by examining the thresholds identified in the previous experiments. For each model, diagnosis, and prognosis, confusion matrices are presented along with normalized results, providing a comprehensive performance overview.

5.4.1 Testing the Pipeline with Old Thresholds

In the initial experiment, the pipeline is evaluated using predefined thresholds: a patch-level threshold (t_p) of 0.999 and a patient-level threshold (t_r) of 0.04. The malignancy rate (M_{rate}) was computed using the ratio between malignant patches and lesions (MB_{rate}).

Figure 5.17 displays the confusion matrices for the diagnosis model in absolute numbers and normalized form. Performance metrics are summarized in Table 5.7. The model demonstrated high recall (1.0000), indicating accurate identification of true positive cases, but low specificity (0.1111), implying a high false-positive rate.



(a) Confusion matrix from the diagnosis model using old thresholds and MB_{rate}. Displaying the true value on the y-axis and the number of predicted WSIs on the x-axis.

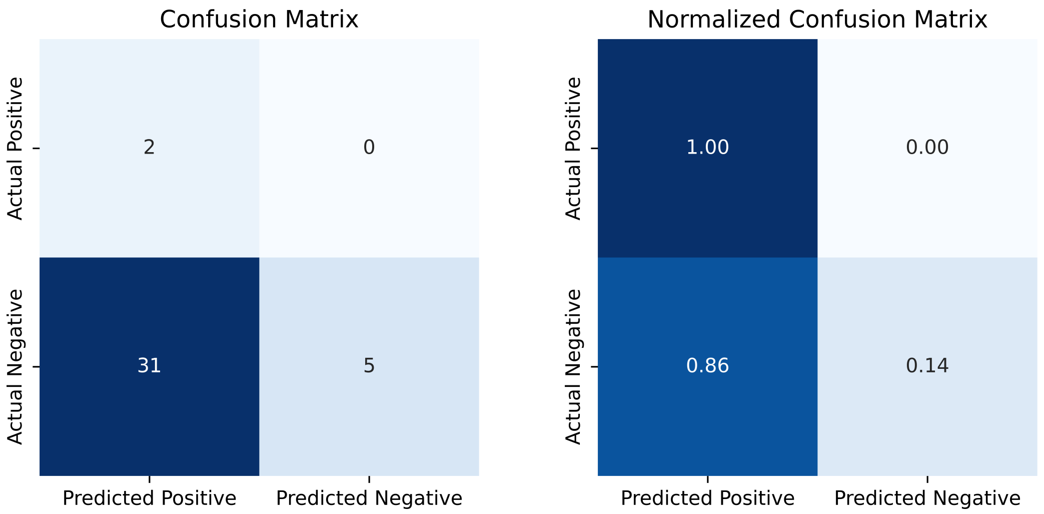
(b) Normalized. Showing percentage in decimal format.

Figure 5.17: Confusion matrix of results from diagnosis model. Settings used was $t_p = 0.999$ and $t_r = 0.04$ on dataset $D_{NewTest}$.

Metric	Specificity	Precision	Recall	Accuracy	F1-Score
Value	0.1111	0.5789	1.0000	0.6000	0.7333

Table 5.7: Performance metrics from the diagnosis model on the $D_{NewTest}$ using old thresholds

The prognosis model's confusion matrices are depicted in Figure 5.18, with corresponding performance metrics in Table 5.8. Like the diagnosis model, the prognosis model also achieved high recall (1.0000) but had lower specificity (0.1389).



(a) Confusion matrix from the diagnosis model using input generated from the diagnosis model. Displaying the true value on the y-axis and the number of predicted WSIs on the x-axis.

(b) Normalized. Showing percentage in decimal format.

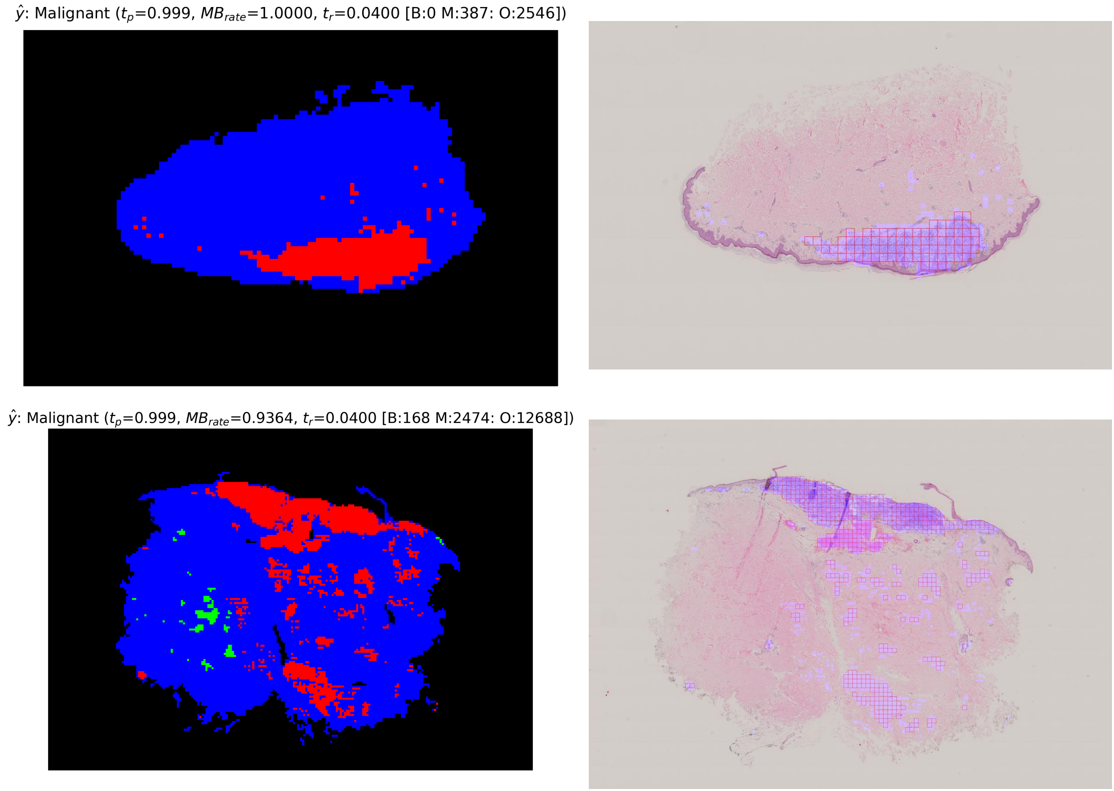
Figure 5.18: Confusion matrix of results from the prognosis model on dataset $D_{NewTest}$. Settings used for the diagnosis model were $t_p = 0.999$ and $t_r = 0.04$.

