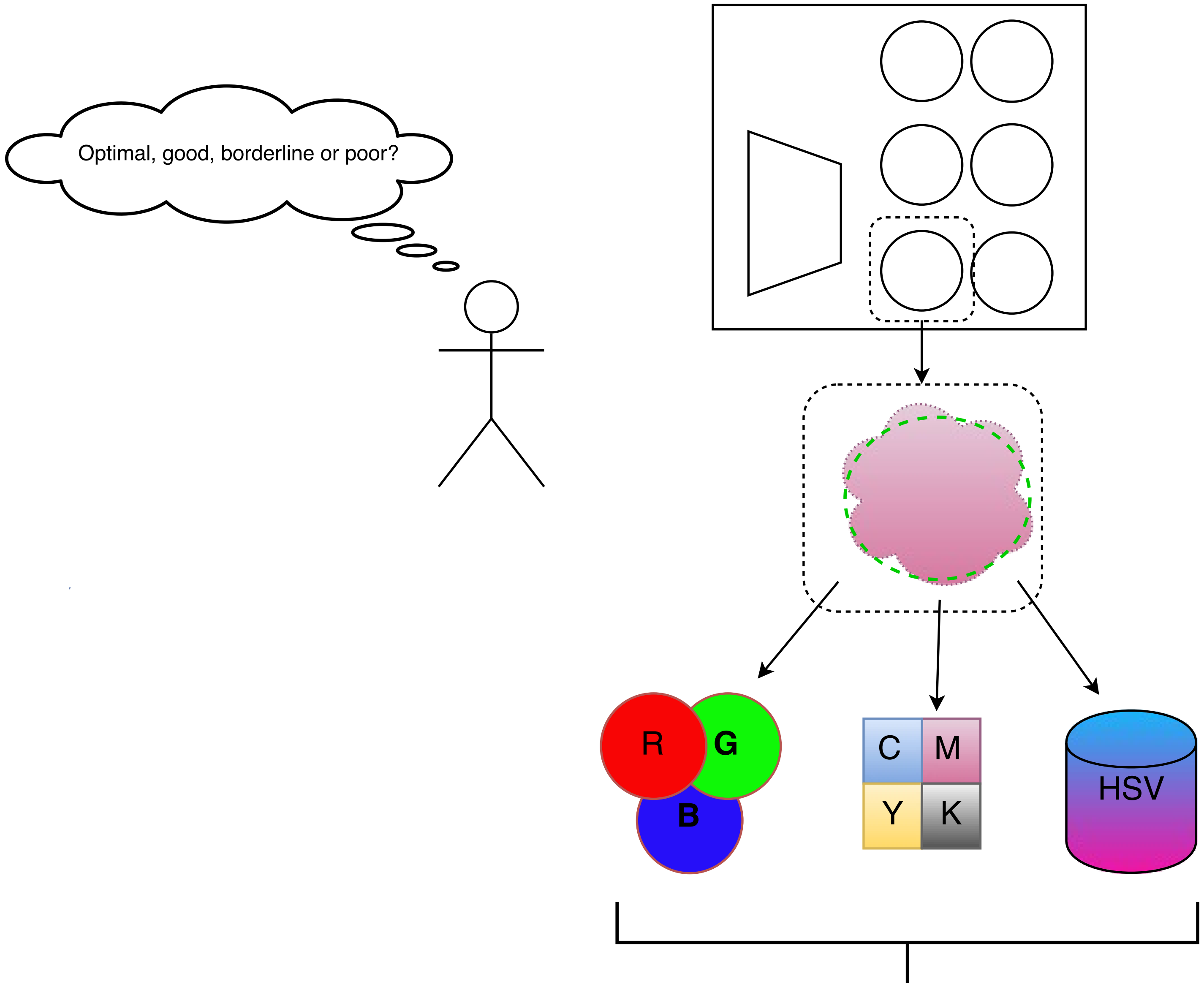
Segmentation and classification of immunohistochemically stained samples based on NordiQC Pan-CK assessment



	Renal clear cell carcinoma
Accuracy	85.74%
Sensitivity	85.92%
Specificity	85.57%

Master Thesis

Christos Zoupis-Schoinas

Søren Larsen

Biomedical Engineering
and Informatics

June 2017

Summary

Introduction

Immunohistochemical staining is one of the tools used by the pathologist to examine tissue samples. This method is used to determine the presence cancerous tissue and its type by applying specific stains. Different cellular components have different responses and this will highlight differences. The pathologist can then base their diagnosis on the observations made on these responses. The immunohistochemical staining itself is a process that involves several steps which can lead to variations. Several studies have covered the topics of inter-laboratory and inter/intra-observer variability, but intra-laboratory seems uncovered. The inter-laboratory variability seems to be commonly acknowledged by most authors. Their primary suggest is that standardisation of protocols should be attempted. However in order to standardise, one should have points of reference as to what a good staining is. Also, at the observer's level the variability is present, both the inter- and intra-. This was found to be a minimum for the inter-observer variability. However for the intra-observer variability studies it was found that the subjective interpretation of the observer could vary on a daily basis, and since it is done manually that it can depend on the alertness and experience of the observer. One of the initiatives that acts as a quality assessment for laboratories worldwide is NordiQC based in Aalborg, Denmark. The laboratories' samples are graded in consensus by 5 clinicians into optimal, good, borderline or poor based on the staining quality. This grading is the qualitative interpretation and is thus still subject to variability. Studies examining the use of Computer-Aided-Decision systems have shown that these can be a future tool within pathology. Therefore this master thesis seeks to investigate the plausibility of grading the samples using quantitative measures by the following study aims:

- To develop an algorithm capable of analysing the NordiQC samples separately
- To investigate the possibility of using image features and classification to describe NordiQC grades
- To validate the results compared to the NordiQC grades
- To explore the plausibility of a CAD system for the NordiQC collaboration

Method

MATLAB® 2016b was used to develop a set of algorithms. Initially the oesophagus was removed from the sample. This was followed by the detection of each tissue core for all samples. This would specify the region of interest for the subsequent snake. A dictionary snake model was applied to segment all the cores and remove the background. From the segmented cores the features were extracted and the Quadratic Discriminant Analysis was applied. Using the Sequential Forward Selection and the Backward Elimination, the feature selection was done for each core. A 5-fold classification was then applied to validate the results comparing them to the NordiQC assessments.

Results

All cores were successfully detected and the deformable model segmented all cores with acceptable results. Excluded features varied between cores suggesting different features describe each core. The maximum and minimum accuracy, sensitivity and specificity were calculated for all cores. From 10 repetitions the average accuracy when validating each core was between 60-85%. Core 2 gave the best results, while core 6 had the lowest accuracy. Tables of misclassifications show also that the classified sample for core 2 rarely shifted more than one grade compared to its assessed grade, while core 6 had a larger frequency of misclassified grades. There was little difference between the average accuracy and the grade specific accuracy for each core.

Discussion

From the results of the validation it is shown that the used classifier can indeed differentiate the samples into their NordiQC grades. This was done by showing an average accuracy ranging from 60-85%. This performance was achieved using only 38 samples and the initial 91 features, primarily intensity-based supplemented by staining colour specific features. This could possible be improved by adding more samples or introducing other feature spaces. The detection and segmentation made it possible to analyse each core in each sample separately. This segmentation process could be refined by introducing core specific deformable models. This thesis shows that CAD can be used to classify staining quality in the case of NordiQC. However to increase the overall performance, further study should be conducted.

Resume

Introduktion

Immunohistokemisk farvning er blot en af de redskaber en patolog kan gøre brug af når vævsprøver skal undersøges. Denne metode bruges til at vurdere tilstedeværelsen og type af cancer gennem anvendelsen af specifikke farvestoffer. Forskellige cellekomponenter reagerer forskelligt på selve farvningen og dette kan tydeliggøre forskelle i vævsprøven. Patologen kan derefter udforme en diagnose ud fra observationer gjort i selve vævsprøven. Selve den immunohistokemiske farvning er en proces der indeholder mange forskellige trin, som kan lede til variationer. Flere studier har undersøgt emnerne omkring inter-laboratorie samt inter-/intra-observatør variabilitet, selvom intra-laboratorie variabilitet synes uberørt. Selve inter-laboratorie variabilitet lader til at være anerkendt af de fleste forfattere. Deres primære forslag til dette er at standardisering af protokoller bør forsøges. Men før man kan standardisere disse, så bør man have referencepunkter til hvad en god farvning reelt er. Også på observatørniveau optræder variabilitet, både som inter- og som intra-. Selve inter-observatør variabilitet blev fundet til at være et minimum. Men for intra-observatør variabilitet fandt flere studier at observatørens subjektive fortolkning kunne variere fra dag til dag, og da dette udføres manuelt så afhænger dette af observatørens årvågenhed og erfaring. En af de initiativerne som fungere som kvalitetsvurdering for laboratorier på verdensplan er NordiQC med beliggenhed i Aalborg. De enkelte laboratoriers prøver bliver vurderet af et panel bestående af 5 klinikere som enten optimal, good, borderline el. poor ud fra kvaliteten af selve farvningen. Denne vurdering er en kvalitativ fortolkning og kan derved stadig påvirkes af variabilitet. Undersøgelser har vist at Computer-Aided-Decision kan blive et fremtidigt redskab for patologer. Derfor søger dette speciale at undersøge muligheden for vurderingen af disse vævsprøver ved hjælp af kvantitative enheder. Dette gøres gennem følgende projektmål:

- Udviklingen af en algoritme der muliggør separat analyse af NordiQC vævsprøver
- Undersøge muligheden for at beskrive NordiQC vurderinger gennem brugen af billedkarakteristika og klassifikationer
- Validering af resultaterne op mod vurderinger foretaget af NordiQC
- Udforske muligheden for et CAD system målrettet NordiQC

Metode

MATLAB$^{®}$ 2016b blev anvendt til at udvikle en serie af algoritmer. Indledningsvist blev øsofagus fjernet fra vævsbilledet. Dette blev efterfulgt af identifikationen af hver vævskerne i alle vævsbilleder. Dette ville specificere området til den efterfølgende deformerbare model. En ’dictionary snake’ blev anvendt til at segmentere alle vævskerne og fjerne billedbaggrunden. Fra disse segmenterede vævskerne blev der udtrukket ’features’ og ’Quadratic Discriminant Analysis’ blev anvendt. Gennem ’Sequential Forward Selection’ og ’Backward Elimination’ blev ’features’ valgt for hver eneste vævskerne. En 5-fold klassifikation blev slutteligt anvendt for at validere resultaterne og sammenligne dem med NordiQC’s vurdering.

Resultater

Alle vævskerne blev identificeres med succes og den deformerbare model segmenterede alle vævskernes med acceptable resultater. De udeladte ’features’ varierede mellem vævskerne hvilket antyder at hver vævskerne beskrives af forskellige ’features’. Den maksimale og minimale nøjagtighed, sensitivitet og specificitet blev beregnet for alle vævskerne. Efter 10 repetitioner var den gennemsnitlige nøjagtighed under valideringen af hver vævskerne mellem 60-85%. Vævskerne 2 havde det bedste

resultat og kerne 6 havde det laveste. Tabeller over fejlklassifikationer viser at de klassificerede vævsprøver for kerne 2 sjældent skifter mere end en grad når denne sammenlignes med NordiQCs vurdering, hvorimod vævskerne 6 havde en større frekvens ang. fejlklassifikation. Der var minimal forskel mellem den gennemsnitlige nøjagtighed og den vurderingsspecifikke nøjagtighed for alle vævskerner.

Diskussion

Fra resultaterne af valideringen kan det ses at den anvendte 'classifier' er i stand til at differentiere vævsprøverne op mod deres NordiQC vurderinger. Dette viste en gennemsnitlig nøjagtighed mellem 60-85%. Dette blev præsteret med kun 38 vævsprøver og de indledende 91 'features'. Dette kunne sandsynligvis forbedres hvis flere prøver blev tilføjet datasættet eller gennem anvendelse af flere 'features'. Identifikationen og segmenteringen gjorde det muligt at analysere hver vævskerne individuelt. Selve segmenteringen kunne bedres ved at introducere deformerbare modeller målrettet den enkelte vævskerne. Dette projekt viser at CAD kan anvendes til at klassificere kvaliteten af farvningen i vævsprøver fra NordiQC. Det anbefales dog at dette undersøges yderligere således klassifikationens præstation kan forbedres.

AALBORG UNIVERSITY
DENMARK

School of Medicine and Health
Fredrik Bajers Vej 7
Phone 96 35 97 31
Fax 98 13 63 93
http://smh.aau.dk

Title: Segmentation and classification of immunohistochemically stained samples based on NordiQC Pan-CK assessment

Keywords: SomeNiceKeyWords

Field of study: Biomedical Engineering and Informatics

Project period ST10, Spring semester 2017

Project group: 17gr10409

Participants:

Christos Zoupis-Schoinas

Søren Larsen

Supervisor: Lasse Riis Østergaard

Supervisor: Alex Skovsbo Jørgensen

Submitted: June 6th, 2017

Preface

This master thesis is the product of the fourth semester during the masters program in Biomedical Engineering and Informatics at Aalborg University. The project period spanned from February 6th 2017 to June 7th 2017.

The initial project proposal was to investigate the possibility of extracting image features from 39 immunohistochemically stained samples in order to classify the quality a given staining protocol according to NordiQC assessment.

The project touches the topics of image analysis and pattern recognition, and a pre-existing knowledge within these fields may prove beneficial to the reader. The thesis itself is primarily aimed towards the investigation of the project proposal. However those with interest within the topics of quality control of IHC staining and computer-aided-decision may find it to their liking as well.

To reference the literature the Harvard method was used in this project. Citations are applied in brackets with author and the year when published. If no date is available the notation n.d. is used instead of the date. If the citation is made after a punctuation mark, it is concerns the text until the start of the paragraph, or until another citation. If the citation is only used for one specific sentence it is written before the punctuation mark. If more than one citation is made for the same statement, citations are comma separated. Figures and tables are also included with citations. All the sources for citations are gathered in a literature list with further details.

Acknowledgements goes to:
Lasse Riis Østergaard and Alex Skovsbo Jørgensen who both have been deeply involved in this thesis and showed the importance of this project even when it felt trivial. Their valuable sparring and continuous support have been greatly appreciated.
Rasmus Røge and the NordiQC for the possibility of contributing to the quality of IHC staining, and for providing the data used in this thesis.

Table of contents

Chapter 1

Introduction

Immunohistochemical (IHC) staining is a technique which utilises antibodies that bind to specific antigens. The IHC staining is used within many fields but its use is highly adopted within the medical field of pathology. The pathologist can use the IHC technology to determine if specific antigens are present within a tissue sample. By using specific cancer-targeting antibodies the pathologist is able to assess whether or not the tissue could contain a type of cancer. Given this technology the pathologist is better able to detect the presence of cancer even if it is not visible under the microscope. This will allow earlier detection and diagnosing of cancer thus enabling better survival chance given earlier treatment. [Sino Biological, Inc., n.d.b]

The quality of the staining is crucial as the future cancer therapy will be greatly influenced by the result of this. If the staining is either too strong or too weak it might result in a false negative or a false positive from the interpretation by the pathologist. If it results in a false negative there would indeed be cancer cells within the patient's tissue sample and most likely within the patient, but the patient would not be diagnosed with cancer which would delay any potentially lifesaving treatment(s). If a false positive would occur the tissue sample would not contain any cancer cell but the patient would be diagnosed with cancer. The patient would then be exposed to the concern of being diagnosed with cancer and unnecessary side effects from the treatment(s). Neither of these are recommendable and should be avoided if possible which adds importance to the quality of the staining in diagnostic purposes. [Reiner-Concin, 2008]

The use of IHC depends not only on the pathologist but also that the staining is done properly. The staining process itself is influenced by the technician, machinery, stains, and protocols. There exists multiple manufacturers each with an assortment of machinery that is used in the staining process. The staining agents are also produced by multiple companies or within the laboratory itself. The protocol by which the staining is performed can also vary between laboratories. Both the machinery, the stains, and the protocol contribute to a source of staining variability. As described by Taylor [2000] the technicians/lab might utilise a qualitative approach to staining quality in order to comply with the pathologist's request [Taylor, 2000]. What is sufficient can vary from pathologist to pathologist, and this approach does not lend itself to any high degree of standardisation thus increasing variability.

Within the given laboratory the Quality Control (QC) should ensure that procedures are done according to protocol ie. monitoring and detection, and similarly the Quality Assurance/Assessment (QA) focus on ensuring the quality of the result [Arthur, n.d.]. In the staining IHC context this relates to: the correct patient, the correct staining/protocol result, and the correct interpretation. Maxwell and McCluggage [2000] also mentions that even through the use of positive and negative control, few recommendations exists for the internal QC to determine the quality of IHC staining [Maxwell and McCluggage, 2000]. One option to overcome this is through the use of external QA, either the UK National External Quality Assurance (UK-NEQAS) or the Nordic immunohistochemical Quality Control (NordiQC) [Maxwell and McCluggage, 2000, Eisen, 2008].

The Nordic immunohistochemical Quality Control was established in 2003 to act as an external QA programme. The NordiQC is a collaboration between the founding countries of Denmark, Norway, Sweden and Finland. As of 2015 the NordiQC serves as a QA for more than 700 laboratories from about 80 countries. Laboratories receive tissue samples, organised into Tissue Micro Arrays (TMA), which they are to stain according to their own protocol and are then graded by NordiQC. An example of such an TMA is Pan-Cytokeratin (Pan-CK). The Pan-CK is important in IHC staining when the pathologist wants to detect and classify carcinomas which is the most common group of cancers, and it will also allow for the differentiation of carcinomas without many morphological characteristics. [NordiQC, 2016, Vyberg et al., 2005, Vyberg and Nielsen, 2016, Jørgensen et al., 2017]

However there are no quantitative measures or definitions for these four different gradings beyond what is mentioned in the assessment.

Chapter 2

Problem Analysis

2.1 The immunohistochemical staining

The process of immunohistochemical staining is based on the sensitive interaction between immunoglobulin (Ig), more commonly known as antibody, with an antigen. The antibody's binding to the antigen occurs at a specific site called the epitope and may be characterised by a specific sequence of amino acids or spatial structure. [Oliver and Jamur, 2009]

The B-lymphocytes are responsible for the production of antibodies found in humans, and can be grouped into five major Ig-classes consisting of IgA, IgD, IgE, IgG and IgM. An illustration of these can be seen in figure 2.1. Each B lymphocyte produces only one of the five Ig antibodies, thereby targeting a single epitope. The IgG is the most used in IHC staining with its prevalence of 75 % in human serum. The IgG can be further divided into four different subclasses in humans, while five subclasses can be found in mice. These subclasses are IgG$_1$, IgG$_{2a}$, IgG$_{2b}$, IgG$_3$ and IgG$_4$. [Oliver and Jamur, 2009]

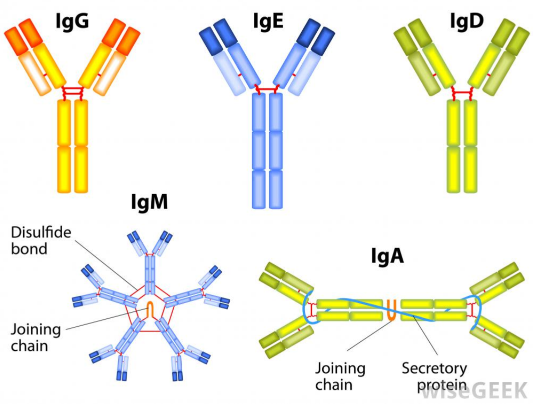


Figure 2.1: An illustration of the types of Ig antibodies. [Serotec, n.d.]

2.1.1 The IHC process

The IHC involves a series of steps based on the protocol. Oliver and Jamur [2009] describes a method when performing IHC staining which is summed up into four main steps; Common fixation, Preparation of tissue blocks, Preparation of slides, Preparation of sections for immunostaining [Oliver and Jamur, 2009]. These steps can be seen in figure 2.2. Note that these steps only focus on the staining procedure and that prior to these steps the tissue in focus would be or is extracted from the patient. Likewise the pathologist, or technician, would inspect and document the findings within the sample leading to the diagnosis.