Metric	Specificity	Precision	Recall	Accuracy	F1-Score
Value	0.1389	0.0606	1.0000	0.1842	0.1143

Table 5.8: Performance metrics from prognosis model on the D_{New} running through the whole pipeline, using the old threshold for finding valid patches.

These results indicate that while both models are highly sensitive and accurately identify all positive cases, they also have high false-positive rates, potentially leading to overdiagnosis or overtreatment in clinical scenarios.

Figure 5.19 displays two examples of output prediction masks from the diagnosis model integrated with the input mask for the prognosis model from the experiment.


$\hat{y}$: Malignant (t_p=0.999, MB_{rate}=1.0000, t_r=0.0400 [B:0 M:387: O:2546])
$\hat{y}$: Malignant (t_p=0.999, MB_{rate}=0.9364, t_r=0.0400 [B:168 M:2474: O:12688])

(a) Prediction masks and patient-level predicting from the diagnosis model A summary of the results and settings are displayed above each prediction mask.

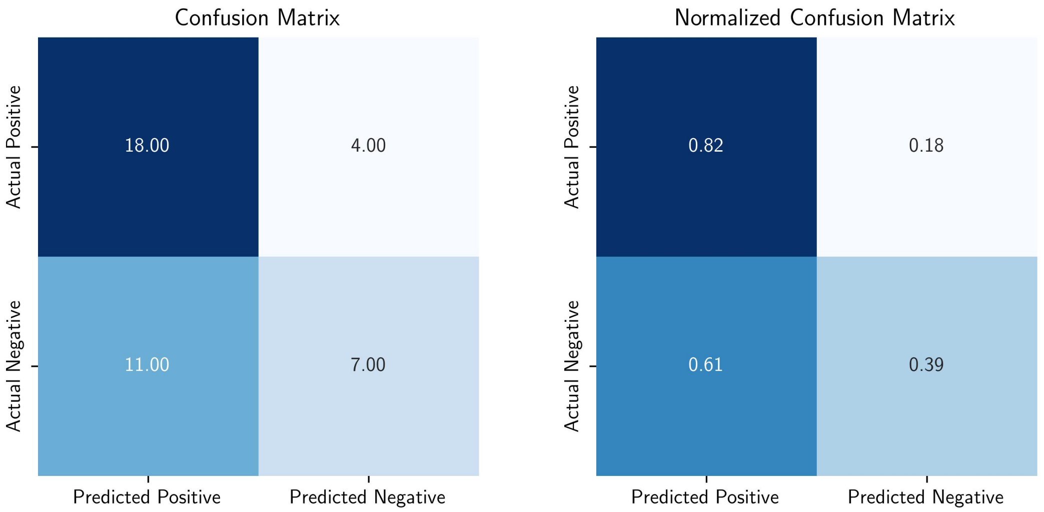
(b) Patch extraction for prognosis model, using diagnosis prediction masks as input.

Figure 5.19: Integration between the diagnosis output prediction mask and the prognosis input mask. Settings used was MB_{rate} method, $t_p = 0.999$ and $t_r = 0.04$ on dataset $D_{NewTest}$.

5.4.2 Testing the Pipeline using New Thresholds ($t_p = 0.999$, $t_r = 0.04$, MT_{rate})

The subsequent experiment tests the pipeline using the MT_{rate} method with a patch-level threshold (t_p) of 0.999 and a patient-level threshold (t_r) of 0.04.

Figure 5.20 displays the confusion matrices for the diagnosis model, both in absolute numbers and normalized form. Performance metrics are summarized in Table 5.9.



(a) Showing the number of predicted WSIs

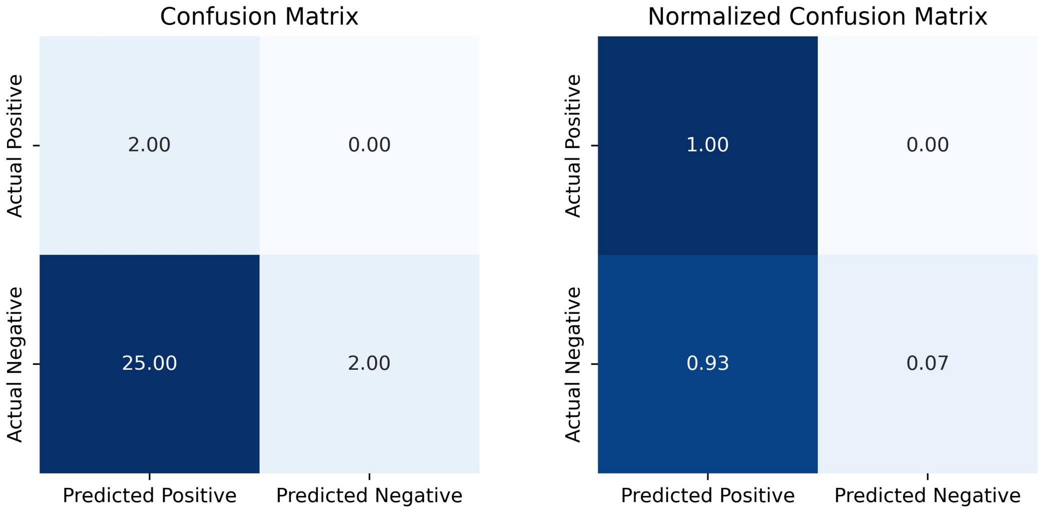
(b) Normalized. Showing percentage in decimal format

Figure 5.20: Confusion matrix of results from diagnosis model with $t_p = 0.999$, $t_r = 0.04$ and MT_{rate} method on dataset $D_{NewTest}$.

Metric	Specificity	Precision	Recall	Accuracy	F1-Score
Value	0.3889	0.6207	0.8182	0.6250	0.7059

Table 5.9: Performance metrics from diagnosis model with $t_p = 0.999$, $t_r = 0.04$ and MT_{rate} method on dataset $D_{NewTest}$

The prognosis model's confusion matrices are depicted in Figure 5.21, with corresponding performance metrics in Table 5.10. The prognosis model achieved a high recall (1) but low specificity (0.0741), indicating a high false-positive rate.



(a) Displaying the true value on the y-axis and the number of predicted WSIs on the x-axis.

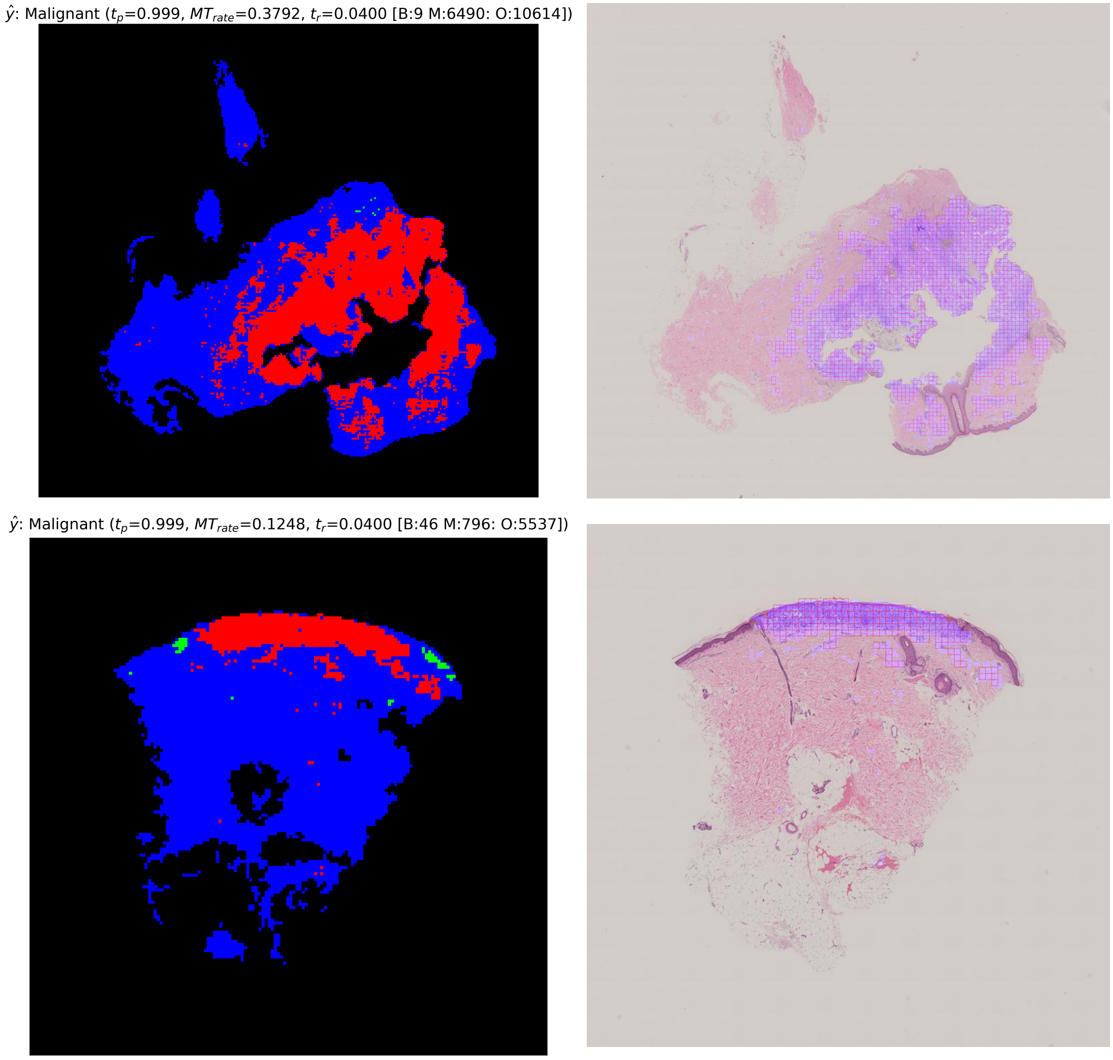
(b) Normalized. Showing percentage in decimal format

Figure 5.21: Confusion matrix of results from prognosis model using patches predicted malignant from diagnosis model. Settings used was $t_p = 0.999$ and $t_r = 0.04$ on dataset $D_{NewTest}$.

Metric	Specificity	Precision	Recall	Accuracy	F1-Score
Value	0.0741	0.0741	1.0000	0.1379	0.1379

Table 5.10: Performance metrics from prognosis model on the D_{New} running through the whole pipeline, using $t_p = 0.999$, $t_r = 0.04$ and MT_{rate} method.

Figure 5.22 displays two examples of output prediction masks from the diagnosis model integrated with the input mask for the prognosis model from the experiment.


$\hat{y}$: Malignant (t_p=0.999, MT_{rate}=0.3792, t_r=0.0400 [B:9 M:6490: O:10614])
$\hat{y}$: Malignant (t_p=0.999, MT_{rate}=0.1248, t_r=0.0400 [B:46 M:796: O:5537])