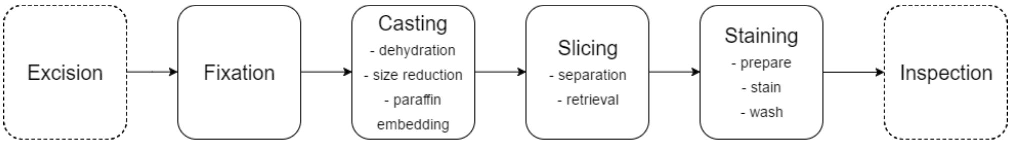


Figure 2.2: An illustration of the generic IHC staining process from start to finish

After the tissue is extracted, it is then placed in a fixative that should equal to 10x the volume of the tissue sample. This is done in order to fixate the cells within the tissue thus halting any decay. If this is not performed the antigen epitopes can deteriorate and therefore antibodies might not bind sufficiently, resulting in a weak staining response. The duration before a satisfiable fixation is acquired is dependent on the used preservative. When using formalin 10 % the duration ranges between 6-12 hours.[Oliver and Jamur, 2009]

After the fixation is complete the tissue placed with cassettes and dehydrated, thereby reducing the water content of the tissue. The tissue sample is then embedded in paraffin. If the tissue is greater than the size of a single cassette it should be cut into smaller portions and embedded separately. After the paraffin casting the samples are sectioned into a paper thin ribbon which is transferred onto a water bath. Each slice is 4 μm in thickness and is carefully separated. These are then placed between two pieces of glass. [Oliver and Jamur, 2009, Chiannilkulchai et al., 1989]

The slices are stained using stains that target specific structure on or within the cell. The choice of staining agent depends on the goal of the staining. The staining agent can either be commercially available or isolated within the laboratory. Where the previous steps illustrate the generic, the staining step can be much more specialised. Depending on the protocol and the use of counterstains, several iterations of washing, preparing, staining may be be needed. As noted by Oliver and Jamur [2009] the use of secondary antibodies that specifically binds to the subclass of the primary antibody gives the best result when using monoclonal antibodies. [Oliver and Jamur, 2009]

An example of agents used in IHC would be the use of Pan-CK, 3,3'-diaminobenzidine (DAB) and Hematoxylin & Eosin (HE). Pan-CK is a mouse monoclonal antibody cocktail and is often used to identify tumor in epithelial cells by binding to specific antigen epitopes [Sorenson et al., 1987]. After the use of Pan-CK it is now easier to locate structures of interest and attach chromogens or other antibodies. DAB is an agent used in staining that stains brown where it is attached. The DAB is activated by the use of peroxidase [Sino Biological, Inc., n.d.a]. This is often used to stain structures of interest. The use of DAB is illustrated in figure 2.3. HE is one of the fundamental stains. Hematoxylin is used to stain the cell nucleus a blue or purple colour whereas the eosin stains the cytoplasm a red or pinkish colour [University of Leeds, n.d.]. The HE is often used as either stain or counterstain in IHC. By this method it is possible to locate tumorous epithelial cell via Pan-CK, stain them brown using DAB, and counterstain nucleus blueish and cytoplasm pinkish. It is especially important in differentiating poorly differentiated carcinomas without morphological characteristics.

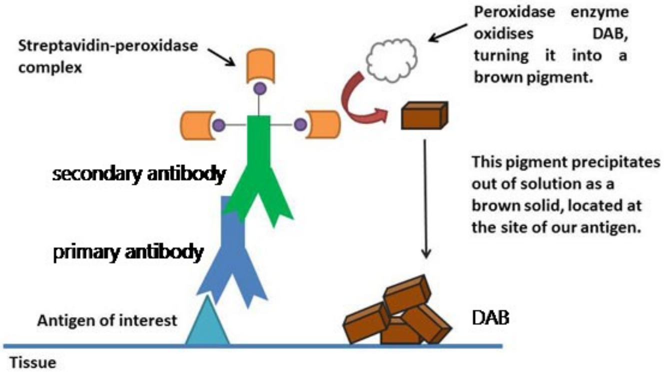

Streptavidin-peroxidase complex
Peroxidase enzyme oxidises DAB, turning it into a brown pigment.
secondary antibody
This pigment precipitates out of solution as a brown solid, located at the site of our antigen.
primary antibody
Antigen of interest
DAB
Tissue

Figure 2.3: An illustration using primary and secondary antibodies to stain using DAB. Image has been modified. [Bitesize Bio, n.d.]

After the staining is complete the pathologist, or in some cases technicians, describe the findings within the sample which ultimately should lead to a diagnosis.

2.2 Variability within IHC staining

The IHC staining is a multi-step process with multiple variables. Therefore it is expected that differences in results to occur even if input is identical. The variability can occur both at a protocol level where the staining results may differ or/and at an interpretation level where the observers' findings deviate. Laboratory variability would concern the protocol level where the observer variability concern the interpretation. This variability has been the subject in multiple studies throughout the last decade. It has been addressed by multiple authors that the variability may lead to difficulties in standardization. [Taylor, 2000, Lin and Chen, 2014, Naik et al., 2008]

Cancer diagnosis depends on the subjective visual inspection of both the clinical and the histopathological information to place tumours in categories based on the tumour's tissue of origin. Clinical information though, can be incomplete or misleading and there is also a wide spectrum in cancer morphology. These complications can result in misjudgements, making second opinions obligatory, leading to an increase of the expense of patient care. [Ramaswamy et al., 2001]

Inter-laboratory variability concerns the difference of IHC results when all laboratories are given identical tissue samples to be stained. This has been examined by several authors and it seems to be a common acknowledgement. [Fitzgibbons et al., 2014, Mengel et al., 2002, von Wasielewski et al., 2002, O'Leary, 2001]

The primary suggestion by most authors has been to attempt standardisation of protocols. However specific indicators of a 'good' staining has yet to be uncovered.

Intra-laboratory variability appears to be a weakly described topic if uncovered at all. However it is possible that the interest in this subject could be based on the future QC and QA of staining. Alternatively this is examined by the laboratory for internal reasons and the results are not published to the community.

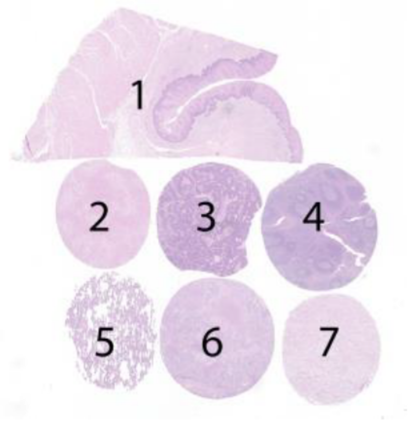
Inter-observer variability was found to be at a minimum by multiple studies. Both Mengel et al. [2002], von Wasielewski et al. [2002] have conducted inter-observer variability studies using data from 172 participants/laboratories. However the number of observers was only 2 perhaps making their inter-observer variability less proof. Also it was suspected that both utilise the same data set in their separate articles. [Mengel et al., 2002, von Wasielewski et al., 2002]

Intra-observer variability does not appear to have been researched. However Ranefall et al. [1997] mentions that assays can depend on the subjectivity of the observer on a daily basis, which does suggest some intra-observer variability is possible. Kirkegaard et al. [2006] also mentions that intra-observer variability can be managed if training and QA is given attention. Another factor, according to Walker [2006], is that as the analysis of immunohistocemistry is mostly manual, it depends a lot on the experience and the alertness of the interpreter suggesting automatic analysis as a more accurate quantifier. [Ranefall et al., 1997, Kirkegaard et al., 2006, Walker, 2006]

2.3 NordiQC Assessments

According to Vyberg2005 the grading of the result of each laboratory's staining protocol is performed in *consensus* by 4 pathologist (DK, NO, SWE, FIN) and 1 (appointed) technician, and the grading scores are *poor*, *borderline*, *good* and *optimal*. [Vyberg et al., 2005, Vyberg and Nielsen, 2016]

In the latest Pan-CK assessment, run 47, the criteria for *optimal* staining are listed. Five criteria have been listed each of which relates to one or two tissue arrays which are seen in figure 2.4.



1. Esophagus, 2. Liver,
3. Small cell lung carcinoma (SCLC),
4. Tonsil, 5. Lung adenocarcinoma,
6. Lung squamous cell carcinoma,
7. Renal clear cell carcinoma (RCC)

Figure 2.4: TMA that a laboratory must stain using their own Pan-Cytokeratin protocols. Image is modified from [NordiQC, 2016]

For TMA-1, the oesophagus, all the squamous epithelial cells should have a strong, distinct cytoplasmic staining throughout all the cell layers. For TMA-2, the liver, there should be a strong, distinct cytoplasmic staining of the epithelial cells in the bile duct, and at least a moderate staining of the cytoplasm of the vast majority of the hepatocytes emphasising membranes. An at least moderate, distinct cytoplasmic, dot-like staining reaction should be achieved in the majority of neoplastic cells from the small cell lung carcinoma in TMA-3 For the lung adenocarcinoma and the lung squamous cell carcinoma, the TMA-5 and TMA-6, the majority of the neoplastic cells should have a strong, distinct staining of the cytoplasm. In TMA-7 at least a weak to moderate, distinct staining of the cytoplasm and membrane should occur in the majority of the neoplastic cells in renal clear cell carcinoma. [NordiQC, 2016]

As previously described the grading is performed as consensus and the criteria to which the samples most comply with in order for the *optimal* grade. However it is not described whether a staining can be graded optimal even if all criteria are not fully met. And if the panel is asked to perform same evaluation of a sample will they reproduce the same over an over again, or can some variation occur?

Figure 2.5 shows two very similar samples from the latest Pan-CK assessment in which the tissue has been graded by NordiQC. The one on the left is graded *optimal* whereas the right is graded *good*.

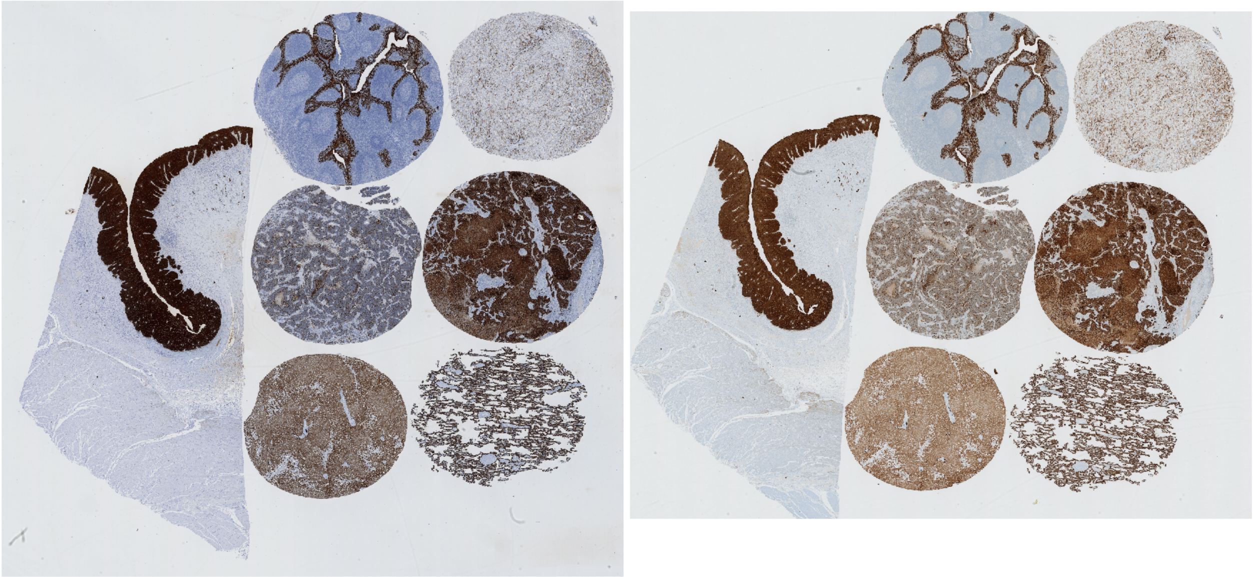


Figure 2.5: *Optimal* sample to the left. *Good* sample to the right.

As with the previous comparison the samples shown in figure 2.6 have been graded. A *borderline* staining to the left and a *poor* to the right

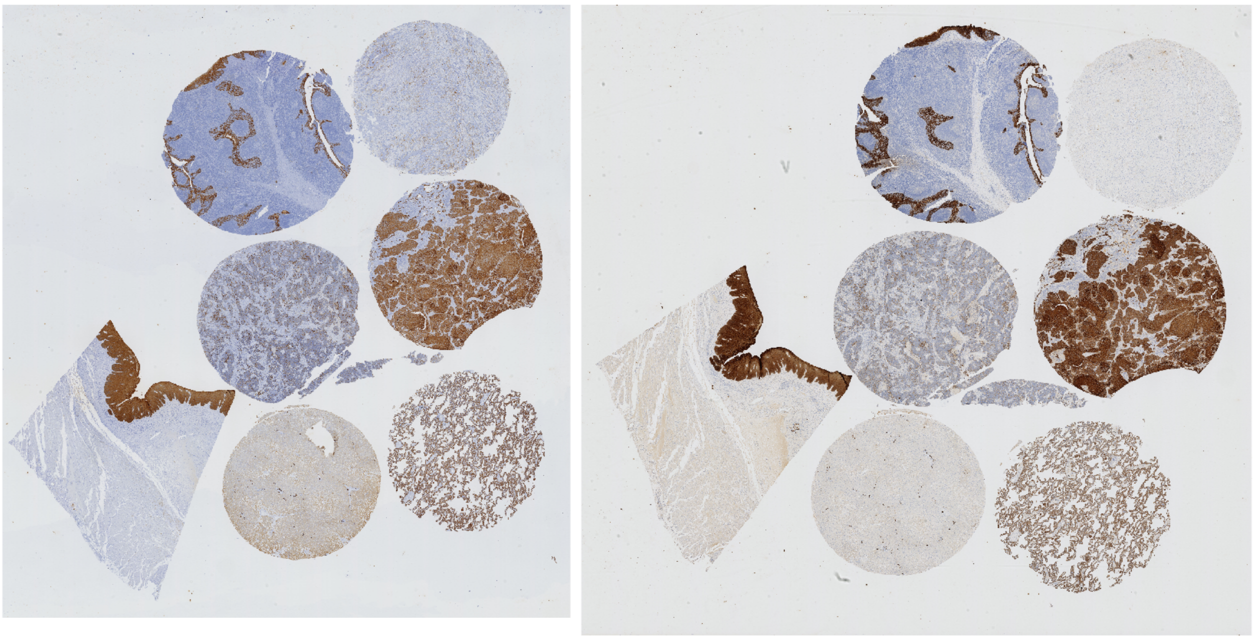


Figure 2.6: *Borderline* sample to the left. *Poor* sample to the right

As mentioned by Torlakovic et al. [2015] both the antibody and the IHC protocol influence the staining result. The use of a high quality antibody with a poor protocol may yield suboptimal results, and the use of difficult low-affinity antibodies with a good protocol may lead to good results. Here the results may come as sensitivity and/or specificity. [Torlakovic et al., 2015]

One approach of determining if the protocol provides a sufficient staining is through the inclusion of tissue with different expression levels in the TMA. High expressors is often tissue with higher amounts of specific proteins whereas low expressors have lower amounts. These are used to demonstrate the overall sensitivity of the protocol and that it is able to detect the targeted expressor. High expressors due to their increased amount of specific proteins are more likely to stain, which can be used to reveal the lower threshold for a positive control. Low expressors can be used as an indicator that the staining protocol is sufficient and that targeted expressors are stained. In the evaluation of the protocol, the high and low expressors can reveal the quality of the staining, and it is this staining quality that can greatly influence the observers ability to detect tissue features leading to cancer diagnosis. [Torlakovic et al., 2014, 2015]

2.4 Computer Aided Decision

Computer Aided Decision (CAD) systems have been researched within multiple different clinical applications. Within the medical field of pathology these are also available. As stated by Kothari et al. [2013] several commercial software tools exist that aid in the objective and quantitative analysis. However in most cases these are not fully automated and the pathologist has to manually select the region(s) of interest (ROI). Furthermore the pathologist is given information feedback based on the ROI which then can lead to a diagnosis. Thus they primarily act as a CAD. To this, one might state that what the commercial tools lack in automating they provide in objectivity. [Kothari et al., 2013]

Tissue biopsies can also have a very high resolution sometimes up to 40000x60000 pixels. These images often contain large amounts of spatial information including low and high-grade tumour, necrosis etc. When manually assessing these images the pathologist identify and study the possible ROI. Due to limitations of technology these often need to be cropped into smaller tiles of 512x512 pixels before image features are extracted when algorithms/systems attempt the same. This however, may be dependent on prior knowledge regarding possible ROI as some tiles can be more interesting than others. Previously multiple researchers have developed supervised models for the identification of these ROI, yet some are now shifting their focus to unsupervised knowledge-based models. [Kothari et al., 2013]

Several researchers have attempted to create algorithms able to detect/diagnose/grade cancer using a standardised grade. These are however often aimed at specific cancer types such as prostate or breast cancer and not general topics such as staining quality. One of such studies is by Doyle et al. [2007] which addresses the inter/intra observer variability of prostate cancer grading using the Gleason scale and the lack of standardisation methodology. In their study 102 graph-based, morphological, and textural features are extracted and used to quantify the tissue information from 4 different tissue classes: benign epithelium, benign stroma, Gleason grade 3 and grade 4 adenocarcinoma. Accuracy between grade 3 & 4 was 76.9 % which according to Doyle et al. [2007] also poses the largest variability and error when pathologists manually score. The remaining 5 combinations had an accuracy ranging between 85.4-92.8 %. [Doyle et al., 2007]

2.5 Problem Statement

IHC is often used by the pathologist to examine tissue in search of cancer as different staining agents can highlight objects/regions of interest. But the staining quality can differ from place to place based on laboratory and observer variability, both inter and intra. The NordiQC collaboration acts as an external QA and provides feedback to laboratory nationwide. This grading process is still performed by clinicians and thus not free from variability. With the progress of technology the next step could be computer aided decision systems targeting IHC staining quality and applied to a NordiQC assessment run.

Therefore the aim of this study is the following:

- To develop an algorithm capable of analysing the NordiQC samples separately
- To investigate the possibility of using image features and classification to describe NordiQC grades
- To validate the results compared to the NordiQC grades
- To explore the plausibility of a CAD system for the NordiQC collaboration

Chapter 3

Data Acquisition

The raw data was a subset of the NordiQC *Assessment Run 47 2016 Pan Cytokeratin*. This was acquired from Pathological Institute at Aalborg University Hospital. The presented data consisted of 39 images and associated NordiQC classification. Each image consisted of one irregular shaped tissue sample and six quasi-circular tissue samples. The subset had been classified as either; optimal, good, borderline or poor. These images are listed in table 3.1.

Table 3.1: Table containing the sample number and its associated NordiQC grade

NordiQC grade and sample no.			
Optimal	Good	Borderline	Poor
1	4	26	15
2	11	27	18
3	12	28	20
5	13	29	21
6	14	30	22
7	16	31	25
8	17	33	32
9	19	34	36
10	23	35	37
	24	38	39

Excluded Data

One of the 39 images was excluded. This was done because it contained to much noise in the form of an air bubble trapped within the sample. Worth mentioning is also that it received a *poor* NordiQC score. The excluded image is shown in figure 3.1.

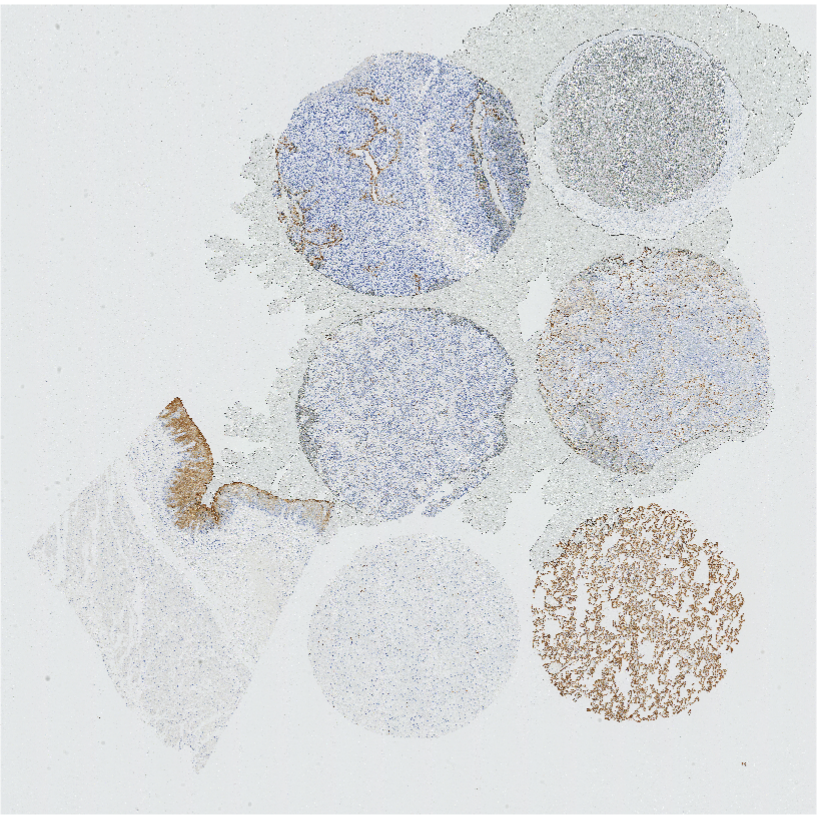


Figure 3.1: The figure shows the sample that was excluded from the data set due to poor quality and high background noise

Chapter 4

Method

4.1 Solution Strategy

This section outlines how the solution strategy was organised. This is done in order to fulfil the aims of this study as described in the problem statement, section 2.5.

The solution strategy consisted of three steps. These are shown in figure 4.1. **Step 1** was the Core Detection which mainly addressed the separation and labelling of NordiQC tissue samples from each other. This makes up section 4.2. **Step 2** was the Core Segmentation which further refined the separated images and prepared for the feature extraction and classification. This is described in section 4.3. Together step 1 and 2 focused on the first study aim of analysing samples separately. **Step 3** focused on classifying the individual cores and validating according to the NordiQC score. This is described in section 4.4. This primarily addressed the second study aim of investigating the possibility of classifying images according to staining quality, and the validation of the results as described in the third study aim. The combination of step 1-3 explored if this project yielded plausibility for a future computer aided decision with this given approach according to the fourth study aim.