(a) Prediction masks and patient-level predicting from the diagnosis model

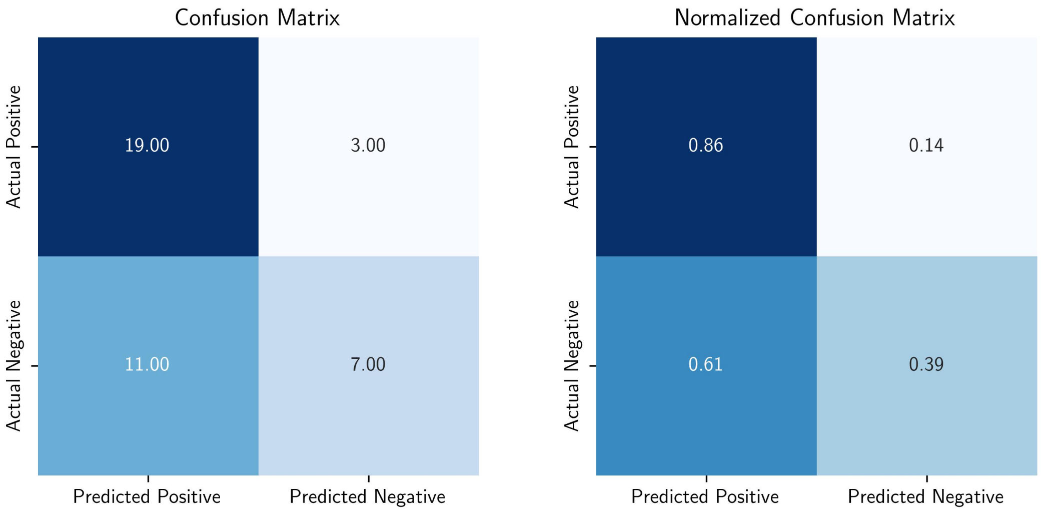
(b) Patch extraction for prognosis model, using diagnosis prediction as input.

Figure 5.22: Integration between the diagnosis output prediction mask and the prognosis input mask. Settings used was MT_{rate} method, $t_p = 0.999$ and $t_r = 0.04$ on dataset $D_{NewTest}$.

5.4.3 Testing the Pipeline using New Thresholds ($t_p = 0.9999$, $t_r = 0.01$, MT_{rate})

The next experiment evaluates the pipeline using the MT_{rate} method with a patch-level threshold (t_p) of 0.9999 and a patient-level threshold (t_r) of 0.01.

Figure 5.23 presents the confusion matrices for the diagnosis model in both absolute and normalized forms. Performance metrics are summarized in Table 5.11. The diagnosis model achieved a recall of 0.8636 and specificity of 0.3889, indicating an improvement in FPR compared to previous settings.



(a) Displaying the true value on the y-axis and the number of predicted WSIs on the x-axis.

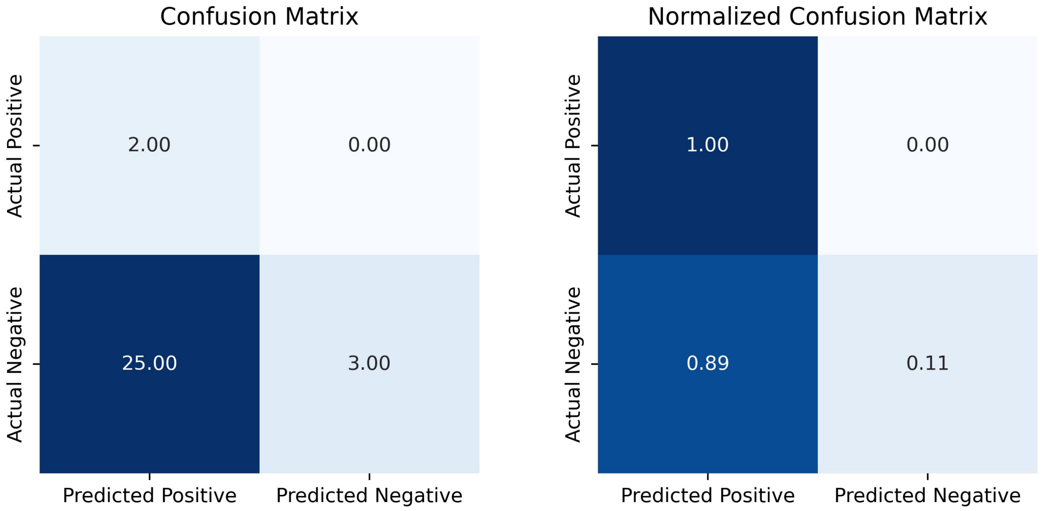
(b) Normalized. Showing percentage in decimal format.

Figure 5.23: Confusion matrix of results from diagnosis model with $t_p = 0.9999$, $t_r = 0.01$ and MT_{rate} method on dataset $D_{NewTest}$.

Metric	Specificity	Precision	Recall	Accuracy	F1-Score
Value	0.3889	0.6333	0.8636	0.6500	0.7308

Table 5.11: Performance metrics from diagnosis model on the D_{New} with $t_p = 0.9999$, $t_r = 0.01$ and MT_{rate}.

The prognosis model's confusion matrices are depicted in Figure 5.24, with corresponding performance metrics in Table 5.12. The prognosis model achieved a high recall (1.) but low specificity (0.0741), indicating a high false-positive rate.



(a) Displaying the true value on the y-axis and the number of predicted WSIs on the x-axis.

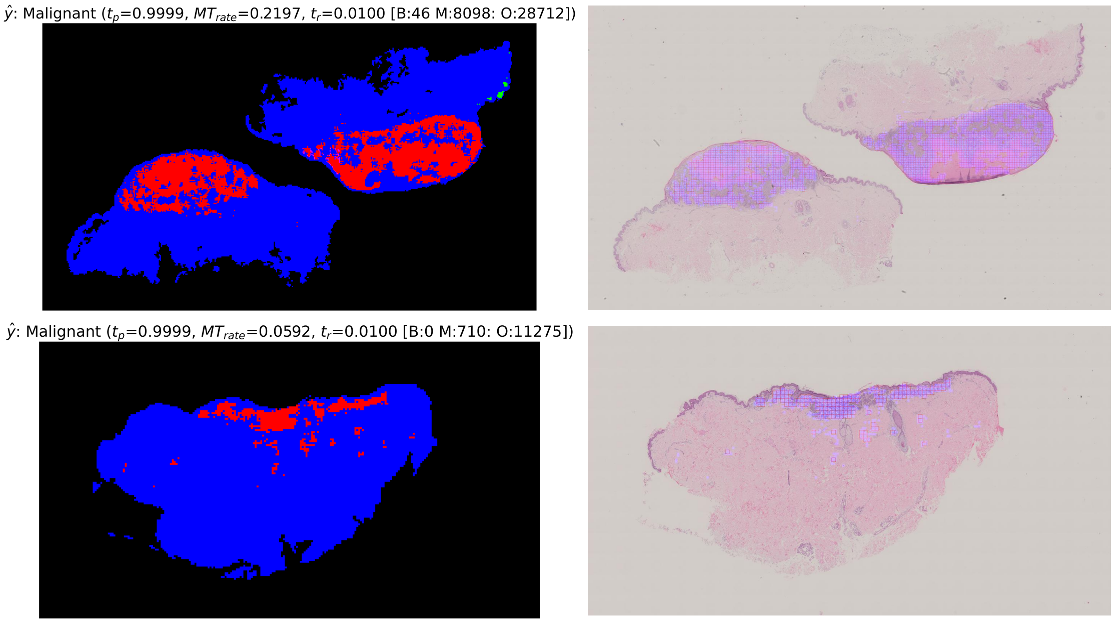
(b) Normalized. Showing percentage in decimal format.

Figure 5.24: Confusion matrix of results from prognosis model using patches predicted malignant from the diagnosis model. Settings used was $t_p = 0.9999$ and $t_r = 0.01$ on dataset $D_{NewTest}$.

Metric	Specificity	Precision	Recall	Accuracy	F1-Score
Value	0.1071	0.0741	1.0000	0.1667	0.1379

Table 5.12: Performance metrics from prognosis model on the D_{New}

Figure 5.25 displays two examples of output prediction masks from the diagnosis model integrated with the input mask for the prognosis model from the experiment.


$\hat{y}$: Malignant (t_p=0.9999, MT_{rate}=0.2197, t_r=0.0100 [B:46 M:8098: O:28712])
$\hat{y}$: Malignant (t_p=0.9999, MT_{rate}=0.0592, t_r=0.0100 [B:0 M:710: O:11275])

(a) Prediction masks and patient-level predicting from the diagnosis model

(b) Patch extraction for prognosis model, using diagnosis prediction as input.

Figure 5.25: Integration between the diagnosis output prediction mask and the prognosis input mask. Settings used was MT_{rate} method, $t_p = 0.9999$ and $t_r = 0.01$ on dataset $D_{NewTest}$.

Chapter 6

Discussion

6.1 Diagnosis Model Limitations

The diagnosis model, developed by R. Amundsen in his master's thesis, was initially trained and validated on a dataset of 90 annotated images. Of these, 73 Whole Slide Images (WSIs) were used for training, eight for validation and threshold determination, and the remaining nine for testing the model. The model achieved an accuracy of 100% when tested on this original dataset, a result that was successfully replicated in this experiment. When the same model and settings were applied to a new set of images, the performance significantly declined, as shown in Table 5.4. The precision score was 0.5989, and the specificity was only 0.1744, although the recall remained high at 0.9550. This discrepancy in performance between the original and new datasets could be attributed to several factors:

- Dataset Characteristics: The new dataset might have different characteristics compared to the WSIs used when training the model. The WSIs provided in the old dataset were selected and carefully annotated by a pathologist, which may result in the training data only containing images with certain types of characters. This can result in the new data containing a variety of different image characters, which the model has not been trained on.
- Color variation: A visual inspection revealed that the WSIs in the new dataset had more color variation from the *H&E* stains than the WSIs in the old dataset. The variation in color can be affected by different staining contrast, variations in image acquisition, and slice thickness. To counteract the color variation, technics such as color transfer could be used. This involves applying the color characteristics from the training set to the new test set.

- Overfitting: The model may have overfitted to the original training data, making it perform exceptionally well on that specific dataset but poorly when introduced to new data.

When examining the patch prediction distribution for each WSIs in the $D_{NewTest}$, a correlation between the number of patches and the number of malignant predictions was revealed. This correlation is shown in 5.10. To counteract this correlation, a new patient-based calculation method was proposed (MT_{rate}). The new method calculates the rate between the number of patches predicted malignant and the number of all valid patches in a given WSI. The equation for the MT_{rate} is presented in Equation 4.8.

ROC plots for MT_{rate} with the patch-level thresholds (t_p) of 0.999 and 0.9999 are displayed in 5.12 and 5.15, respectively. The ROC AUC score increased for MT_{rate} for both $t_p = 0.999$ (0.6106 to 0.6820) and $t_p = 0.9999$ (0.6173 to 0.7107) compared to the old malignant-rate calculation (MB_{rate}). Based on the findings, it was concluded that using MT_{rate} was preferred in this context.

6.2 Choosing parameters

Several experiments were performed to find the optimal parameters for the diagnosis model using a larger dataset. The dataset was divided into a validation set to find thresholds (D_{NewVal}) and a testing set ($D_{NewTest}$) to evaluate the new thresholds. From the visual inspection, it was decided that patch-level thresholds of 0.999 and 0.9999 gave a promising prediction mask. Some sample images from the visual inspection are displayed in 5.7. ROC plots for both patch-level thresholds presented a range of promising patient-level thresholds. It was expressed by pathologists from Stavanger University Hospital that a high sensitivity was preferred. Considering the pathologists' preferences, the model's ability to accurately predict positive melanoma cases was prioritized. The relatively low threshold of 0.04 and 0.01 was found to give good sensitivity while still maintaining some specificity. The new parameter pairs chosen for the experiments were [$t_p = 0.999$, $t_r 0.01$, MT_{rate}] and [$t_p = 0.9909$, $t_r 0.04$, MT_{rate}].

Since the prognosis model provides patient-level prediction in a binary fashion, the ratio between metastasis and valid patches has to be fixed. For our experiments, we fixed this threshold to 0.3720. However, this threshold can be further tuned based on a prior model that is integrated into the pipeline. It is complex to find the optimal threshold by selecting a possible combination of thresholds from diagnostic and prognostic models due to the computational cost affiliated with inference of the new dataset. In short, we need to find the best thresholds from diagnostic models first and then optimize the threshold for the prognostic model.