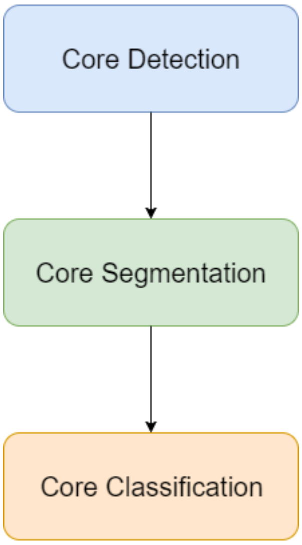


Figure 4.1: An illustration of the 3 steps of the solution strategy

Matlab version 2016b was used in order to design and develop the algorithm in during this project. The algorithm was, just like the solution strategy, made up of several steps allowing for stepwise application if needed during future study.

4.2 Core Detection

The Core Detection separated the individual cores[1] within the slides and the irregular shaped oesophagus tissue. Afterwards an initial segmentation was applied to each tissue core which was followed by a labelling. This was done to make sure that the cores were organised in the correct order for later use. This order is also shown in figure 4.2.

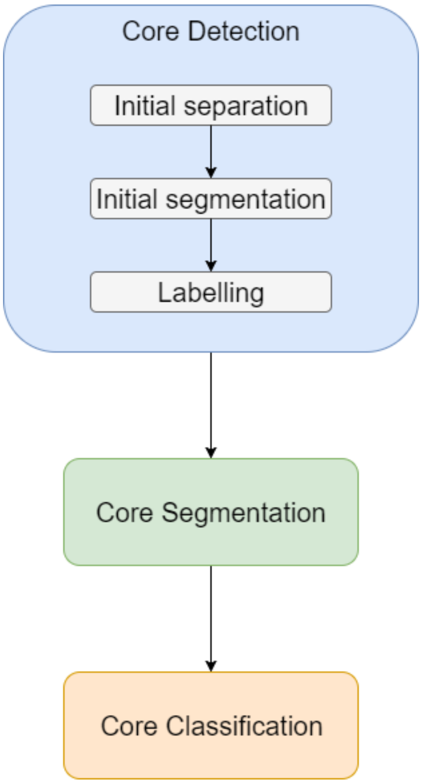


Figure 4.2: Figure showing the first step of the solution strategy and its contents

The flowcharts for the Core Detection are illustrated in appendix A where figure A.1 shows the general flowchart for step 1.

4.2.1 Initial Separation

Before the initial segmentation of the tissue could be done, the image was divided into a region containing mostly oesophagus tissue and a region mostly containing cores. The reason behind was that throughout the 38 images there was inconsistency with respect to the size and irregular shape of the oesophagus tissue whereas the remaining cores were highly consistent in shape and size. Therefore, only the 6 cores were considered for further processing and the oesophagus region was deemed to be beyond the scope of this project. This process is illustrated in the following pseudocode.

```
for all sample do
    load image;
    convert RGB to binary;
    /* image is binary. Identify object using 8-connected */
    use 4/8-connectivity to label image elements;
    find the 7 biggest objects;
    /* assume leftmost object is the oesophagus */
    exclude leftmost object
end
```

Algorithm 1: The pseudocode illustrating how the oesophagus tissue is excluded from the sample

[1] quasi-circular shaped tissue samples

Using 8-connectivity made it possible to identify the largest objects in the image which corresponded to the tissues. When converting the image to binary the use of the grayscale intensity of the background should make sure that it did not interfere when the neighbouring elements were labelled.

In appendix A.2 the flowchart for the algorithm regarding the initial separation is shown in greater detail.

4.2.2 Initial Segmentation

The initial segmentation made use of several steps in order to detect each core which also included that each core should only be detected once per sample.

To identify the cores the approximate area (of a tissue core) was calculated based on visual inspection of the data. It was noted that the image height was roughly 3.5 times the core diameter and that each core almost had the same size. This area should however not differ as the core should be extracted using similar methods. From this the approximate area could be derived as

$$area = \pi * radius^2$$

where

$$radius \approx \frac{image_{height}}{7}$$

This area was furthermore increased by 10% and would describe the maximum area that would be expected of a tissue core. From the approximate area the two other cases could occur. The first was that tissue cores could touch or to some degree overlap which was observed in several samples in the data. This would result in an area greater than the approximated area. The other case was that the core could have a lower area from weak staining, a low expressor or background noise. This led to the following definition of radius and thereby areas:

- touching cores > approx radius
- approx radius > expected core size > 0.5x approx radius
- 0.5x approx radius > weak core > 0.3x approx radius

Any radius and thereby area below the weak core should be considered fragments or a special case that might require a different approach. In figure 4.3 a comparison between core type and their size is shown.

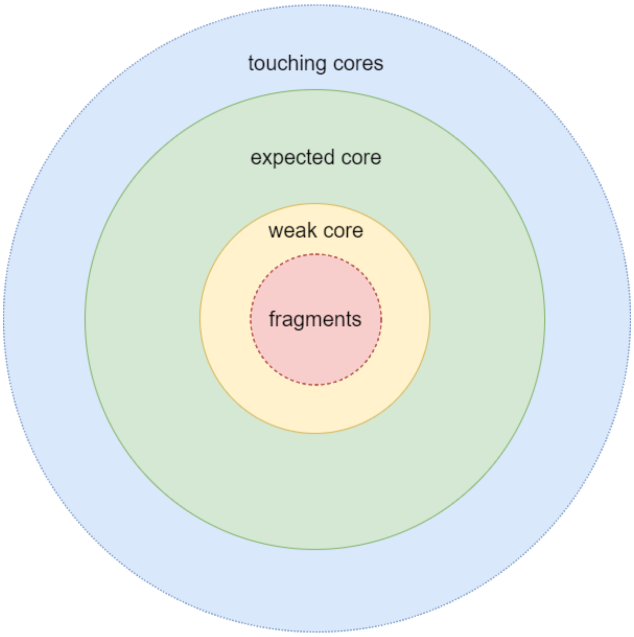


Figure 4.3: A comparison between the size of the defined core types

The process can also be described using the following pseudocode. This can be supplemented by the flowchart figure A.3 where the process is described in greater detail.

```
for all sample do
    load image;
    convert RGB to black/white;
    find image height, approx radius and area;
    while cores identified <6 do
        use 8-connectivity to label image elements;
        if expected core is found then
            get position;
            exclude object/core from image
        else
            if touching cores is found then
                separate cores;
                get position;
                exclude object/core from image
            else
                if weak core is found then
                    get position;
                    exclude object/core from image
                else
                    /*must be fragments or special case*/
                end
            end
        end
    end
end
```

Algorithm 2: Pseudocode that illustrates how the core sizes should be identified and in which order

First the image was converted to grayscale before it was converted into binary using a grayscale threshold. The threshold value would be calculated as the mean of the background. This was done using two squares with the size of 10% image width by 10% image height, which was placed in the top left and in the bottom right of the image. If the threshold value exceeded upper or lower limit, a default value of 0.8 was used.

After the binary conversion the *regionprops* was used to find the *boundary boxes* in Matlab. This would find binary clusters and by sorting these by area in descending order would reveal the different types of cores as shown in figure 4.3. This would mimic the labelling through 8-connected neighbours. When segmenting the core from the image it would have its radius increase by 10% to ensure that all information about this core should be included. The process of classifying and excluding the core is illustrated in greater detail in figure A.4

An issue could be that the cores had a weak staining, and therefore parts of the core would be considered background when the image was binarised. This was a recurring issue for two of the cores. This would make the area fall below the threshold for a weak core and thus not be included. To counter this the threshold would shifted thus making the extraction of weak cores possible, through the error handling.

4.2.3 Labelling

From the Initial segmentation it was expected that a core wouldn't always be segmented in the same order when iterating through the data set. Therefore each core was sorted by coordinates and labelled accordingly. The labelling of cores is shown in figure 4.4

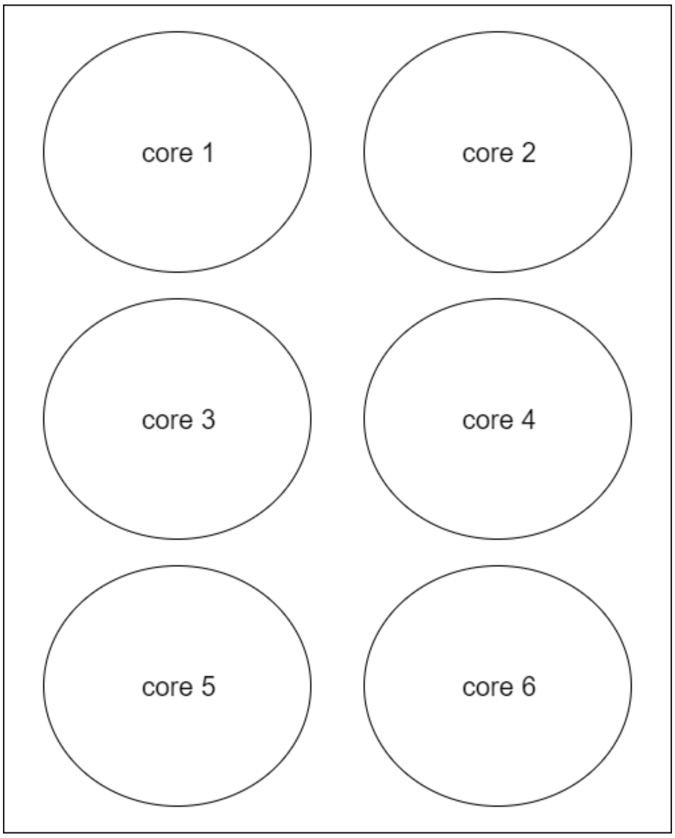


Figure 4.4: The labelling of each core with a specific number

4.3 Core Segmentation

After the Core Isolation each set of images, ranging in category from 1 to 6, would now consist of 38 images. These images would mainly contain the tissue core of interest together with surrounding background and/or parts of adjacent tissue. The tissue core of interest would be at the centre of the image, and the image itself would be 10 % taller and/or wider than the cores radii, which ever was the largest. To segment the tissue core from the background in the remaining image a deformable model was used. This would allow for the creation of a mask which was used to extract the majority of the core and minimize inclusion of background. The image segmentation process is illustrated in figure B.1.

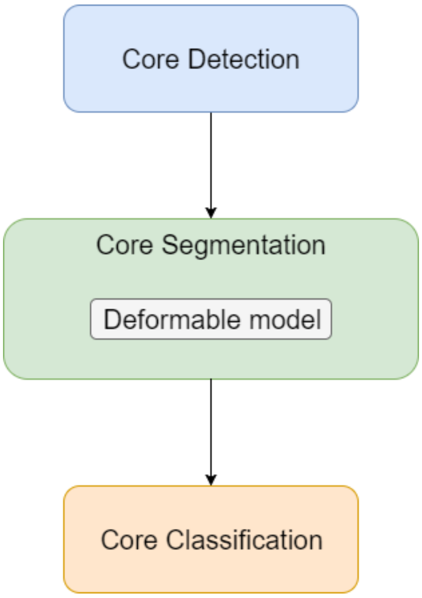


Figure 4.5: Figure showing the second step of the solution strategy and its contents

The process of the Core Segmentation and its deformable model can also be described using the following pseudocode.

```
for all cores do
    for all samples do
        load image;
        add 500 point to form a circle in the image;
        for 1:150 do
            apply deformable model;
            move points in image
        end
        create mask;
        extract ROI
    end
end
```

Algorithm 3: The pseudocode illustrating the Core Segmentation procedure

4.3.1 Deformable Model

Deformable models are used to segment object using energy formulations. A contour shape is defined within the image where it is deformed to the object in order through determining the lowest energy level. When the contour encloses the object the energy level should be at its lowest. There are several different types of contour representations. Given the data a paramteric contour representation would be more suitable to use. This is due to the fact that all the cores show circular or quasi-circular shape, which would not define the shape of the object as complex. Also no self-entangling was seen possible. This would make the parametric representation more ideal compared to the geometric representation. Had there been topological changes then the geometric representation would be preferred. [McInerney and Terzopoulos, 1996, Sonka and Fitzpatrick, 2000]

The mathematical expression of the deformable model is a curve C(s)=(X(s),Y(s)), s ∈ [0 1]. This curve moves across the image where its goal is to minimize an energy function E(C) consisting of internal and external energy. Through iterations the lowest total energy is the goal.

$$E(C) = E_{internal}(C) + E_{external}(C)$$

The internal and external energies can be further elaborated and weighting parameters can be added. The internal energy can also be split into two terms where α weights the tension/elasticity and β weights the rigidity/bending. For the external energy the γ can be used for weighting.

$$E(C) = \alpha E_{tension}(C) + \beta E_{rigidity}(C) + \gamma E_{external}(C)$$

The internal energy is responsible for holding the contour together and determining the elasticity and the rigidity. The E$_{tension}$ is described by the first order partial derivative and the E$_{ridigity}$ is by the second order partial derivative of the curve at point s. [Demir and Yener, 2004, McInerney and Terzopoulos, 1996, Sonka and Fitzpatrick, 2000]

$$E_{internal}(C) = \int_0^1 \alpha(s)|\frac{\partial C}{\partial s}|^2 + \beta(s)|\frac{\partial^2 C}{\partial s^2}|^2 ds$$

The external energy is responsible for pulling the contour towards the desired objects boundaries. This is done by integrating the potential external energy function P(x,y) along the contour C(s).

$$E_{external}(C) = \int_0^1 P(C(s))ds$$

The potential external energy function P is driven by a gradient operator ∇ and a Gaussian filter with the standard deviation σ which is convoluted with the gray-level image I(x,y).

$$P(x,y) = -w_e|\nabla[G_\sigma(x,y) * I(x,y)]|^2$$

To find the curve C(s) that holds the minimum energy can be done by using Euler-Lagranges equation as the curve that minimizes E(C) satisfies this. [Sonka and Fitzpatrick, 2000]

$$\frac{\partial}{\partial s}(\alpha\frac{\partial C}{\partial s}) - \frac{\partial^2}{\partial s^2}(\beta\frac{\partial^2 C}{\partial s^2}) - \nabla P(C) = 0$$

Snake Model

A modified version of the deformable model made by Dahl and Dahl [2016] was implemented into the algorithm. The deformable model was a dictionary snake and was based on their previous research from the paper Dahl and Dahl [2014]. The original version was modified to take a RGB image matrix as an input, and return the points that made up the final contour. The parameters α and β were set as 3 and 1.5 respectively which were determined by empirical process. The γ did not appear as a parameter in their work.

The a priori knowledge used with the snake was that the image would be 10 % larger than the previously isolated core, and since this core itself primarily was qausi-circular in shape, it was assumed that the starting radius of the snake could be 70 % of image radius. This could reduce amount of iterations needed before the boundary would be reached. Subsequently the number of iterations was set as 150. The starting shape of the snake itself would be a circle made up of 500 points equidistantly spaced along the circumference. This would result in a circle inflating from within the targeted core and it should stop at the boundary between tissue and background.

From the snake model, 500 points were retrieved which made up the boundary. Linking these points together represent the deformable model with the lowest total energy level.

The points alone made up a discrete boundary between the core and the background. Before the extraction mask could be created the points were connected. This was done by determining the euclidean distance between the current point and the previous point. Using the maximum distance would ensure that all points were connected. This was followed by filling the inside of the contour.

The mask itself was made binary with background having the value of 0 and the ROI being 1. This was then used to extract the majority of the core within the image. The result was then placed into a new image where the background had no RGB values. By doing so, it would be possible to determine the total area of the extracted by measuring the amount of non-zero elements in the image matrix.

4.4 Core Classification

From the Core Segmentation the cores should now mainly consist of tissue, which also included the exclusion of other tissue cores. From these the features could be extracted which would later be used in the classification and validation of the samples.

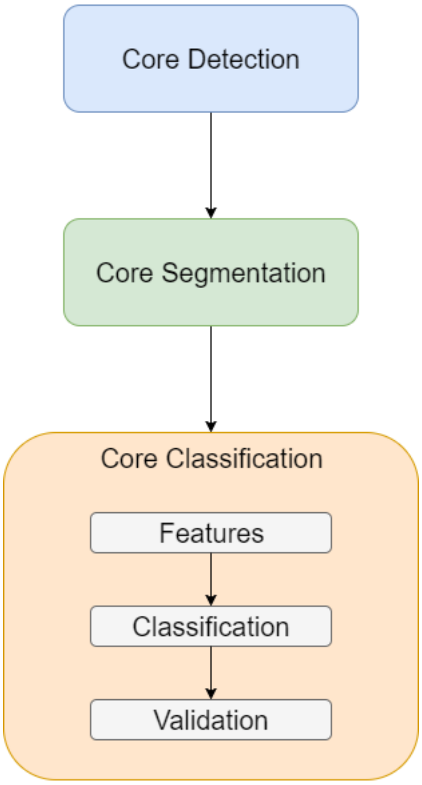


Figure 4.6: Figure showing the third step of the solution strategy and its contents

4.4.1 Features

The selected features were split into two groups. The first was the features that were based on intensity, while the second group of features were based on the dominance of specific colours.

Intensity-based features

The features can be defined at either a cellular or a tissue-level. The cellular aims to quantify properties within the cell but does not consider the spatial information between each cell. At a tissue-level the distribution of the cells throughout the tissue is quantified thus the spatial information would be considered. Throughout the whole data set it's safe to assume that the tissue was cancerous given its nature, and thus no spatial information was necessary. As no spatial information should be needed the choice of features should rely on cellular-level features. When visually inspecting the images the preliminary difference between the different grades; optimal, good, borderline, poor, was the colour intensity. This corresponds well with the fact that tissue staining being the main difference. This led to the approach of using intensity-based features as mention by Demir and Yener [2004]. In their paper it's mentioned that colour value can be obtained at a pixel level in one or more colour channels. The mean value, standard deviation (SD) and ratio between channels were some of the features mention. [Demir and Yener, 2004]

This was expanded upon by including multiple colour spaces which were Red/Green/Blue (RGB), Cyan/Magenta/Yellow/Black (CMYK) and Hue/Saturation/Value (HSV) as these were all well established colour spaces.

Table 4.1: Table of the initially planned features and the corresponding colour spaces

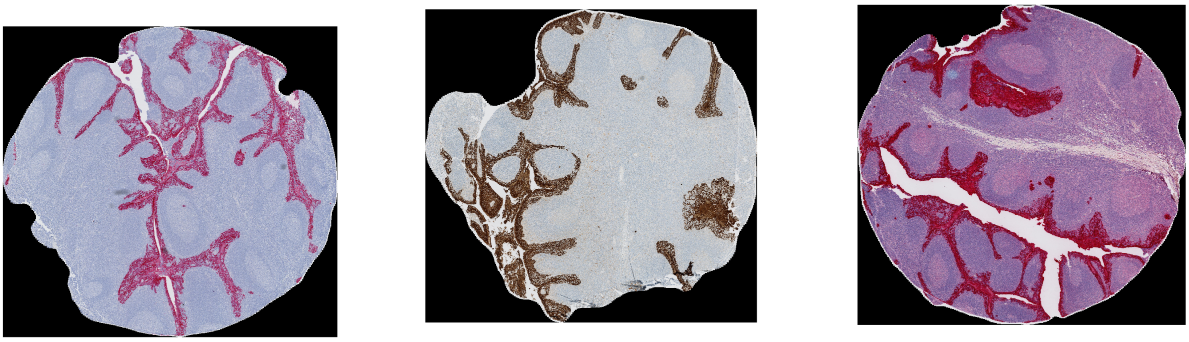
Feature	RGB colour space	CMYK colour space	HSV colour space
Non-zero elements	RGB image Red channel Green channel Blue channel	CMYK image Cyan channel Magenta channel Yellow channel Black channel	HSV image Hue channel Saturation channel Value channel
Ratio	Red/Green Red/Blue Green/Blue	Cyan/Magenta Cyan/Yellow Cyan/Black Magenta/Yellow Magenta/Black Yellow/Black	Hue/Saturation Hue/Value Saturation/Value
Mean	RGB image Red channel Green channel Blue channel	CMYK image Cyan channel Magenta channel Yellow channel Black channel	HSV image Hue channel Saturation channel Value channel
Standard Deviation	RGB image Red channel Green channel Blue channel	CMYK image Cyan channel Magenta channel Yellow channel Black channel	HSV image Hue channel Saturation channel Value channel
Entropy	RGB image Red channel Green channel Blue channel	CMYK image Cyan channel Magenta channel Yellow channel Black channel	HSV image Hue channel Saturation channel Value channel
3rd moment (skewness)	Red channel Green channel Blue channel	Cyan channel Magenta channel Yellow channel Black channel	Hue channel Saturation channel Value channel
4th moment (kurtosis)	Red channel Green channel Blue channel	Cyan channel Magenta channel Yellow channel Black channel	Hue channel Saturation channel Value channel

Using these features combined with the three colour spaces results in 84 features each of which was applied to the different cores throughout the data set.