6.3 Prognosis Model

The prognosis model was trained solely on malignant patches, leading to a lack of representation of other tissue types in the training data. This becomes a significant issue when utilizing the prediction masks from the diagnostic model. Specifically, regions inaccurately identified as melanoma by the diagnostic model will be processed by the prognosis model. This could potentially skew the prognosis model's predictions, as it has not been trained to recognize and differentiate non-melanoma tissues. Consequently, the model might provide inaccurate prognoses based on these false positives, which could lead to unnecessary treatments or interventions.

6.4 Integration of Both Models

The table 6.1 presents the performance metrics from three different runs of the integrated pipeline, each employing different thresholds for the diagnosis on the $D_{NewTest}$.

In the first run (A), the diagnosis model (D) achieved perfect recall (1.0000) and a high F1-Score (0.7333), indicating that it successfully identified all actual positive melanoma cases. However, its specificity was low (0.1111), suggesting a high false positive rate. The prognosis model (P) in this run also achieved perfect recall but had a low F1-Score (0.1143), indicating a high number of false positives.

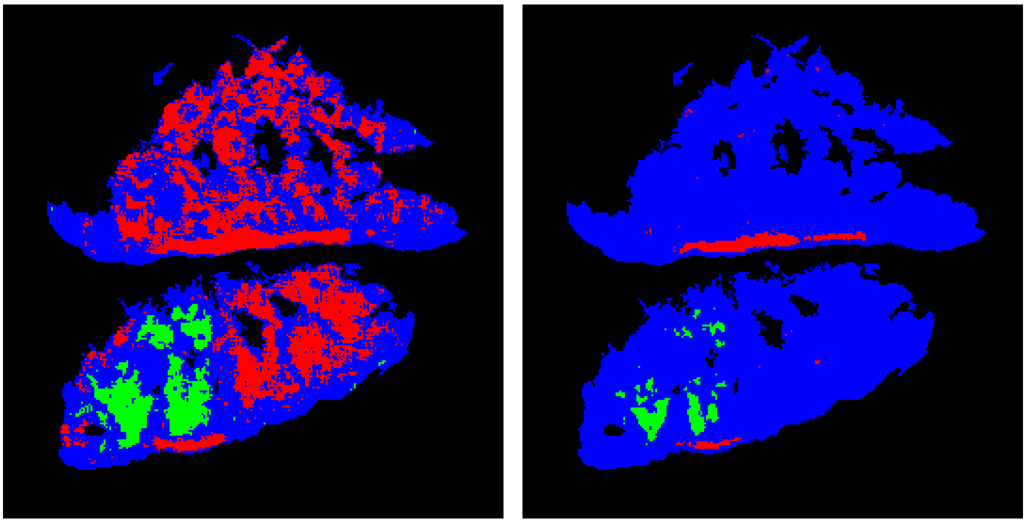
In the second run (B), the diagnosis model improved its specificity to 0.3889 and maintained a high F1-Score (0.7059). The prognosis model, however, did not show significant improvement in its performance.

In the third run (C), the diagnosis model further improved its specificity to 0.3889 and precision to 0.6333, while maintaining a high recall (0.8636) and accuracy (0.6500). The prognosis model's performance remained relatively unchanged.

Overall, the diagnosis model showed consistent improvement across the runs, particularly in terms of specificity and precision, without significant loss in recall or accuracy. This suggests that adjusting the thresholds improved the model's ability to correctly identify non-melanoma cases and reduce false positives. Run C gave the best performance with respect to the tradeoff between specificity and recall. However, the best settings for the model are highly dependent on the use case. Using a higher patch-level threshold might result in improved accuracy, but the malignant tissue mask will be reduced. Contrary, choosing a lower patch-level threshold might result in other tissue areas being present in the melanoma prediction mask. Figure 6.1 presents a prediction from the diagnosis model with different patch-level thresholds.

Run (Ch)	**Model**	$\mathbf{t_p}$	$\mathbf{t_r}$	$\mathbf{M_{rate}}$	**Specificity**	**Precision**	**Recall**	**Accuracy**	$\mathbf{F_1}$**-score**
A. Ch. 5.2	D	0.9990	0.04	MB	0.1111	0.5789	**1.0000**	0.6000	**0.7333**
	P	-	-	-	0.1389	0.0606	1.0000	0.1842	0.1143
B. Ch. 5.4.2	D	0.9990	0.04	MT	**0.3889**	0.6207	0.8182	0.6250	0.7059
	P	-	-	-	0.0741	0.0741	1.0000	0.1379	0.1379
C. Ch. 5.4.3	D	**0.9999**	**0.01**	**MT**	**0.3889**	**0.6333**	0.8636	**0.6500**	0.7308
	P	-	-	-	0.1071	0.0741	1.0000	0.1667	0.1379

Table 6.1: Performance metrics from every experiment using different thresholds from the diagnosis and prognosis model on the $D_{NewTest}$. Rows with the same run name belong to the same run. The letter in *Model* refers to if the metrics in the row are from the diagnosis (D) or prognosis (P) model from the respective run.



(a) Prediction mask from the diagnosis model using a patch-level threshold value of 0.99

(b) Prediction mask from the diagnosis model using a patch-level threshold value of 0.9999

Figure 6.1: Representation of malignant tissue masks from different patch-level thresholds for the prognosis model.

The prognosis model's performance did not change significantly in the experiments. The recall was 1.0000 for all runs, and experiments A and C resulted in increased specificity. Run A and B uses the same t_p, which results in identical prediction masks from the diagnosis model. The only factor differentiating the two experiments is the calculation of M_{rate}. In other words, patient-level prediction created two different sets of WSIs resulting in different performances.

Chapter 7

Conclusion

The main objective of this thesis was to develop a pipeline for predicting the diagnosis and prognosis of WSIs. A dataset consisting of 243 WSIs from different patients, each with a weak label was provided. The pipeline includes two independent models. The first model is responsible for predicting and localizing melanoma in a WSI. The second model is responsible for predicting the prognosis by using information about malignant regions provided by the diagnostic model.

The initial result for the diagnostic model on the old datasets was promising but was not generalizable on the new data set with the current parameters. By tuning the parameters, the specificity increased, but recall decreased. The best result for the diagnosis model gave a recall of 0.8636 and a specificity of 0.3889. From internal conversations with pathologists from SUH, it was decided to prioritize minimizing false negatives, as this could potentially lead to underdiagnosis or undertreatment in real clinical scenarios. Consequently, it is suggested to use the diagnosis model with thresholds that give it high sensitivity. This is achieved by using low patch-level and patient-level thresholds. In a real clinical scenario, the majority of cases are benign. By using a diagnosis model with good sensitivity, it can be used as a tool to help pathologists prioritize by excluding non-urgent cases.

A correlation between the number of patches predicted malignant and the total number of patches was discovered in the new dataset provided. To counteract this correlation, a calculation method for finding the melanoma ratio in a WSI was proposed. The proposed calculation method takes into account all valid patches rather than only the predicted lesion. This has the added benefit of not depending on the models' ability to correctly predict benign patches. This method was utilized when giving a patient-level diagnostic prediction.

The pipeline integration displayed some promising results. The diagnosis model presented considerable potential in localizing the malignant lesion within the WSIs, resulting in the prognosis model being provided with relevant patches. It is important to note that as the dataset lacked annotations of the lesions, the final confirmation of the predicted lesions still relies on the expertise of a pathologist. The region presented in the prognosis model can be adjusted by carefully tuning the patch-level threshold for the diagnosis model. The patient-level threshold can be selected to include or exclude WSIs but does not impact the prediction mask. In all three experiments, the prognosis model predicted all positive cases correctly. It is important to acknowledge that the test set used only contained two malignant cases, which is not enough samples to give a conclusive evaluation. In addition, the best accuracy from the experiments was only 0.1842. The performance of the prognosis model might be increased by tuning the patient (good/bad patches) ratio threshold for the model.

7.1 Future Work

This section will present suggestions for future work that were not within the scope of this thesis.

- Optimizing parameters for the prognosis model: Optimizing parameters for the prognosis model is a crucial step for improving the pipeline. This can be achieved by generating prediction masks with the diagnostic model using promising parameters, followed by evaluating patient-level thresholds for the prognosis model. This approach could potentially enhance the accuracy of prognostic predictions by ensuring that the model is tuned to the specific characteristics of the data.

- Incorporating preprocessing techniques for artifact detection and color variation. These preprocessing techniques, as presented in [18], could significantly enhance the model's performance. These techniques can help mitigate the effects of variations in image quality and color. This could potentially improve the model's generalizability and robustness to different imaging conditions and artifacts.

- Implementing a multiple instance learning (MIL) classifier for prognostic prediction. MIL allows the model to extract feature vectors and apply them effectively even when only weak labels are available. This is especially beneficial in cases where metastasis labels do not provide localized information or annotations. By identifying and learning from these weakly labeled instances, MIL can potentially improve the model's ability to predict prognosis with greater accuracy. In essence, this approach could enhance the model's learning capability from less detailed or incomplete data, thereby improving its overall predictive performance.

Bibliography

[1] Naheed R. Abbasi et al. “Early Diagnosis of Cutaneous MelanomaRevisiting the ABCD Criteria”. In: *JAMA* 292.22 (Dec. 8, 2004), pp. 2771–2776. ISSN: 0098-7484. DOI: 10.1001/jama.292.22.2771. URL: https://doi.org/10.1001/jama.292.22.2771 (visited on 05/25/2023).

[2] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. “Understanding of a convolutional neural network”. In: *2017 International Conference on Engineering and Technology (ICET)*. 2017 International Conference on Engineering and Technology (ICET). Aug. 2017, pp. 1–6. DOI: 10.1109/ICEngTechnol.2017.8308186.

[3] Roger Amundsen. “Melanoma Diagnosis and Localization from Whole Slide Images using Convolutional Neural Networks”. Accepted: 2022-09-02T15:51:24Z. Master thesis. uis, 2022. URL: https://uis.brage.unit.no/uis-xmlui/handle/11250/3015481 (visited on 06/13/2023).

[4] Christopher Andreassen. “Melanoma prognosis prediction using image processing and machine learning”. Accepted: 2022-11-11T16:51:12Z. Master thesis. uis, 2022. URL: https://uis.brage.unit.no/uis-xmlui/handle/11250/3031489 (visited on 06/13/2023).

[5] Christopher Andreassen et al. *Deep Learning for Predicting Metastasis on Melanoma WSIs*. Mar. 10, 2023. DOI: 10.48550/arXiv.2303.05752. arXiv: 2303.05752[cs, eess]. URL: http://arxiv.org/abs/2303.05752 (visited on 06/23/2023).

[6] Kaustav Bera et al. “Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology”. In: *Nature Reviews Clinical Oncology* 16.11 (Nov. 2019). Number: 11 Publisher: Nature Publishing Group, pp. 703–715. ISSN: 1759-4782. DOI: 10.1038/s41571-019-0252-y. URL: https://www.nature.com/articles/s41571-019-0252-y (visited on 06/11/2023).

[7] Thomas M. Brown and Karthik Krishnamurthy. “Histology, Dermis”. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2023. URL: http://www.ncbi.nlm.nih.gov/books/NBK535346/ (visited on 06/29/2023).

[8] *Cancer in Norway 2021 - Cancer incidence, mortality, survival and prevalence in Norway*. 2022. URL: https://www.kreftregisteret.no/globalassets/cancer-in-norway/2021/cin_report.pdf.

[9] Janice Cormier et al. "Improving outcomes in patients with melanoma: strategies to ensure an early diagnosis". In: *Patient Related Outcome Measures* (Nov. 2015), p. 229. ISSN: 1179-271X. DOI: 10.2147/PROM.S69351. URL: https://www.dovepress.com/improving-outcomes-in-patients-with-melanoma-strategies-to-ensure-an-e-peer-reviewed-article-PROM (visited on 06/29/2023).

[10] *Definition of metastasis - NCI Dictionary of Cancer Terms - NCI*. Archive Location: nciglobal,ncienterprise. Feb. 2, 2011. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/metastasis (visited on 05/22/2023).