Staining Colour specific features

Besides the 84 features listed in table 4.1 a few more features were added. These features target the primary staining colours in the stained sample from the Pan-CK protocol.

In figure 4.7 the colour combinations are shown for Core 1. Here the difference between the dominant and the secondary colour was most obvious. One case would be where Core 1 had a blue and purple combination as shown in figure 4.7a. Here the blue colour would be the dominant colour and purple is the secondary colour. Another case would be where Core 1 had a blue and brown colour combination as shown in figure 4.7b, blue being the dominant and brown being the secondary. The last case was when Core 1 had a purple and pink colour combination as illustrated in figure 4.7c



(a) Blue-Pink colour combination.

(b) Blue-Brown colour combination.

(c) Purple-Pink colour combination.

Figure 4.7: Illustration of cores with a different colour combination as a result of the staining.

For ease of reference blue and purple are being referred to as dominant colour while pink and brown are being referred to as secondary colour.

The colour combination was recognized by comparing the maximum intensity of the magenta and the yellow channel in the CMYK colour space. The channel with the largest value was chosen. A threshold of twice the mean intensity of the chosen channel was set. The secondary colour of the core was a result from the pixels with a value less than the threshold. The dominant colour was a result of the subtraction of the black channel and the secondary colour. The steps can be seen in Algorithm 4 and illustrations are available in subsection 5.3.1.

```
for all cores do
    for all samples do
        calculate the amount of non-zero pixels;
        convert RGB to CMYK;
        if Yellow channel < Magenta channel then
            threshold = 2x mean of Magenta channel;
            select Magenta channel
        else
            threshold = 2x mean of Yellow channel;
            select Yellow channel
        end
        for the selected channel make all pixels < threshold = 0;
        /*the resulting is the Secondary colour*/
        get Secondary colour;
        /*limit found by inspection*/
        for the K channel make all pixels < limit = 0;
        /*get the Dominant colour*/
        Dominant = selected channel - K channel
    end
end
```

Algorithm 4: The pseudocode illustrating how the Dominant and Secondary colour should be found.

The mean, standard deviation, and ratio were then calculated for the dominant and secondary colours as well. This led to the following stain specific features:

- Dominant colour ratio of Non-zero pixels
- Secondary colour ratio of Non-zero pixels
- Dominant / Secondary colour ratio
- Mean of Dominant colour
- Mean of Secondary colour
- Standard deviation of Dominant colour
- Standard deviation of Secondary colour

As a result, 91 final features were selected.

4.4.2 Classification

Classification is the process of selecting which class an observation best fits by using training datasets of observations where it is already known where the classes belong [Kim, 2010]. In the case of this thesis the aim of the classification is to determine the staining quality of each core and compare this to the assessment performed by NordiQC. Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are the most commonly used approaches in classification tasks [Franklin, 2005]. Both have advantages and disadvantages though. LDA requires Gaussian distributed classes and homogeneous variances of the classes leading to linear decision boundaries. In QDA, even if a Gaussian distribution is also assumed, the variances are expected to be heterogeneous which provides arbitrary decision boundaries taking any form [Dougherty, 2012]. Since the core types are not identical the covariance between the features and thereby between the cores was assumed to differ among the four gradings. For this reason, the QDA was chosen for classification purposes.

Features Excluded

During the feature extraction it was noticed that there were some features having the same value for all the different samples. Having zero variance or producing singular covariance matrices would lead to errors during cross validation thus it was decided to exclude these. Moreover, there were some Inf[2] values or NaN[3] for some features that would give an error as well. Therefore it was decided to exclude all the features with at least one Inf or NaN value.

4.4.3 Dimensionality Reduction

Dimensionality of the feature space as well as the number of training samples are important factors for the complexity of the classification. When having a small amount of training datasets and large dimensionality then some of the features may be irrelevant or unnecessary, thereby affecting the performance of the classifier by over fitting the classifier to the data. This is called the *curse of dimensionality.* [Dougherty, 2012]

The decrement of the amount of features improves the prediction accuracy, simplifies the data collection, reduces both storage space, and classification procedure time [Neumann et al., 2005]. There are two major methods for feature reduction: feature selection and feature transformation. The objective of the first is to pick an optimal subset of features, while the latter aims to find an optimal transformation of the initial feature space for a new space with lower dimensions. [Dougherty, 2012]

The feature selection method was selected, as it keeps the clarification of the data. [Janecek et al., 2008]

[2] infinite

[3] Not-a-Number

Feature Selection

Feature selection techniques usually consists of three methods: filtering, wrapping, and embedded. [Stańczyk and Jain, 2015]

In the filtering method a feature subset is selected by statistical measures and not by using a classifier learning algorithm. Filter methods can easily be scaled to high-dimensional datasets and can be computed fast. However, each feature is considered independently without taking possible correlations between features into consideration, which compared to other techniques may result to worse classification performance. [Stańczyk and Jain, 2015]

The wrapping method, on the other hand, tests the performance of each feature subset by using a learning algorithm, which leads to an optimal performance of the classification algorithm. For N features there are 2N-1 potential feature subsets [Guyon et al., 2008]. However, it often goes through all the subsets, making it necessary to combine it with additional methods to limit the search space of the feature subsets [Stańczyk and Jain, 2015]. The two techniques mainly used to reduce the search space are the sequential forward selection (SFS), where features are added one at a time, and the backward elimination (BE), where features are removed gradually [Dougherty, 2012].

In the embedded methods the training of the classifier as well as the feature selection are done at the same time. This way, the selected features are only accurate for the specific classifier. [Stańczyk and Jain, 2015]

The wrapping method can be implemented easier than the embedded method and compared to the filtering method, it performs better [Saeys et al., 2007]. For this reason a wrapping method was selected. One of the problems of the SFS is that when a feature has been selected it is not possible to be unselected even if it may be unnecessary later in the process [Guyon and Elisseeff, 2003]. To avoid this, both SFS and BE were implemented.

Sequential forward selection is an algorithm that starts searching with an empty feature subset and adds features one at a time which is decided by the learning algorithm used [Marcano-Cedeño et al., 2010]. The classification was done using QDA, meaning that in the SFS approach this was the learning algorithm of interest. The features of the SFS were calculated from the QDA accuracy, meaning that the selected feature should increase the classification accuracy if it was to be included in the feature subset. On the first iteration the algorithm selects the feature that gives the highest accuracy and on the following iterations a feature is added if it produces a higher accuracy in combination with the already existing set. In case an iteration doesn't include any features that increase the accuracy, the algorithm stops without adding any additional features to the feature subset.

Backward elimination is similar to SFS. It initially starts with the whole feature space and in every iteration it discards the feature whose removal will result in an increase of the accuracy. The BE algorithm stops when the removal of any more features will result in a decreased accuracy. The remaining feature space will then only consist of features that contribute to the accuracy in a positive manner. [Guyon and Elisseeff, 2003]

4.4.4 Validation

To estimate the accuracy of a classifier generated by the supervised learning algorithms, the estimation of its future prediction accuracy is important. For this estimation, a method with low bias and also low variance is needed. [Wolpert, 1992]

Even though the Leave-One-Out Cross-Validation (LOOCV) method is nearly unbiased, it may produce unreliable estimates because of the high variance [Efron, 1986]. For that reason the M-fold method was selected. In M-fold method the dataset is split randomly into M subsets about the same size. The inducer uses all the subsets except from one to train and is then tested with the remaining subset. This will happen M times, with the subset left out changing every time. [Kohavi et al., 1995]

The accuracy of a classifier is the possibility of predicting correctly a random instance's class. The cross validation was totally based on accuracy, but this measurement only represents the number of errors and not what sort of errors that occur. This is the reason why sensitivity and specificity were also calculated. [Dougherty, 2012]

Sensitivity describes how correctly a model identifies the class an instance belongs to, while specificity describes the model's ability to correctly identify the instances that don't belong in a certain class. [Dougherty, 2012]

Accuracy, sensitivity and specificity is calculated using the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) as stated in the following formulas [Baratloo et al., 2015].

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

For our method, as a TP was when the classifier produced the same result in compliance with the NordiQC grade. A TN was the avoiding of a misclassification to a given grade. An example for the optimal classifier can be seen in table 4.2.

Table 4.2: Example of confusion matrix for Optimal Classifier. 'o' = optimal grade, 'g' = good grade, 'b' = borderline grade, 'p' = poor grade.

Optimal Classifier

	Positive	**Negative**
True	Classifier: 'o' NordiQC grade: 'o'	Classifier: 'g'/'b'/'p' NordiQC grade: 'g'/'b'/'p'
False	Classifier: 'o' NordiQC grade: 'g'/'b'/'p'	Classifier: 'g'/'b'/'p' NordiQC grade: 'o'

One-versus-all

In order to reduce the classifying problem of four classes One-Versus-all (OVA) was used. By doing so the classification was reduced into four binary problems. Each one of them, distinguishes a given class from the remaining three classes. For this approach, four binary classifiers were required, where the 4th classifier was trained with positive examples that belonged to class 4 and negative examples that belonged to the other three classes. [Bishop, 2006]

Even though this approach was simple, its performance could provide results that were often as accurate as other methods. [Rifkin and Klautau, 2004]

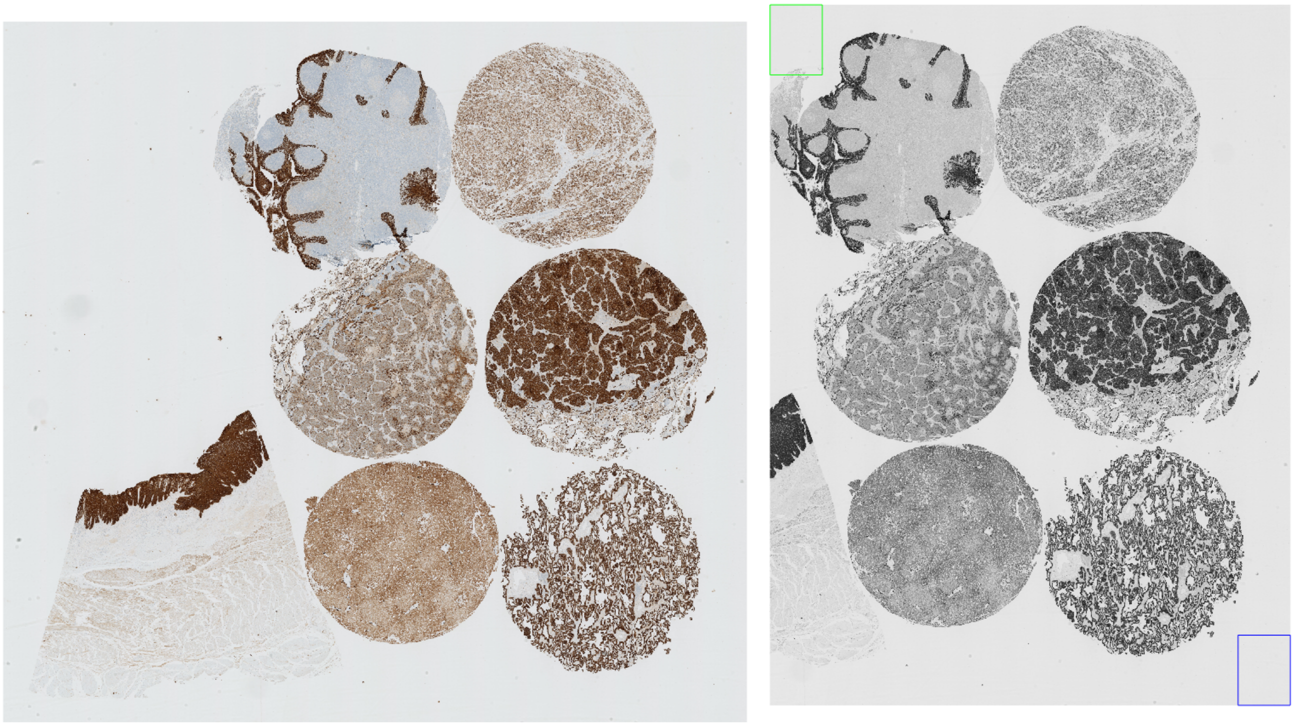
Chapter 5

Results

5.1 Core Detection

5.1.1 Initial Separation

Before the initial segmentation of the original image (5.1a), the oesophagus was cropped from the whole sample. The result is shown in figure 5.1b.



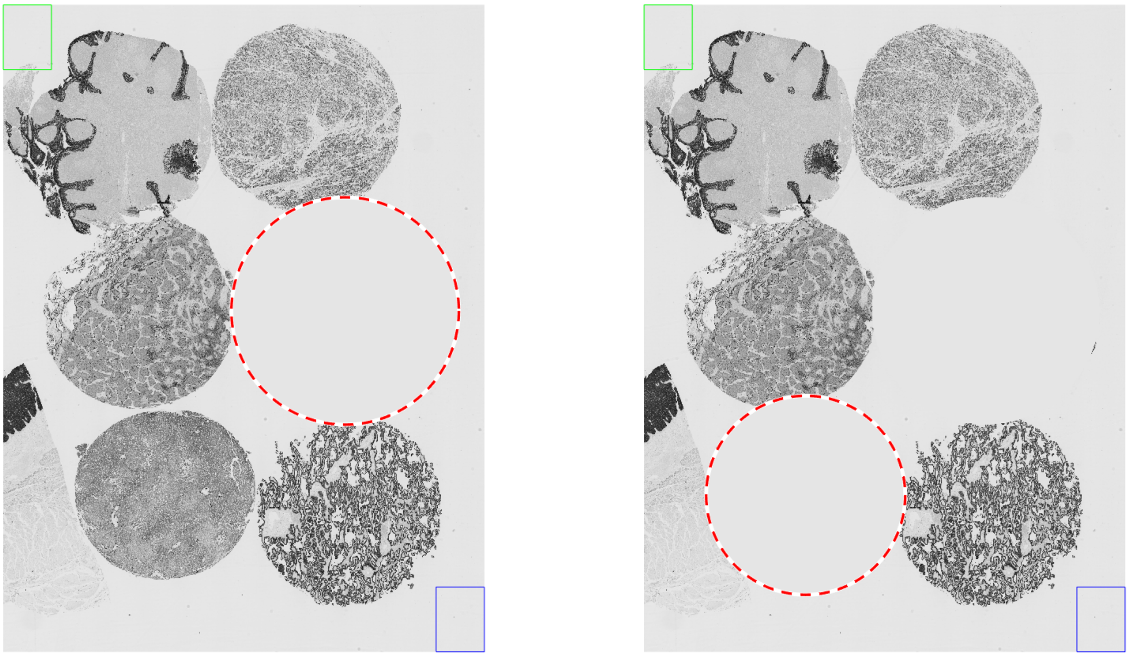
(a) Illustration of the original image in RGB

(b) Illustration of the cropped image without the oesophagus tissue.

Figure 5.1: Illustration of the removal of oesophagus.

5.1.2 Initial Segmentation

When isolating the remaining six cores the cropped image would also use the two rectangles to calculate the background after the image was grayscaled. Once the core with the 'biggest' area was identified, it was then copied into a separate image file from the original RGB image (figure 5.1a), and the area would be concealed in the grayscale image (figure 5.1b) using the background intensity values found earlier. The removal of the extracted core and the visible remaining cores are shown in figure 5.2a and figure 5.2b for the first two in sample 1.

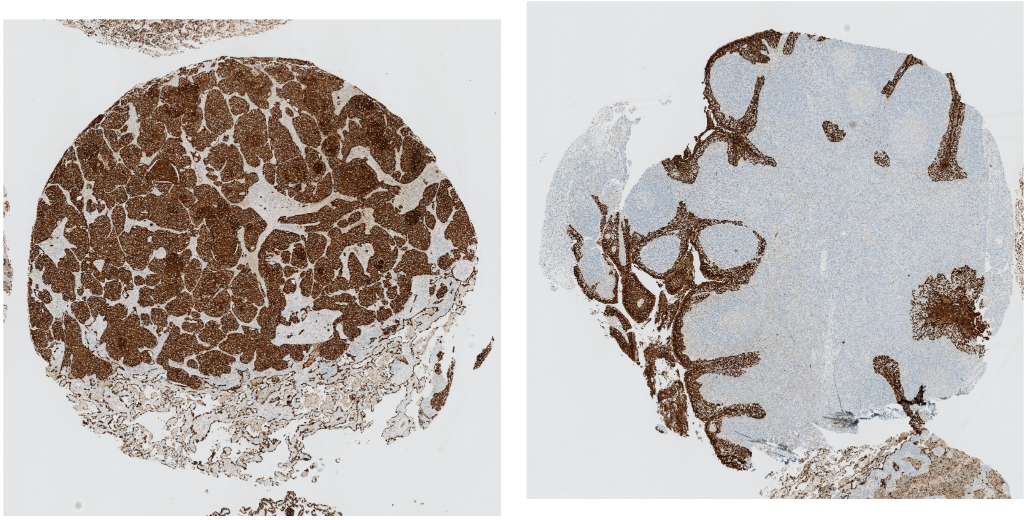


(a) First core detected, core 4.

(b) Second core detected, core 5.

Figure 5.2: Detection of the cores in sample 1. The red dashed line illustrates the extracted area as it has been excluded.

The extracted cores from the RGB image could fall into two types. The first would be regular cores where they would be separated from other cores. The other would be irregular where tissue cores would/could touch or even overlap. In figure 5.3a a regular core has been segmented, and in figure 5.3b a irregular core is shown as it touches core 2 and 3.



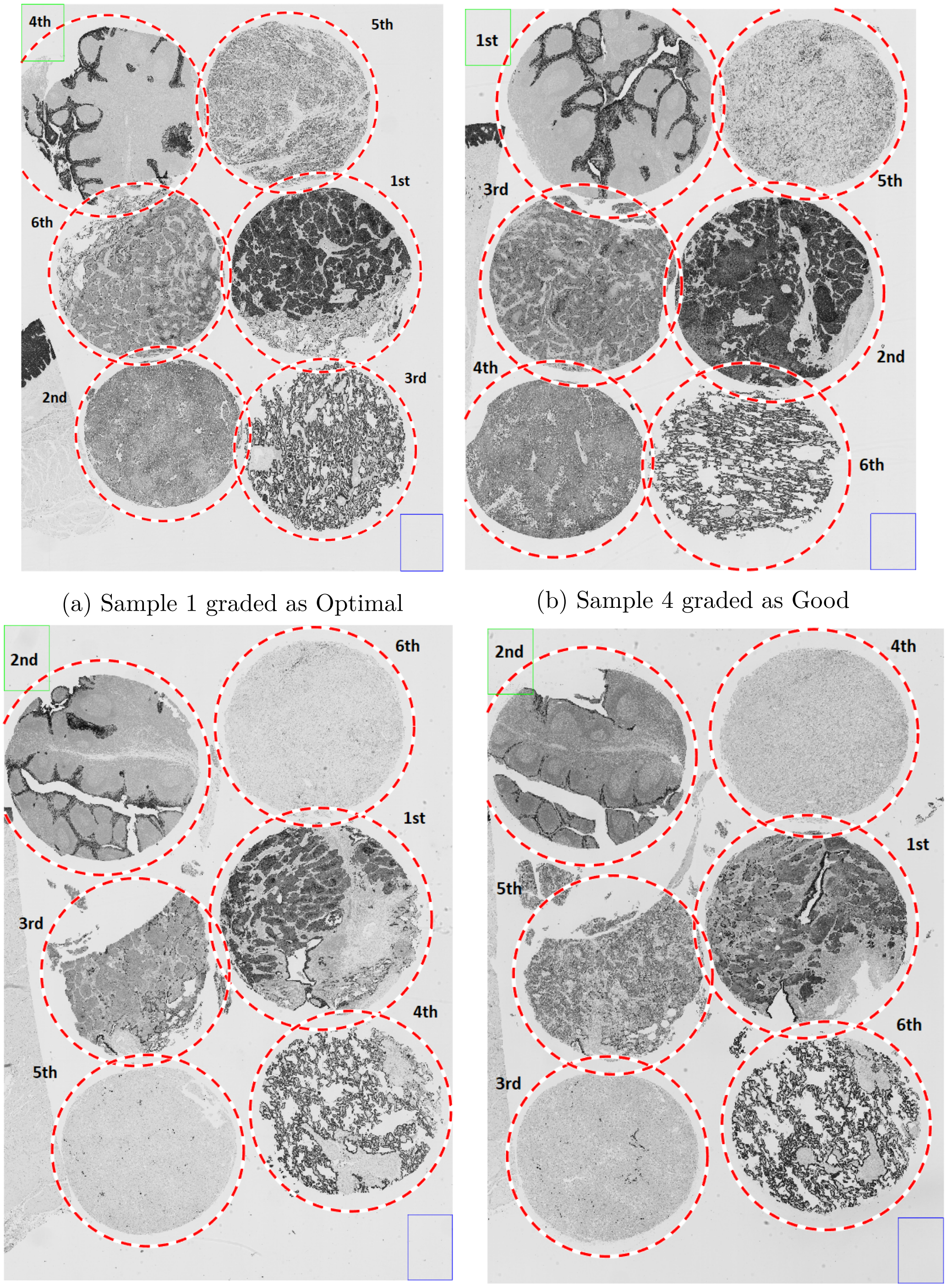
(a) The figure shows core 4 that was extracted from figure 5.1a

(b) The figure shows the core 1 that was extracted from figure 5.1a. The core touches core 2 to the right and core 3 below.