[11] R. J. Friedman, D. S. Rigel, and A. W. Kopf. "Early Detection of Malignant Melanoma: The Role of Physician Examination and Self-Examination of the Skin". In: *CA: A Cancer Journal for Clinicians* 35.3 (May 1, 1985), pp. 130–151. ISSN: 0007-9235. DOI: 10.3322/canjclin.35.3.130. URL: http://doi.wiley.com/10.3322/canjclin.35.3.130 (visited on 05/25/2023).

[12] Lidiya Georgieva, Tatyana Dimitrova, and Nicola Angelov. "RGB and HSV colour models in colour identification of digital traumas images". In: (2005).

[13] Megha Goyal. "Morphological Image Processing". In: 2.4 (2011).

[14] Jonathan B. Heistein, Utkarsh Acharya, and Shiva Kumar R. Mukkamalla. "Malignant Melanoma". In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2023. URL: http://www.ncbi.nlm.nih.gov/books/NBK470409/ (visited on 05/20/2023).

[15] *How biopsy procedures are used to diagnose cancer*. Mayo Clinic. URL: https://www.mayoclinic.org/diseases-conditions/cancer/in-depth/biopsy/art-20043922 (visited on 06/29/2023).

[16] Jean Kanitakis. "Anatomy, histology and immunohistochemistry of normal human skin". In: *European journal of dermatology : EJD* 12 (Nov. 2001), 390–9, quiz 400.

[17] Neel Kanwal et al. "Detection and localization of melanoma skin cancer in histopathological whole slide images". In: *arXiv preprint arXiv:2302.03014* (2023).

[18] Neel Kanwal et al. "The Devil is in the Details: Whole Slide Image Acquisition and Processing for Artifacts Detection, Color Variation, and Data Augmentation: A Review". In: *IEEE Access* 10 (2022). Publisher: IEEE, pp. 58821–58844.

[19] *Learn About Melanoma*. IMPACT Melanoma. URL: https://impactmelanoma.org/learn-about-melanoma/ (visited on 06/29/2023).

[20] *Melanoma cancer statistic in Norway 1930-2022 (men, women)*. URL: https://sb.kreftregisteret.no/ (visited on 06/11/2023).

[21] *Metastatic Cancer: When Cancer Spreads - NCI*. Archive Location: nciglobal,ncienterprise. May 12, 2015. URL: https://www.cancer.gov/types/metastatic-cancer (visited on 05/22/2023).

[22] Arlo J. Miller and Martin C. Mihm. “Melanoma”. In: *New England Journal of Medicine* 355.1 (July 6, 2006), pp. 51–65. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMra052166. URL: http://www.nejm.org/doi/abs/10.1056/NEJMra052166 (visited on 05/07/2023).

[23] Sandra Morales, Kjersti Engan, and Valery Naranjo. “Artificial intelligence in computational pathology – challenges and future directions”. In: *Digital Signal Processing* 119 (Dec. 1, 2021), p. 103196. ISSN: 1051-2004. DOI: 10.1016/j.dsp.2021.103196. URL: https://www.sciencedirect.com/science/article/pii/S1051200421002359 (visited on 06/26/2023).

[24] Maryam M. Najafabadi et al. “Deep learning applications and challenges in big data analytics”. In: *Journal of Big Data* 2.1 (Feb. 24, 2015), p. 1. ISSN: 2196-1115. DOI: 10.1186/s40537-014-0007-7. URL: https://doi.org/10.1186/s40537-014-0007-7 (visited on 06/27/2023).

[25] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. “Digital pathology and artificial intelligence”. In: *The Lancet Oncology* 20.5 (May 1, 2019), e253–e261. ISSN: 1470-2045. DOI: 10.1016/S1470-2045(19)30154-8. URL: https://www.sciencedirect.com/science/article/pii/S1470204519301548 (visited on 06/26/2023).

[26] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[27] Robin Reid, Fiona Roberts, and Elaine MacDuff. *Pathology Illustrated E-Book*. Google-Books-ID: OWnHAAAAQBAJ. Elsevier Health Sciences, Oct. 24, 2011. 683 pp. ISBN: 978-0-7020-4829-6.

[28] *Skin cancer statistics | World Cancer Research Fund International*. WCRF International. URL: https://www.wcrf.org/cancer-trends/skin-cancer-statistics/ (visited on 06/29/2023).

[29] Srinath Sundararajan et al. “Metastatic Melanoma”. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2023. URL: http://www.ncbi.nlm.nih.gov/books/NBK470358/ (visited on 06/27/2023).

[30] *Timing of Melanoma Diagnosis, Treatment Critical to Survival*. Consult QD. Section: Cancer. Oct. 30, 2017. URL: https://consultqd.clevelandclinic.org/timing-of-melanoma-diagnosis-treatment-critical-to-survival/ (visited on 06/29/2023).

[31] wenxing su wenxing et al. “Bioinformatic analysis reveals hub genes and pathways that promote melanoma metastasis”. In: 20.1 (Sept. 2020). DOI: https://doi.org/10.1186/s12885-020-07372-5. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7487637/.

[32] Rune Wetteland. “Parameterized Extraction of Tiles in Multilevel Gigapixel Images”. In: *Intritute of Electrical and Electronics Engineers* (2021). In collab. with Kjersti Engan and Trygve Eftesol. URL: https://ieeexplore.ieee.org/document/9552104/authors#authors.

[33] *What is melanoma?* URL: https://www.cancerresearchuk.org/about-cancer/melanoma/about (visited on 06/29/2023).

[34] World Health Organication. *Skin cancer – WHO*. URL: https://www.iarc.who.int/cancer-type/skin-cancer/ (visited on 06/29/2023).

[35] Hani Yousef, Mandy Alhajj, and Sandeep Sharma. “Anatomy, Skin (Integument), Epidermis”. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2023. URL: http://www.ncbi.nlm.nih.gov/books/NBK470464/ (visited on 05/08/2023).

[36] Xiangyu Zhang et al. *Accelerating Very Deep Convolutional Networks for Classification and Detection*. Nov. 18, 2015. DOI: 10.48550/arXiv.1505.06798. arXiv: 1505.06798[cs]. URL: http://arxiv.org/abs/1505.06798 (visited on 06/27/2023).

Appendix

Link to GitHub repository: https://github.com/Mariebs97/Master23.git