Figure 5.3: Extracted cores of sample 1.

5.1.3 Labelling

The order in which the tissue cores were extracted could vary greatly in between the 38 samples. The expected sized cores would be extracted first, then overlapping/touching cores followed by weak cores. As illustrated in figure 5.4a it is clear that core 1 was not always the first core to be extracted. This therefore had to be addressed such that the cores were labelled following the order shown in figure 5.5



(a) Sample 1 graded as Optimal

(b) Sample 4 graded as Good

(c) Sample 30 graded as Borderline

(d) Sample 22 graded as Poor

Figure 5.4: Illustration of the order in which the cores were extracted for samples with different grading.

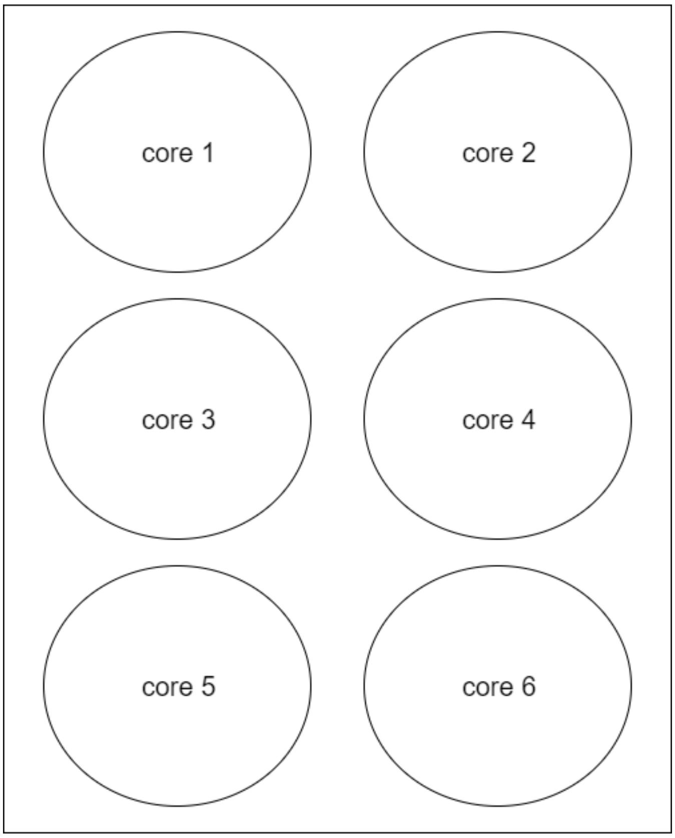


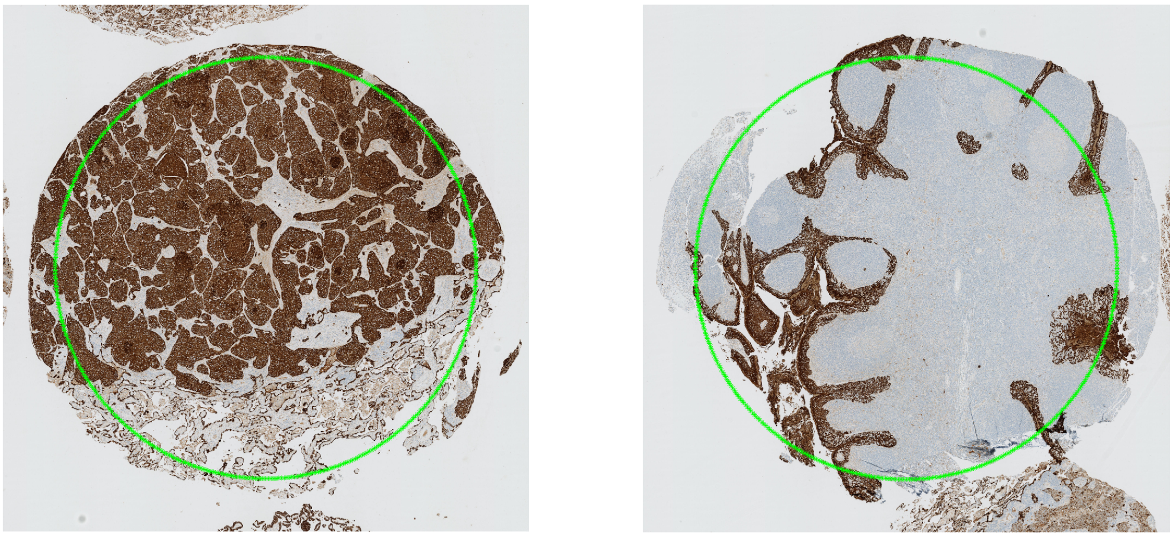
Figure 5.5: The labelling of each core with a specific number

5.2 Core Segmentation

5.2.1 Deformable Model

The use of the snake function resulted in the extracting of the majority of the tissue core while excluding as much background as possible. In some cases the snake function omitted minute parts of the tissue core. However an overall acceptable segmentation was achieved for every 38 images regarding each tissue core. This was consistent throughout both the regular and irregular cores.

The starting position of the deformable model is shown in figure 5.6a for a regular core and figure 5.6b for an irregular core.

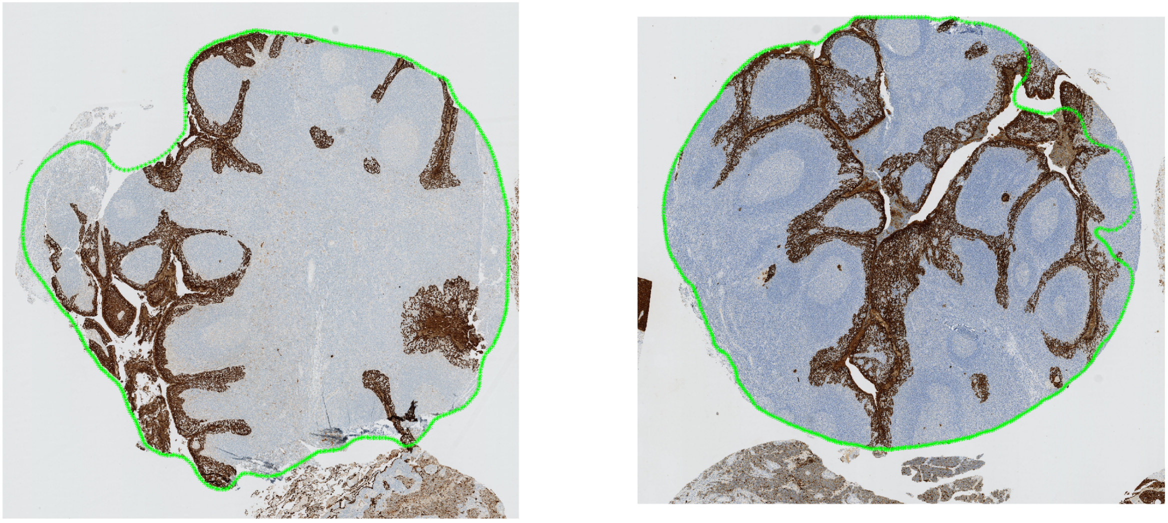


(a) Initial contour for core 4 of sample 1.

(b) Initial contour for core 1 of sample 1.

Figure 5.6: Initial contour did not depend on whether the selected core was connected to others or not.

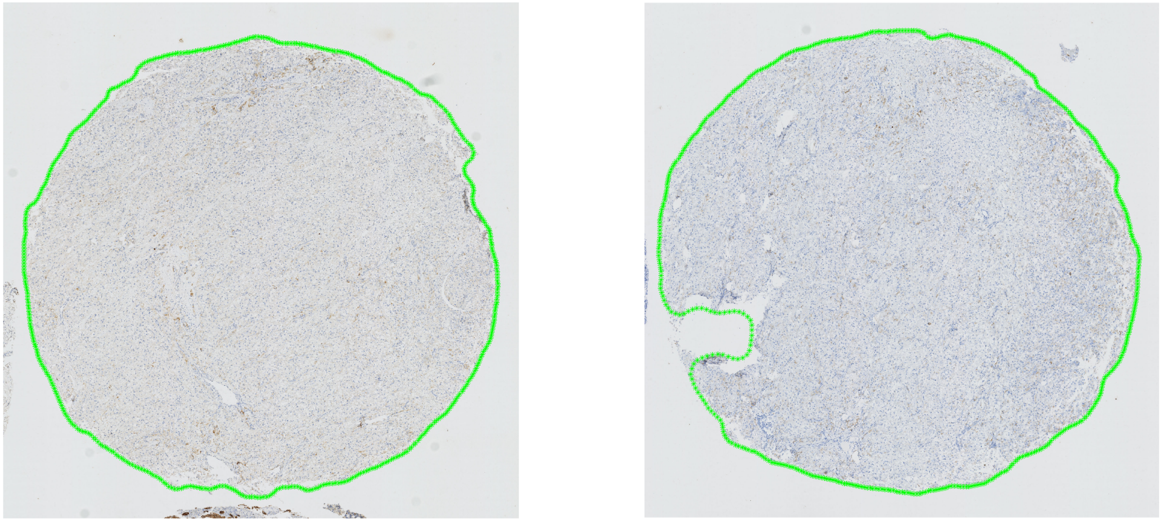
The outcome of the used deformable model is shown in figure 5.10a for the regular and figure 5.7a for the irregular core. The figures 5.7-5.12 all show a good versus a sub-optimal segmentation for every core. The green boundary visualises the minimum total energy after 150 iterations. The green boundary itself consisted of 500 points and at a pixel level it only created a discrete boundary between the core and the background.



(a) Good example of the deformable model in sample 1.

(b) Sub-optimal example of the deformable model in sample 19.

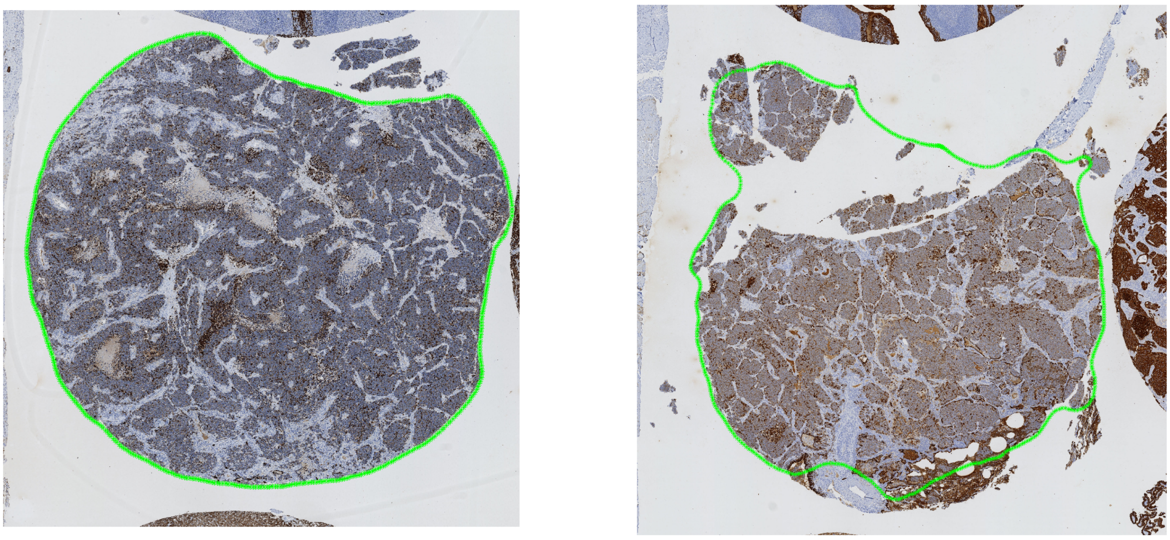
Figure 5.7: The final contour on core 1.



(a) Good example of the deformable model in sample 31.

(b) Sub-optimal example of the deformable model in sample 15

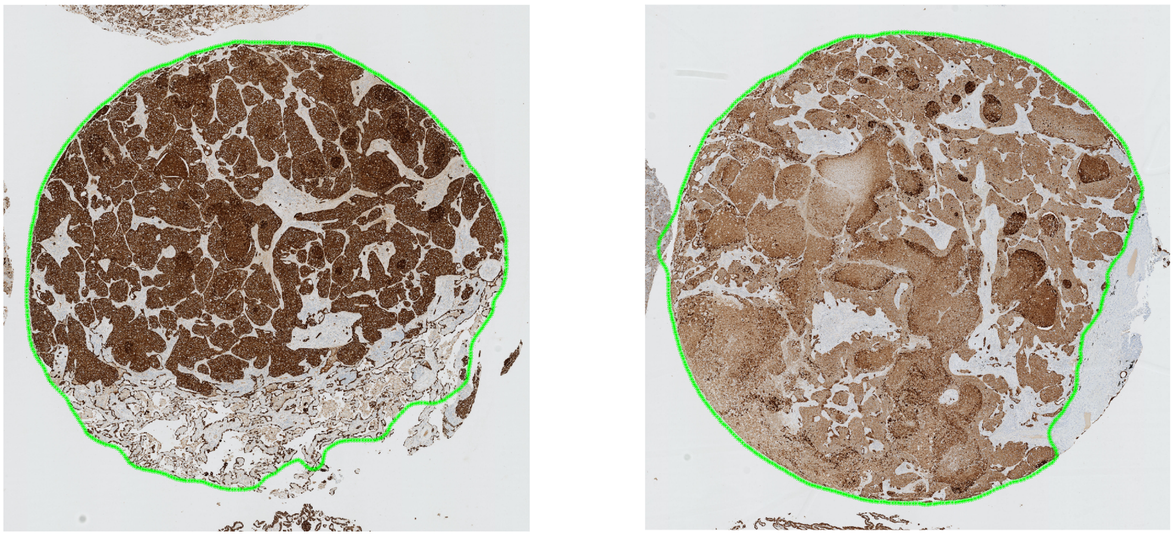
Figure 5.8: The final contour on core 2.



(a) Good example of the deformable model in sample 5.

(b) Sub-optimal example of the deformable model in sample 23.

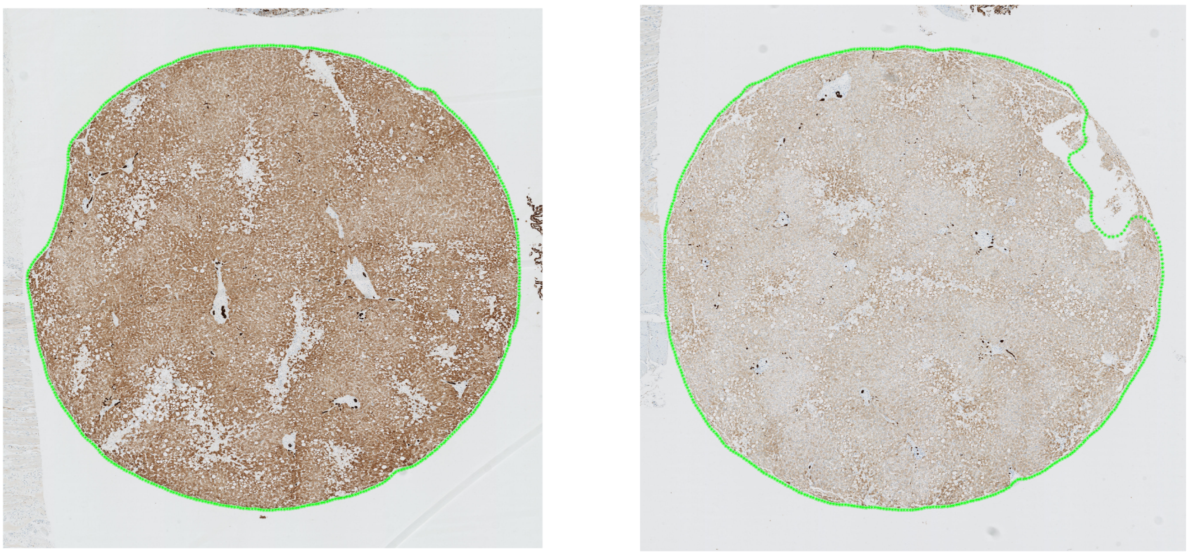
Figure 5.9: The final contour on core 3.



(a) Good example of the deformable model in sample 1.

(b) Sub-optimal example of the deformable model in sample 14.

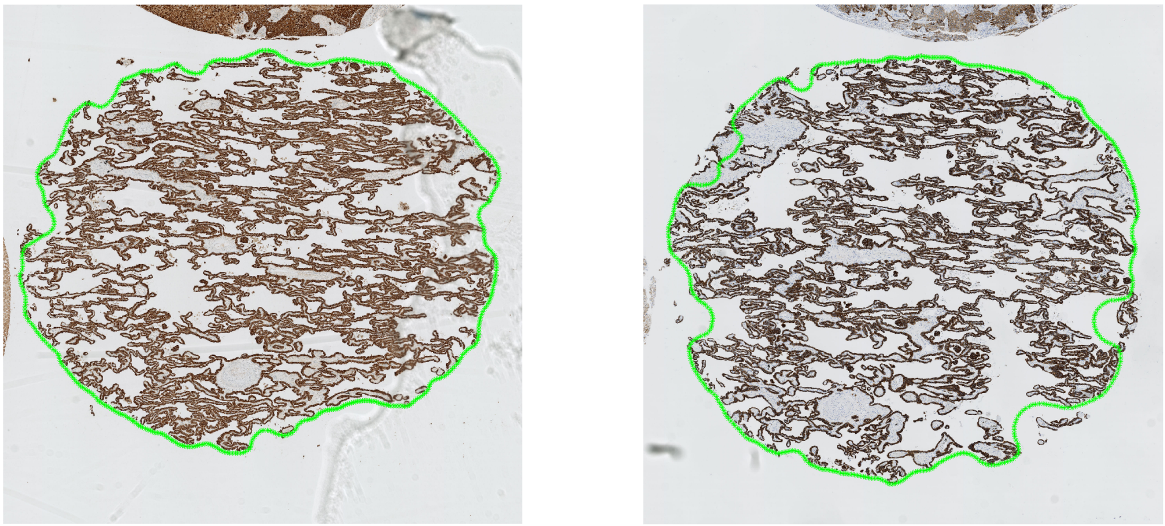
Figure 5.10: The final contour on core 4.



(a) Good example of the deformable model in sample 2.

(b) Sub-optimal example of the deformable model in sample 26.

Figure 5.11: The final contour on core 5.



(a) Good example of the deformable model in sample 8.

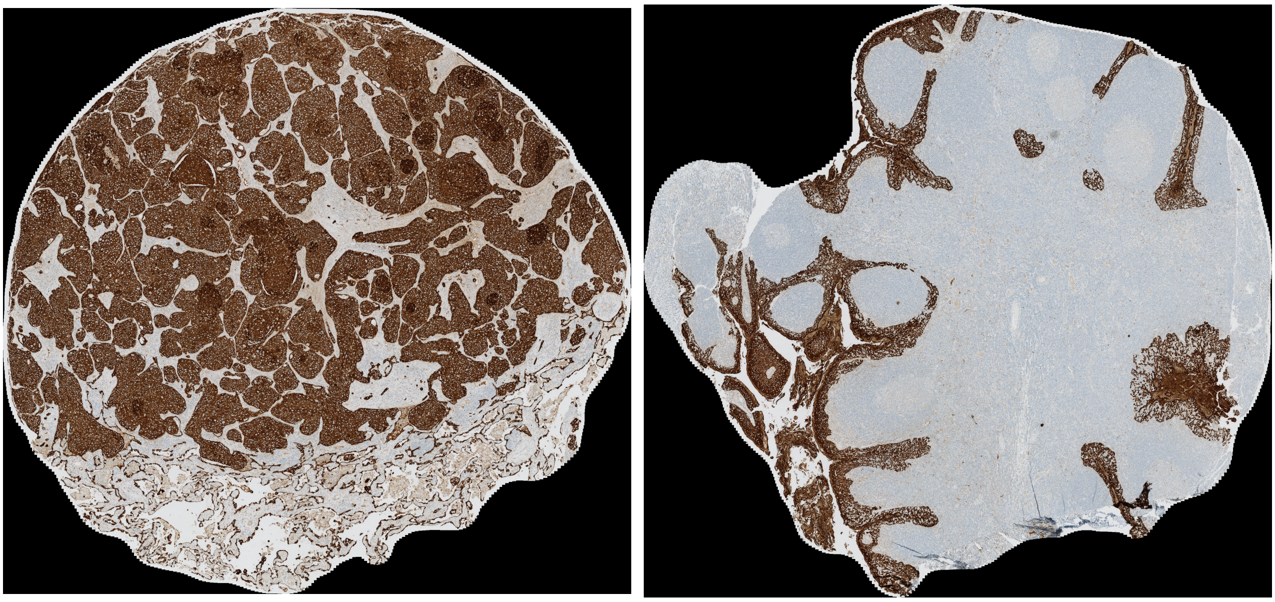
(b) Sub-optimal example of the deformable model in sample 18.

Figure 5.12: The final contour on core 6.

5.2.2 Extracted Image

The green contour illustrated in previous figures was used to create the mask used for extraction of the core. The resulting mask itself was a binary image indicating which indexes of the segmented image that should be used.

Using the extraction mask the large majority of the tissue core has been extracted. The extracted portion is shown in figure 5.13a for the regular core for sample 1. Figure 5.13b shows the extracted irregular core in sample 1.



(a) Figure showing the extracted portion of core 4 in sample 1. Core was of the regular type

(b) Figure showing the extracted portion of core 1 in sample 1. Core was of the irregular type

Figure 5.13: Extracted cores 1 & 4 after the snake model for sample 1. The resulting image from figure 5.6

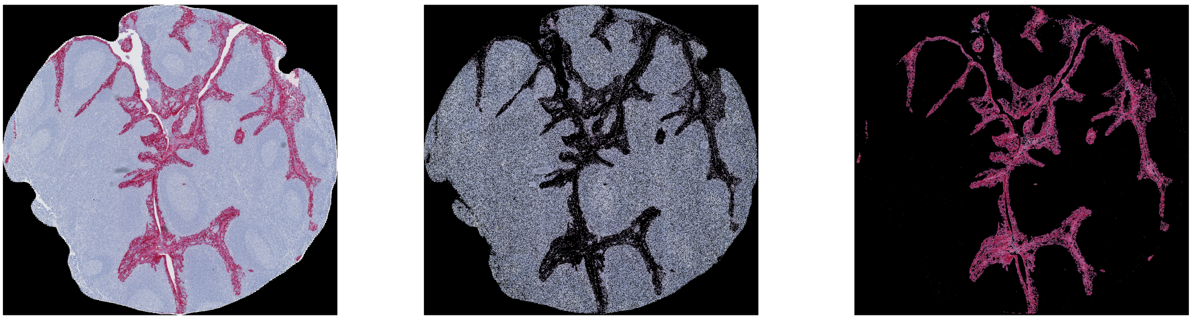
5.3 Core Classification

The results of the Core Classification consist of the features and the validation. The excluded features were based on the criteria mentioned in the method subsection 4.4.2. The Selected Features was the result of 10 repetitions. This approach should introduce randomness. The features and the validation have been combined to show the maximum and minimum result for each core. This is further expanded by confusion matrices and distribution of misclassifications that can be seen in subsection 5.3.4.

Four classifiers were used, one for each grade. These were then validated by comparing each classifier's results with the NordiQC assessments. This was done for the true positive and true negative being accurate classification, and the false positive and false negative being the misclassification.

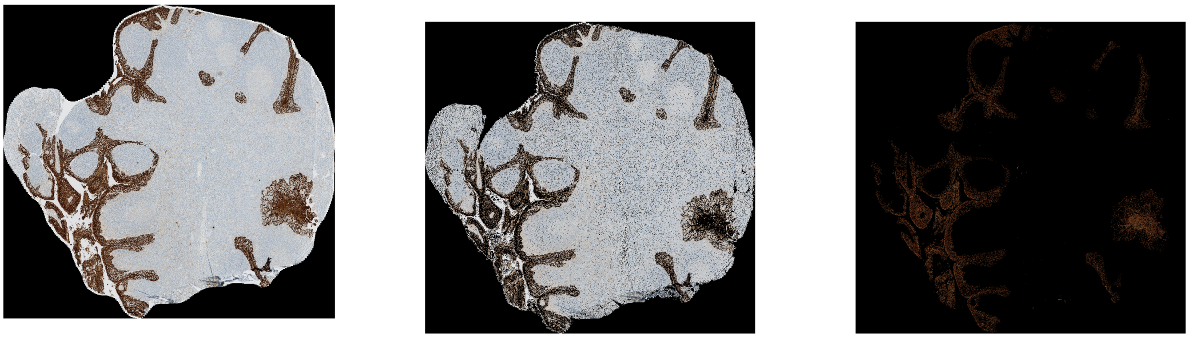
5.3.1 Stain Colour Separation

Illustrations of stain colour separation for the 3 different colour combinations are presented in figures 5.14 - 5.16.



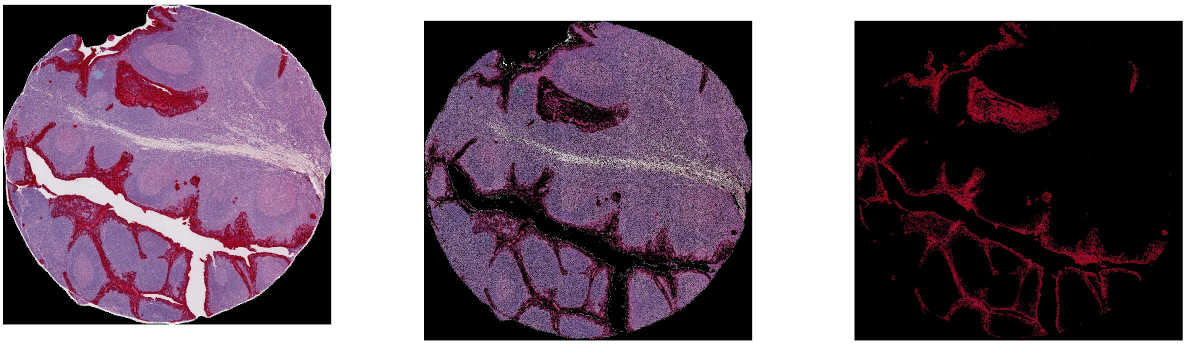
(a) Blue-Pink colour combination. (b) Dominant Colour. (c) Secondary Colour.

Figure 5.14: Illustration of colour separation for the Blue-Pink combination.



(a) Blue-Brown colour combination. (b) Dominant Colour. (c) Secondary Colour.

Figure 5.15: Illustration of colour separation for the Blue-Brown combination.



(a) Purple-Pink colour combination.

(b) Dominant Colour.

(c) Secondary Colour.

Figure 5.16: Illustration of colour separation for the Purple-Pink combination.

5.3.2 Features Excluded

For all cores, the *mean of the RGB Image* was excluded. The *Non-zero pixels in the Volume channel* was excluded from core 1, 2, 3 and 5. The *Non-zero pixels in the Red channel* was excluded for cores 2 and 5. The *Dominant / Secondary colour ratio*, *Mean of Secondary colour* and the *Standard deviation of Secondary colour* were excluded for core number 4 and 5. The *Non-zero pixels in the Green channel* and the *Red/Green Ratio* were also excluded for core 2. For core 4, the last 2 features that were excluded were the *Cyan/Magenta ratio* and the *Secondary colour ratio of the non-zero pixels*. The specific core and its excluded features are shown in table 5.1.

Table 5.1: Table contains a list of the features that were excluded for each core.

	Features Excluded
Core 1	• Mean of RGB image • Non-zero pixels in Volume channel
Core 2	• Mean of RGB image • Non-zero pixels in Volume channel • Non-zero pixels in Red channel • Non-zero pixels in Green channel • Red/Green Ratio
Core 3	• Mean of RGB image • Non-zero pixels in Volume channel
Core 4	• Mean of RGB image • Dominant / Secondary colour ratio • Mean of Secondary colour • Standard deviation of Secondary colour • Cyan/Magenta ratio • Secondary colour ratio of non-zero pixels
Core 5	• Mean of RGB image • Non-zero pixels in Volume channel • Non-zero pixels in Red channel • Dominant / Secondary colour ratio • Mean of Secondary colour • Standard deviation of Secondary colour
Core 6	• Mean of RGB image

5.3.3 Features Selected

After 10 repetitions, the features that had a selection frequency more than 50% were taken as the most important. These features can be seen in table 5.2.

Table 5.2: Selected features after 10 repetitions. The first percentage is the frequency of the feature being selected from SFS and the second the frequency of the feature being selected from BE. If both techniques gave the same percentage, it's stated only once.

	Features
Core 1	Standard Deviation of Blue channel (90%, 80%) Non-zero pixels in Yellow channel(60%) Dominant colour ratio of non-zero pixels (50%)
Core 2	Dominant / Secondary Ratio (90%) Standard Deviation of Volume Channel (70%)
Core 3	Mean of Blue channel (100%) 4th moment of Volume channel (60%, 0%)
Core 4	3rd moment of Saturation channel (70%) 4th moment of Green channel (60%)
Core 5	Standard Deviation of Blue channel (90%) Non-zero pixels in Cyan channel (80%)
Core 6	Standard Deviation of Green channel (80%) Secondary colour ratio of non-zero pixels (80%)

5.3.4 Validation

All cores had an average accuracy over 60%. Core 2 had the highest average accuracy with 85.74%, while core 6 had the lowest with 60.65%. Core 1 and 5 have both an accuracy higher than 75%, while core 3 and 4 have accuracies between 65-70%. The average accuracy, sensitivity and specificity for each core can be seen in table 5.3 in which the four classifiers are compared to each other.

Table 5.3: Table showing the average result of the cores regarding the accuracy, sensitivity and specificity.

	Core 1	**Core 2**	**Core 3**	**Core 4**	**Core 5**	**Core 6**
Accuracy	76.01%	85.74%	69.35%	66.51%	77.56%	60.65%
Sensitivity	75.97%	85.92%	69.67%	66.67%	77.72%	60.75%
Specificity	76.84%	85.57%	69.03%	66.35%	77.41%	60.55%

Apart for the general average of accuracy for each core, the accuracy for each grading of the cores was calculated. The highest accuracy for the optimal, borderline and poor grading was in core 2, while for the good grading was in core 5 with 79.5%. The minimum accuracy for all grading was in core 6, having 66.86% for the optimal, 56.18% accuracy for the good, 62.29% for the borderline and 57.28% for the poor grading. All the results can be seen in table 5.4.

Table 5.4: Average accuracy for the 4 different grading of each core.

	Core 1	**Core 2**	**Core 3**	**Core 4**	**Core 5**	**Core 6**
Accuracy for Optimal grading	73.58%	86.69%	71.99%	70.73%	78.28%	66.86%
Accuracy for Good grading	74.75%	79.25%	60.29%	57.46%	79.50%	56.18%
Accuracy for Borderline grading	78.29%	87.29%	69.29%	69.68%	71.14%	62.29%
Accuracy for Poor grading	77.41%	89.75%	75.82%	63.83%	81.34%	57.28%

For core 1, the maximum accuracy achieved was 78.95% using 5 features that can be seen in table 5.5. Both SFS and BE selected the same features.

Table 5.5: Table showing the maximum and minimum result for Core 1.

	Maximum Result	**Minimum Result**
Accuracy	78.95%	65.79%
Sensitivity	78.33%	65.83%
Specificity	78.88%	65.79%
Features Selected	• Standard Deviation of Volume channel • Standard Deviation of Blue channel • Entropy of Blue channel • Non-zero pixels in Yellow channel • 3rd moment of Black channel	• Standard Deviation of Blue channel

In table 5.6, the error matrix can be seen. For core 1, the biggest error percentage was in the optimal grading. In average, 2.6/9 of the optimal graded cores were classified as good. More results about the errors can be seen in table 5.7. Figure 5.17 shows two examples of a TP result from the borderline classifier (5.17a) and an FN from the poor classifier (5.17b).

Table 5.6: Confusion matrix for all grades in core 1.

Optimal Grading		
	Positive	**Negative**
True	68.89%	78.28%
False	31.11%	21.72%

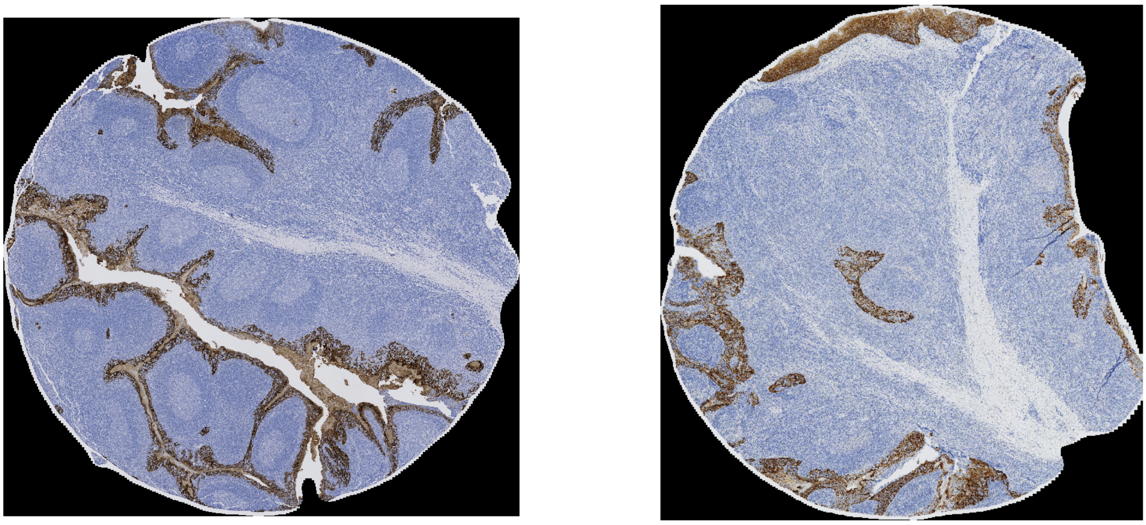
Good Grading		
	Positive	**Negative**
True	72.00%	77.50%
False	28.00%	22.50%

Borderline Grading		
	Positive	**Negative**
True	83.00%	73.57%
False	17.00%	26.43%

Poor Grading		
	Positive	**Negative**
True	80.00%	74.83%
False	20.00%	25.17%

Table 5.7: Table showing the distribution of misclassifications for core 1 between the NordiQC grade (rows) versus the classifier (columns).

grade / classification	**Optimal**	**Good**	**Borderline**	**Poor**	**n total = 38**
Optimal	6.2	2.6	0.0	0.2	**9**
Good	1.8	7.2	0.6	0.4	**10**
Borderline	0.1	0.1	8.3	1.5	**10**
Poor	0.0	0.6	1.2	7.2	**9**



(a) True positive of borderline classifier.

(b) False negative of poor classifier, graded as good.

Figure 5.17: TP and FN examples from classifiers for Core 1.

For core 2, the minimum accuracy was 81.58% with 3 features selected from SFS and 2 from BE. The minimum and maximum results achieved for core 2 can be seen in table 5.8.

Table 5.8: Table showing the maximum and minimum result for Core 2. * not selected from the Backward Elimination

	Maximum Result	**Minimum Result**
Accuracy	89.47%	81.58%
Sensitivity	89.72%	82%
Specificity	89.50%	81.59%
Features Selected	• Mean of Saturation channel • Standard Deviation of Volume channel • Dominant / Secondary Ratio	• Hue/Saturation Ratio* • Standard Deviation of Volume channel • Dominant / Secondary Ratio

Core 2, had an average accuracy for poor grading of 97.78% while the lowest was in good grading with 66%. All the results can be seen in table 5.9 and more detailed in table 5.10. Figure 5.18 shows two examples of a TP result from the poor classifier (5.18a) and an FN from the good classifier (5.18b).

Table 5.9: Confusion matrix for all grades in core 2.

Optimal Grading		
	Positive	**Negative**
True	88.89%	84.48%
False	11.11%	15.52%

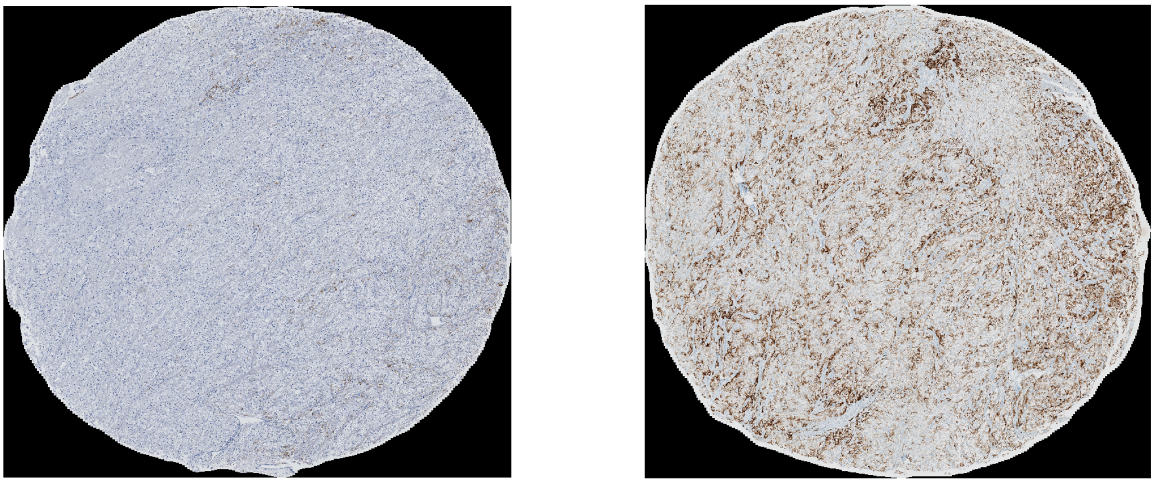
Good Grading		
	Positive	**Negative**
True	66.00%	92.50%
False	34.00%	7.50%

Borderline Grading		
	Positive	**Negative**
True	91.00%	83.57%
False	9.00%	16.43%

Poor Grading		
	Positive	**Negative**
True	97.78%	81.72%
False	2.22%	18.28%

Table 5.10: Table showing the distribution of misclassifications for core 2 between the NordiQC grade (rows) versus the classifier (columns).

grade / classification	**Optimal**	**Good**	**Borderline**	**Poor**	**n total = 38**
Optimal	8.0	1.0	0.0	0.0	**9**
Good	3.3	6.6	0.1	0.0	**10**
Borderline	0.0	0.8	9.1	0.1	**10**
Poor	0.0	0.0	0.2	8.8	**9**



(a) True positive of poor classifier.

(b) False negative of good classifier, graded as optimal.

Figure 5.18: TP and FN examples from classifiers for Core 2.

For core 3, as it can be seen in table 5.11, the minimum and the maximum results had a difference of about 5% in accuracy.

Table 5.11: Table showing the maximum and minimum result for Core 3. * not selected from the Backward Elimination

	Maximum Result	**Minimum Result**
Accuracy	73.68%	68.42%
Sensitivity	74.14%	69.17%
Specificity	73.74%	68.50%
Features Selected	• 4th moment of Volume channel • Mean of Blue channel • 3rd moment of Blue channel* • 4th moment of Green channel	• 4th moment of Volume channel* • Mean of Blue channel

For all gradings, in core 3, the true negative's percentage was above 60% as table 5.12 shows, while the most errors were in the good grading with an average of 4.8/10 identified correctly (table 5.13). In figure 5.19 two examples of a TP result from the poor classifier (5.19a) and an FN from the good classifier (5.19b) can be seen.

Table 5.12: Confusion matrix for all grades in core 3.

Optimal Grading		
	Positive	**Negative**
True	77.78%	66.21%
False	22.22%	33.79%

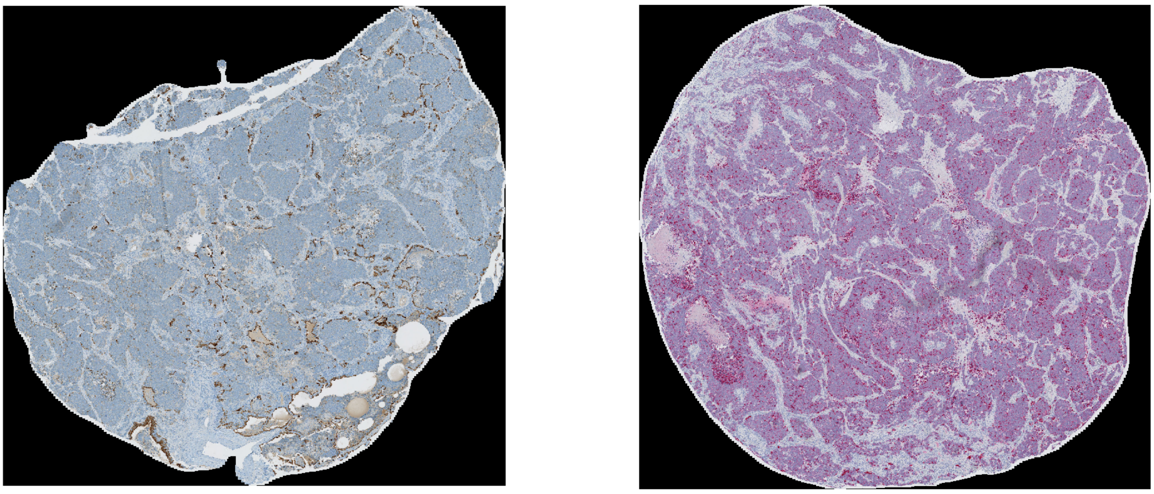
Good Grading		
	Positive	**Negative**
True	42.00%	78.57%
False	58.00%	21.43%

Borderline Grading		
	Positive	**Negative**
True	70.00%	68.57%
False	30.00%	31.43%

Poor Grading		
	Positive	**Negative**
True	88.89%	62.76%
False	11.11%	37.24%

Table 5.13: Table showing the distribution of misclassifications for core 3 between the NordiQC grade (rows) versus the classifier (columns).

grade / classification	**Optimal**	**Good**	**Borderline**	**Poor**	**n total = 38**
Optimal	7.0	2.0	0.0	0.0	**9**
Good	2.3	4.8	1.9	1.0	**10**
Borderline	1.9	0.1	7.0	1.0	**10**
Poor	0.0	0.0	1.0	8.0	**9**



(a) True positive of poor classifier.

(b) False negative of good classifier, graded as poor.

Figure 5.19: TP and FN examples from classifiers for Core 3.

Table 5.14, shows that for the maximum results of core 4, three features were selected from SFS and BW giving an accuracy of 71.05%.

Table 5.14: Table showing the maximum and minimum result for Core 4.

	Maximum Result	**Minimum Result**
Accuracy	71.05%	60.53%
Sensitivity	71.39%	61.39%
Specificity	71.09%	60.62%
Features Selected	• Standard Deviation of Saturation channel • Mean of Blue channel • 4th moment of Green channel	• Standard Deviation of Saturation channel

Core 4, had 41% average accuracy for good grading as table 5.15 shows. In table 5.16 can be seen that in average, 4.1/10. Figure 5.20 shows two examples of a TP result from the optimal classifier (5.20a) and an FN from the poor classifier (5.20b).

Table 5.15: Confusion matrix for all grades in core 4.

Optimal Grading		
	Positive	Negative
True	81.11%	60.34%
False	18.89%	39.66%

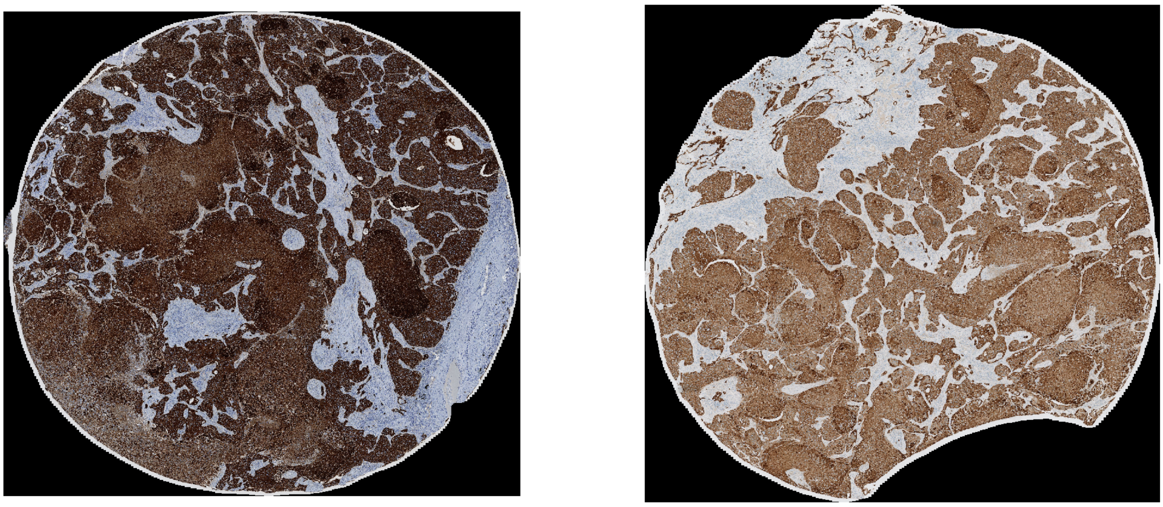
Good Grading		
	Positive	Negative
True	41.00%	73.93%
False	59.00%	26.07%

Borderline Grading		
	Positive	Negative
True	79.00%	60.36%
False	21.00%	39.64%

Poor Grading		
	Positive	Negative
True	61.11%	66.55%
False	38.89%	33.45%

Table 5.16: Table showing the distribution of misclassifications for core 4 between the NordiQC grade (rows) versus the classifier (columns).

grade / classification	**Optimal**	**Good**	**Borderline**	**Poor**	**n total = 38**
Optimal	7.3	1.5	0.2	0	**9**
Good	3.1	4.1	1.9	0.9	**10**
Borderline	0.1	1.2	7.9	0.8	**10**
Poor	0.9	0.8	1.8	5.5	**9**



(a) True positive of optimal classifier.

(b) False negative of poor classifier, graded as good.

Figure 5.20: TP and FN examples from classifiers for Core 4.

For core 5, two features selected both from SFS and BE gave the best result with 78.95% accuracy (table 5.17.

Table 5.17: Table showing the maximum and minimum result for Core 5.

	Maximum Result	**Minimum Result**
Accuracy	78.95%	71.05%
Sensitivity	79.17%	71.94%
Specificity	78.97%	71.15%
Features Selected	• Standard Deviation of Blue channel • Non-zero pixels in Cyan channel	• Entropy of Yellow channel

Accuracy in all 4 different gradings was above 50%, as table 5.18 shows. Especially poor grading, table 5.19 presents that the average error was 1 out of 9 which was classified as borderline grading. In figure 5.21 are presented two examples of a TP result from the good classifier (5.21a) and an FN from the borderline classifier (5.21b).

Table 5.18: Confusion matrix for all grades in core 5.

Optimal Grading		
	Positive	**Negative**
True	80.00%	76.55%
False	20.00%	23.45%

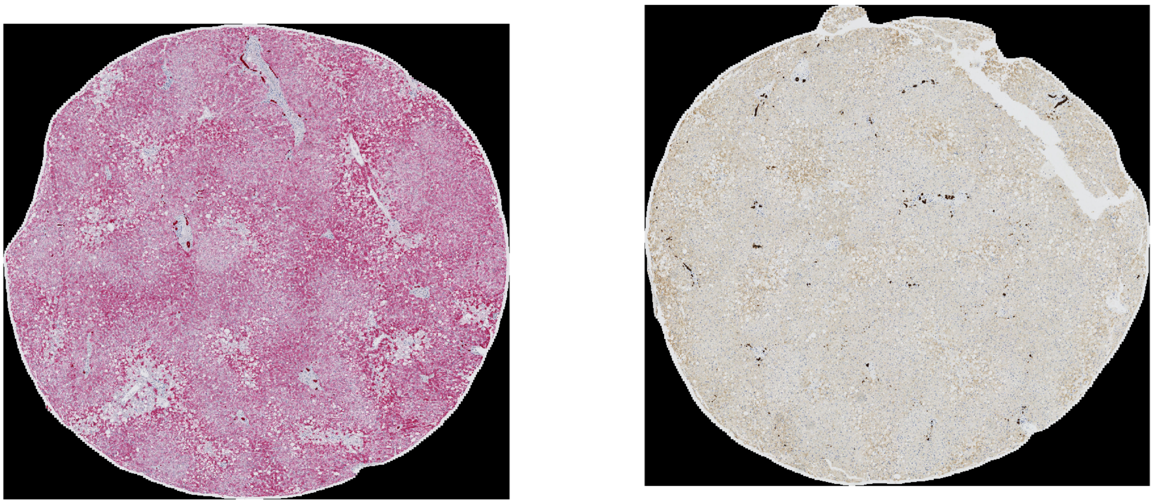
Good Grading		
	Positive	**Negative**
True	84.00%	75.00%
False	16.00%	25.00%

Borderline Grading		
	Positive	**Negative**
True	58.00%	84.29%
False	42.00%	15.71%

Poor Grading		
	Positive	**Negative**
True	88.89%	73.79%
False	11.11%	26.21%

Table 5.19: Table showing the distribution of misclassifications for core 5 between the NordiQC grade (rows) versus the classifier (columns).

grade / classification	**Optimal**	**Good**	**Borderline**	**Poor**	**n total = 38**
Optimal	7.2	1.8	0.0	0.0	**9**
Good	1.4	8.4	0.0	0.2	**10**
Borderline	0.0	1.4	5.8	2.8	**10**
Poor	0.0	0.0	1.0	8.0	**9**



(a) True positive of good classifier.

(b) False negative of borderline classifier, graded as poor.

Figure 5.21: TP and FN examples from classifiers for Core 5.

Core 6 had as maximum result and accuracy less than 70% and a minimum a bit more than 50% (table 5.20.

Table 5.20: Table showing the maximum and minimum result for Core 6.

	Maximum Result	**Minimum Result**
Accuracy	65.79%	52.63%
Sensitivity	66.11%	52.50%
Specificity	65.83%	52.62%
Features Selected	• Standard Deviation of Green channel • Secondary colour ratio of non-zero pixels	• Mean of Volume channel

The best accuracy was for the optimal grading with 78.89%, as table 5.21 shows. More information about the error can be seen in table 5.22. Figure 5.22 shows two examples of a TP result from the optimal classifier (5.22a) and an FN from the poor classifier (5.22b).

Table 5.21: Confusion matrix for all grades in core 6.

Optimal Grading		
	Positive	**Negative**
True	78.89%	54.83%
False	21.11%	45.17%

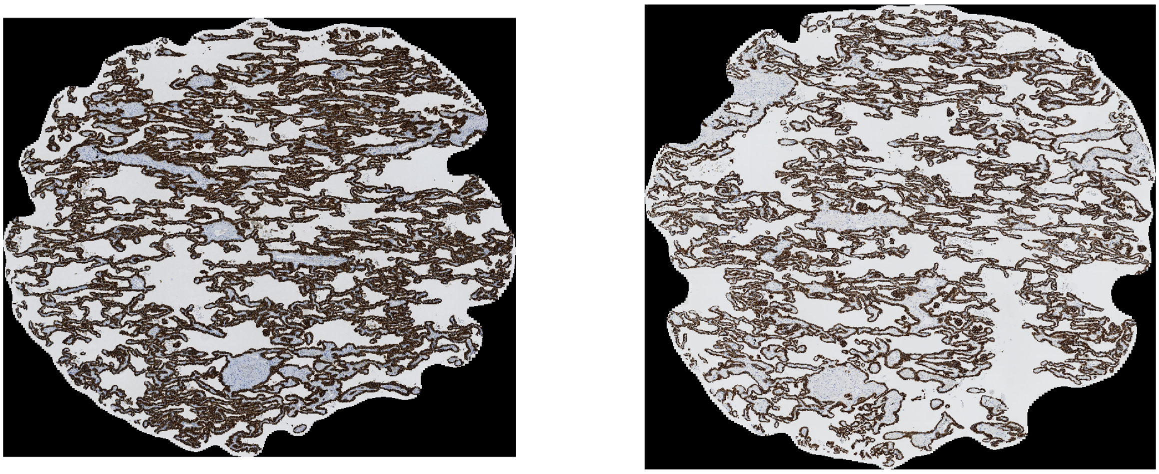
Good Grading		
	Positive	**Negative**
True	47.00%	65.36%
False	53.00%	34.64%

Borderline Grading		
	Positive	**Negative**
True	66.00%	58.57%
False	34.00%	41.43%

Poor Grading		
	Positive	**Negative**
True	51.11%	63.45%
False	48.89%	36.55%

Table 5.22: Table showing the distribution of misclassifications for core 6 between the NordiQC grade (rows) versus the classifier (columns).

grade / classification	**Optimal**	**Good**	**Borderline**	**Poor**	**n total = 38**
Optimal	7.1	0.7	0.7	0.5	**9**
Good	1.2	4.7	3.4	0.7	**10**
Borderline	0.9	1.8	6.6	0.7	**10**
Poor	1.2	1.3	1.9	4.6	**9**



(a) True positive of optimal classifier.

(b) False negative of poor classifier, graded as optimal.

Figure 5.22: TP and FN examples from classifiers for Core 6.

Chapter 6

Discussion

This project is the result of a master thesis which seeks to uncover if image analysis combined with pattern recognition is able to segment and classify immunohistological samples with comparison to the assessment done by NordiQC. Presently the assessment is done in consensus between 5 clinicians but as revealed in the problem analysis, variability is present both at a laboratory and observer level. And these assessments are still qualitative even though guidelines exists. Therefore this project investigated the possibility of performing quantitative measures with reference to the assessments by NordiQC.

6.1 Core Classification

The features used for classification were not consistent throughout the six cores. This is shown in the results table 5.1 where the excluded features are presented for each core. These were excluded either because there was little/no variance between the cores or exclusion due to the covariance matrix regarding each class/grade. The feature that proved most important was the Blue channel as it was included as a feature in Core 1, 3 & 5. Here it was selected as a features for each core in more than 80% of the repetitions.

Choosing to primarily rely on intensity-based features might have limited the variance in the features. However this is unsure since the initial impression of the data only revealed differences with regards to the colour and intensity between the NordiQC assessments. Also, by combining 3 colour spaces with comparable features, added to the ability of describing the individual images in a quantitative manner. The inclusion of staining specific features namely the Dominant and Secondary colour added 7 additional features. However the Dominant and Secondary colour only appeared as included features twice, thus making them less useful.

Increasing the amount of features might not always produce more accurate results as this also may lead to the curse of dimensionality. However compared to the NordiQC Pan-CK assessment criteria it could increase the results by adding spatial information into the feature space.

The validation of the classifier shows that it is possible to accurately classify the 38 samples. This is shown in table 5.3 for the average accuracy, sensitivity and specificity. The results range from 60% for core 6 up to 85% for core 2. When comparing this with the elaborated table 5.4 that also shows the grade, it shows that it is possible to categorise the 38 samples using a quantitative classifier. Table 5.7, 5.10, 5.13, 5.16, 5.19 & 5.22 all show the distribution of misclassifications and from this it is clear that the result rarely shifts more than one grade if the classifier was wrong compared to the NordiQC assessment.

For this study 5-fold validation was used. According to Kohavi et al. [1995] a 10-fold or 20-fold cross validation would give better results and be less biased. Due to the limited amount of data available this would not prove a viable option, as 10-fold would create 3-4 samples per group and 20-fold would create only 2 samples per group [Kohavi et al., 1995]. However considering the amount of data, a 5-fold would give lower variance when compared to 10- or 20-fold.

It is easily noticed that core 6 had the lowest accuracy in average. This may be caused by the large amount of holes within the tissue of core 6. In order to avoid this, one could try implementing deformable models that are specialized towards a single core type like core 6.

6.2 Core Segmentation

The results in 5.7-5.12 show how the deformable model performed for each different core regarding a good and a bad segmentation. These results can in general be considered acceptable. In some of the segmentations the small outliers have not been included, which reduces the amount of information by very little. On the other hand, this has also omitted small portions of core which often was caused by holes in the periphery of the tissue sample. In some other cases the deformable model did not include tissue fragments that were separated from the main tissue. This leads to an issue where one must determine if the lack of including core tissue or fragments can outweigh the exclusion of what can be considered background or noise.

The use of a single deformable model which was used to segment all six different cores gave acceptable results. However these cores were not completely identical and differences did occur between them, which also appeared in some segmentations. Another approach would be to have a deformable model that would be specifically tailored to each core. This could lead to a segmentation that could consider each core's special characteristics such as sharper edges, separated fragments, or even a shrinking snake instead. Based on the visual inspection of the cores their shape was not considered complex and this led to the choice of a parametric deformable model. As time can be a precious resource it was chosen to search for existing deformable models for Matlab and modify this to our purpose. A handful of deformable models were found, yet only one was in such a state that it allowed seamless integration. The integrated model was made by [Dahl and Dahl, 2016] and features a dictionary snake which is different from the planned parametric model. It is possible that a parametric snake could yield different segmentation results. However this has not been evaluated due to time limitations.

The segmentation performance of the deformable model is limited by its ability to segment the core from the background in multiple different tissue cores. Since there is a large structural variability between the cores, the snake model cannot be finely tuned for a specific core type. This places a constraint on the overall performance, as the snake has to segment multiple different core types. This limitation becomes clear as fragments of tissue were not included in the segmentation. However, it is unsure how large of an effect the inclusion of this would have on the classification results. But by utilising a deformable model adjusted for each specific core it would become possible to achieved a more close fitting contour as opposed to using only one. This could be reasonable when considering the cavitis within core 6. These would require different weighting parameters than e.g. core 5.

6.3 Core Detection

The core detection was used to identify the position of each tissue core in every sample. To describe the position, the cores center point and radius were used. As seen in the figure 5.4 this was performed to a very satisfying degree even when some of the cores connected with each other, which could make the distinction between core centre points difficult as these cores could be considered as one.

The performance of the core detection could be attributed to its use of a background specific intensity measure which ensured that the background was used as the threshold when converting grayscale into binary. This was performed to each sample individually which could take the colour intensity into account as this varied between samples and their NordiQC score. This approach lent itself very useful as 8-connectivity was used to find the cluster throughout the image and the clusters present in the binary image was sensitive to the threshold used when converting from grayscale to binary. One reason why the Core detection proved useful was that it specified the location where the Core Segmentation should start which could vary between samples. As a byproduct it also reduced the computation required as only a portion of the image was used.

As discovered early in the process the Core Detection had one instance where it was completely unable to identify each of the cores. This happened when the Core Detection was performed on the excluded sample. Because of its high amount of noise due to an air bubble the Core Detection was unable to identify the individual cores and the results hereof were far from satisfying. However after consulting with a NordiQC clinician it was agreed that such a sample should rarely or even at all represent the result from a laboratory given its very poor quality which justified its exclusion.

The isolation of the oesophagus from the cores was done to a satisfying degree as more than half of the oesophagus tissue was removed throughout all the samples. By doing so, it enabled a more precise detection of the tissue cores. This would prove important as the oesophagus' irregular shape and size as a starting point would require a different approach if it should contribute to the classification results at the same level as the tissue cores.

The 8-connectivity was also used to find the leftmost object which should match the oesophagus. However this approach is based on some assumptions on the position of the oesophagus in all the samples. Also the removal could be considered rough as border between oesophagus and tissue cores wasn't aligned to fit between them but rather was a vertical border. This would often lead to the inclusion of small portions of oesophagus but this did not interfere with the core detection.

One limitation would be the orientation of the sample as it was assumed that the oesophagus would always be the leftmost object. Another approach could be to detect a line that would allow the rotation of the image. Then a vertical border could be placed more precisely. Identifying the correct line would be challenging though, as the background had plenty of hidden lines that were visible only after turning the image into the grayscale space and using histogram equalisation. Another limitation of excluding the oesophagus is that the classification is based solely on the six tissue cores and it cannot be ruled out that these results could become better if this image information was added.

Chapter 7

Conclusion

This project sought to investigate the possibility of classifying IHC samples using image features. To achieve this the following four study aims were defined. These are the following:

- To develop an algorithm capable of analysing the NordiQC samples separately
- To investigate the possibility of using image features and classification to describe NordiQC grades
- To validate the results compared to the NordiQC grades
- To explore the plausibility of a CAD system for the NordiQC collaboration

From the results it was shown that an algorithm could be developed which allowed for the separate analysis of NordiQC samples. This was achieved through the Core Detection and the Core Segmentation. Using two groups of features to describe the NordiQC grades proved possible. The two groups were intensity-based features and staining colour specific features, which made up 91 features in total. These features were validated against the NordiQC grades. The end average accuracy for the cores was between 60-85% when considering all possible NordiQC grades. These results did not differ significantly for the classifier of each grade within the cores. From these results it is therefore concluded that CAD systems are plausible.

The provided data set consisted of 39 samples, one of which was excluded. The 38 samples were grouped into four different grades, thereby only leaving 9-10 samples per group. It should therefore be noted that a larger amount of data should be used in future studies to supplement this study and its results. Furthermore including features able to differentiate cellular components within each sample should be considered as to comply with NordiQC assessment criteria. Adding more samples and including additional feature may lead to improved results.

Literature

Arthur, B.: n.d., The difference between quality assurance and quality control.
URL: *https://www.dialog.com.au/open-dialog/the-difference-between-quality-assurance-and-quality-control/*

Baratloo, A., Hosseini, M., Negida, A. and Ashal, G. E.: 2015, Part 1: Simple definition and calculation of accuracy, sensitivity and specificity, *Emergency* **2**(3), 48–49.

Bishop, C. M.: 2006, Pattern recognition, *Machine Learning* **128**, 1–58.

Bitesize Bio: n.d.
URL: *http://bitesizebio.s3.amazonaws.com/wp-content/uploads/2011/10/Oxidisation.jpg*

Chiannilkulchai, N., Driouich, Z., Benoit, J., Parodi, A. and Couvreur, P.: 1989, Doxorubicin-loaded nanoparticles: increased efficiency in murine hepatic metastases, *Selective cancer therapeutics* **5**(1), 1–11.

Dahl, A. B. and Dahl, V. A.: 2014, Dictionary snakes, *2014 22nd International Conference on Pattern Recognition*, pp. 142–147.

Dahl, V. A. and Dahl, A. B.: 2016, Deformable models for texture segmentation v1.0. Last visited 30th May 2017.
URL: *https://se.mathworks.com/matlabcentral/fileexchange/56445-deformable-models-for-texture-segmentation?focused=6159306&tab=function*

Demir, C. and Yener, B.: 2004, Automated cancer diagnosis based on histopathological images: a systematic survey, *Technical report*, RENSSELAER POLYTECHNIC INSTITUTE, DEPARTMENT OF COMPUTER SCIENCE.

Dougherty, G.: 2012, *Pattern recognition and classification: an introduction*, Springer Science & Business Media.

Doyle, S., Hwang, M., Shah, K., Madabhushi, A., Feldman, M. and Tomaszeweski, J.: 2007, AUTOMATED GRADING OF PROSTATE CANCER USING ARCHITECTURAL AND TEXTURAL IMAGE FEATURES.

Efron, B.: 1986, How biased is the apparent error rate of a prediction rule?, *Journal of the American statistical Association* **81**(394), 461–470.

Eisen, R. N.: 2008, Quality management in immunohistochemistry, *DIAGNOSTIC HISTOPATHOLOGY* **14**, 299–307.

Fitzgibbons, P. L., Bradley, L. A., Fatheree, L. A., Alsabeh, R., Fulton, R. S., Goldsmith, J. D., Haas, T. S., Karabakhtsian, R. G., Loykasek, P. A., Marolt, M. J., Shen, S. S., Smith, A. T. and Swanson, P. E.: 2014, Principles of analytic validation of immunohistochemical assays, *Archives of Pathology & Laboratory Medicine* **138**.

Franklin, J.: 2005, The elements of statistical learning: data mining, inference and prediction, *The Mathematical Intelligencer* **27**(2), 83–85.
URL: *http://dx.doi.org/10.1007/BF02985802*

Guyon, I. and Elisseeff, A.: 2003, An introduction to variable and feature selection, *Journal of machine learning research* **3**(Mar), 1157–1182.

Guyon, I., Gunn, S., Nikravesh, M. and Zadeh, L. A.: 2008, *Feature extraction: foundations and applications*, Vol. 207, Springer.

Janecek, A., Gansterer, W. N., Demel, M. and Ecker, G.: 2008, On the relationship between feature selection and classification accuracy., *FSDM* **4**, 90–105.

Jørgensen, A. S., Østergaard, L. R. and Røge, R.: 2017, Quality control of immunohistochemically stained slides for diagnosis of cancer.

Kim, M.: 2010, Statistical classification.

Kirkegaard, T., Edwards, J., Tovey, S., McGlynn, L. M., Krishna, S. N., Mukherjee, R., Tam, L., Munro, A. F., Dunne, B. and Bartlett, J. M. S.: 2006, Observer variation in immunohistochemical analysis of protein expression, time for a change?, *Histopathology* pp. 787–794.

Kohavi, R. et al.: 1995, A study of cross-validation and bootstrap for accuracy estimation and model selection, *Ijcai*, Vol. 14, Stanford, CA, pp. 1137–1145.

Kothari, S., Phan, J. H., Stokes, T. H. and Wang, M. D.: 2013, Pathology imaging informatics for quantitative analysis of whole-slide images, *Journal of the American Medical Informatics Association* **20**(6), 1099–1108.

Lin, F. and Chen, Z.: 2014, Standardization of diagnostic immunohistochemistry, *Archives of Pathology & Laboratory Medicine* **138**.

Marcano-Cedeño, A., Quintanilla-Domínguez, J., Cortina-Januchs, M. and Andina, D.: 2010, Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network, *IECON 2010-36th Annual Conference on IEEE Industrial Electronics Society*, IEEE, pp. 2845–2850.

Maxwell, P. and McCluggage, W. G.: 2000, Audit and internal quality control in immunohistochemistry, *Journal of Clinical Pathology* **53**(12), 929–932.

McInerney, T. and Terzopoulos, D.: 1996, Deformable models in medical image analysis: a survey, *Medical Image Analysis* **1**(2), 91–108.

Mengel, M., von Wasielewski, R., Wiese, B., Rüdiger, T., Müller-Hermelink, H. K. and Kreipe, H.: 2002, Inter-laboratory and inter-observer reproducibility of immunohistochemical assessment of the ki-67 labelling index in a large multi-centre trial, *Journal of Pathology* pp. 292–299.

Naik, S., Doyle, S., Agner, S., Madabhushi, A., Feldman, M. and Tomaszewski, J.: 2008, Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology, *IEEE International Symposium on Biomedical Imaging* .
URL: *https://www.researchgate.net/publication/4341682*

Neumann, J., Schnörr, C. and Steidl, G.: 2005, Combined svm-based feature selection and classification, *Machine learning* **61**(1), 129–150.

NordiQC: 2016, Assessment Run 47 2016 Pan Cytokeratin (CK-PAN).

O'Leary, T. J.: 2001, Standardization in immunohistochemistry, *Applied Immunohistochemistry & Molecular Morphology* .

Oliver, C. and Jamur, M. C.: 2009, *Immunocytochemical Methods and Protocols*, third edn, Humana Press.
URL: *http://www.ebook.de/de/product/8579740/immunocytochemical_methods_and_protocols.html*

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P. et al.: 2001, Multiclass cancer diagnosis using tumor gene expression signatures, *Proceedings of the National Academy of Sciences* **9**8(26), 15149–15154.

Ranefall, P., Egevad, L., Nordin, B. and Bengtsson, E.: 1997, A new method for segmentation of colour images applied to immunohistochemically stained cell nuclei, *Analytical Cellular Pathology* pp. 145–156.

Reiner-Concin, A.: 2008, External quality assurance in immunohistochemistry - is it the solution to a complex problem?, *Breast Care* .

Rifkin, R. and Klautau, A.: 2004, In defense of one-vs-all classification, *Journal of machine learning research* **5**(Jan), 101–141.

Saeys, Y., Inza, I. and Larrañaga, P.: 2007, A review of feature selection techniques in bioinformatics, *bioinformatics* **2**3(19), 2507–2517.

Serotec: n.d.
URL: *https://www.serotec.co.uk/images/types-of-antibodies.jpg*

Sino Biological, Inc.: n.d.a, DAB immunohistochemistry.
URL: *http://www.immunohistochemistry.us/what-is-immunohistochemistry/DAB-Immunohistochemistry.html*

Sino Biological, Inc.: n.d.b, Immunohistochemistry (IHC) Principle.
URL: *http://www.immunohistochemistry.us/IHC-principle.html*

Sonka, M. and Fitzpatrick, J. M.: 2000, *Handbook of Medical Imaging*, Vol. 2, SPIE–The International Society for Optical Engineering.

Sorenson, S. C., Asch, B. B., Connolly, J. L., Burstein, N. A. and Asch, H. L.: 1987, Structural distinctions among human breast epithelial cells revealed by the monclonal antikeratin antibodies AE1 and AE3, *The Journal of Pathology* **1**53(2), 151–162.
URL: *http://dx.doi.org/10.1002/path.1711530208*

Stańczyk, U. and Jain, L. C.: 2015, *Feature selection for data and pattern recognition*, Springer.

Taylor, C. R.: 2000, The Total Test Approach to Standardization of Immunohistochemistry, *Archives of Pathology & Laboratory Medicine* **1**24.

Torlakovic, E. E., Francis, G., Garratt, J., Gilks, B., Hyjek, E., Ibrahim, M., Miller, R., Nielsen, S., Petcu, E. B., Swanson, P. E., Taylor, C. R. and Vyberg, M.: 2014, Standardization of negative controls in diagnostic immunohistochemistry: Recommendations from the international ad hoc expert panel, *Applied Immunohistochemistry & Molecular Morphology* pp. 241–252.

Torlakovic, E. E., Nielsen, S., Vyberg, M. and Taylor, C. R.: 2015, Getting controls under control: the time is now for immunohistochemistry, *Journal of Clinical Pathology* **6**8(11), 879–882.

University of Leeds: n.d., What is H&E?
URL: *http://histology.leeds.ac.uk/what-is-histology/H_and_E.php*

von Wasielewski, R., Mengel, M., Wiese, B., Rüdiger, T., Müller-Hermelink, H. K. and Kreipe, H.: 2002, Tissue array technology for testing interlaboratory and interobserver reproducibility of immunohistochemical estrogen receptor analysis in a large multicenter trial, *American Journal of Clinical Pathology* .

Vyberg, M. and Nielsen, S.: 2016, Proficiency testing in immunohistochemistry - experiences from Nordic Immunohistochemical Quality Control (NordiQC), *Virchows Archiv* **4**68, 19–29.

Vyberg, M., Torlakovic, E., Seidal, T., Risberg, B., Helin, H. and Nielsen, S.: 2005, Nordic Immunohistochemical Quality Control, *Croatian Medical Journal* **46**, 368–371.

Walker, R.: 2006, Quantification of immunohistochemistry - issues concerning methods, utility and semiquantitative assessment I, *Histopathology* **49**(4), 406–410.

Wolpert, D. H.: 1992, Stacked generalization, *Neural networks* **5**(2), 241–259.

Appendix A

Flowchart - Step 1

A.1 Overview

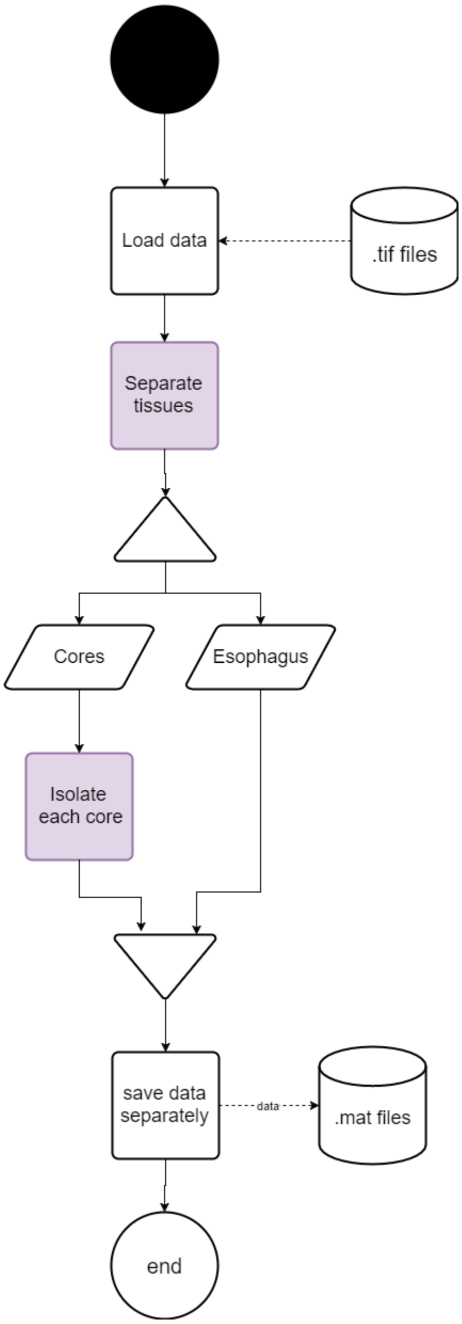


Figure A.1: Flowchart of the general preprocessing. The Separate tissues is expanded in INITIAL SEPARATION, and the Isolate each core is expanded in INITIAL SEGMENTATION

A.2 Initial separation

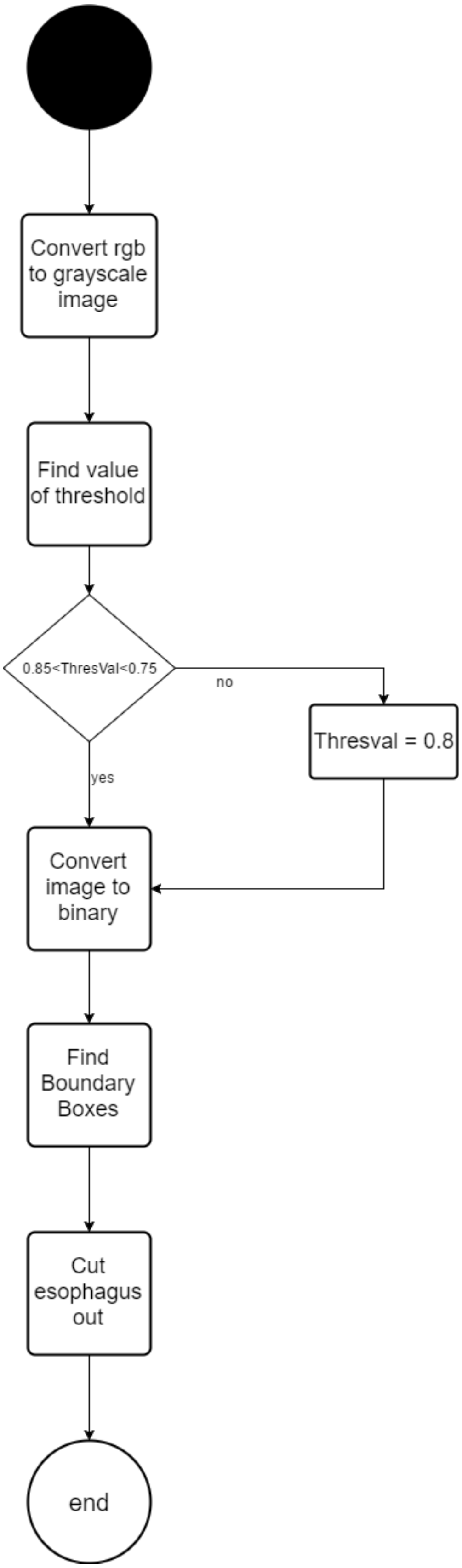


Figure A.2: Flowchart of the initial separation of oesophagus and the cores

A.3 Initial segmentation

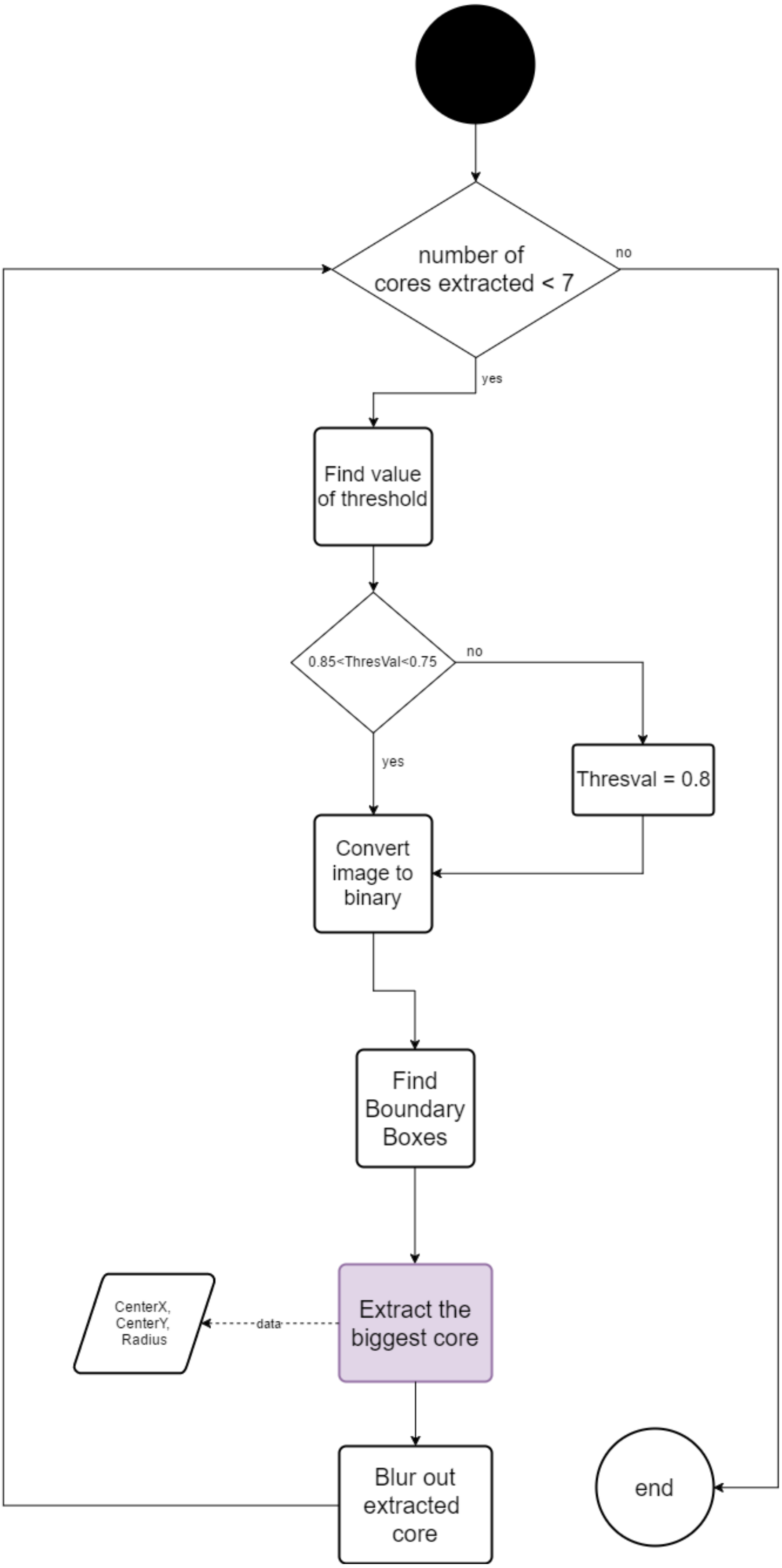


Figure A.3: Flowchart of the isolation of each core

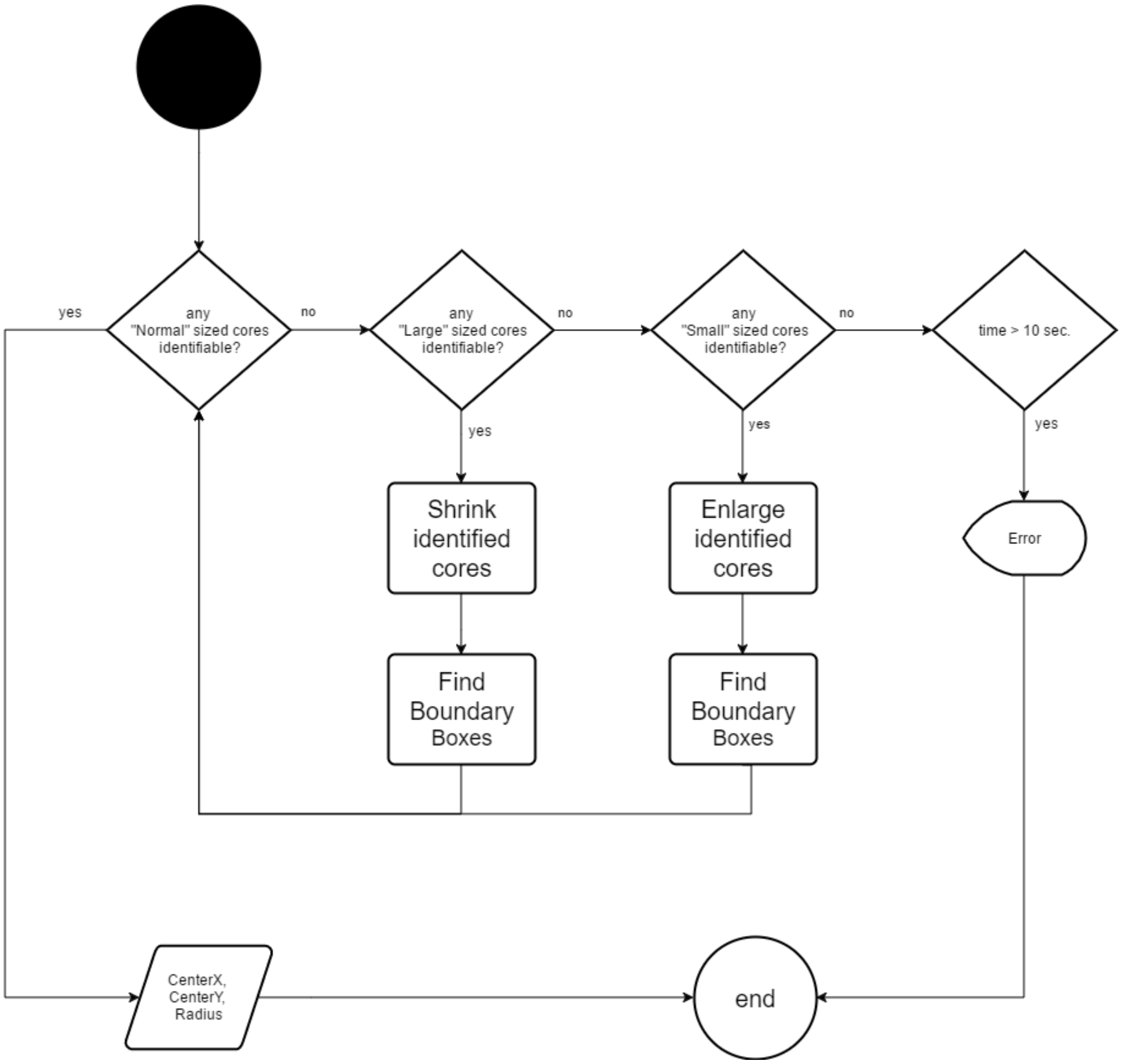


Figure A.4: Flowchart of the extraction of each core.

Appendix B

Flowchart - Step 2

B.1 Overview

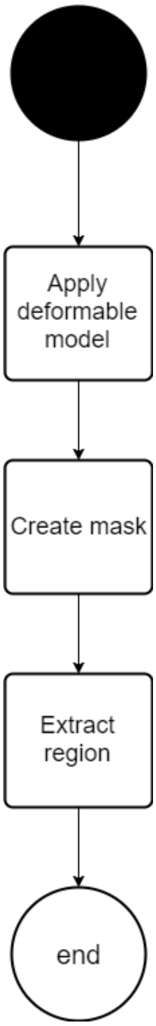


Figure B.1