Digitizing pathology lab workflows using image processing and OCR

Author: Markus Hatlem
Supervisor: Fazle Rabbi
Co-supervisor: Patrick Stünkel
Co-supervisor: Friedemann Leh

Department of Information Science and Media Studies
University of Bergen

May 31, 2023

Scientific environment

This study is carried out at the faculty of Social Sciences, University of Bergen. The work is supported by Helse Vest through the pathology lab at Haukeland university hospital.

Acknowledgements

A big thank you to my supervisor Fazle Rabbi for making this thesis possible. I appreciate all the time and effort you put into guidance for this project.

I would also like to thank my co-supervisors Patrick Stünkel and Friedemann Leh for their time and feedback. I also appreciate their assistance with accessing pathology lab resources and domain knowledge.

Thank you to Yngve Lamo the work package leader for the project. I am grateful for the time spent on meetings and the constructive feedback provided.

Thank you Maren Sundt Solheim for the help with designing illustrations of pathology lab cassettes.

Markus Hatlem
Bergen, 24.04.2023

Abstract

The pathology lab at Haukeland University Hospital is currently facing a few challenges. In the lab they process different specimens of tissue samples for various reasons such as looking for signs of cancers and tumours. The last six years the lab has had an average increase of 3.42% in the amount of test samples which needs to be analysed each year. This increase has led to queues forming at various stages of sample analysis. Specific samples are hard to locate within these queues and the queues lead to slowdowns of sample processing. The pathology lab require a better solution to the way and order in which samples are processed and tracked, to do so they need to gather data about the processing by implementing a tracking solution. This project aims to help them achieve this goal by looking for a potential software solution to part of the problem. This solution aims to take advantage of technologies such as optical character recognition (OCR) for detecting and tracking samples. The goal for this research is to create and test solutions for cassette detection and identification using pre-trained image processing libraries. Testing two different methods, these being edge detection and the EAST neural network, they achieved an accuracy of 77.84% and 93.41% respectively regarding cassette detection. Tesseract OCR performance of detected cassettes also varies between the two methods, giving an accuracy score of 36.1% when using edge detection and 62.1% using EAST. The increase in accuracy comes at a cost in runtime. In addition to these evaluations an in-lab trial compares the sorting time for the current solution of manual sorting versus the efficiency of sorting using the proposed digital solution. The trial concluded that the proposed digital solution is able to increase the amount of cassettes sorted within a set amount of time by 54% decreasing the time spent on manual sorting activities by 35%. This thesis also covers some of the interaction design decisions for the proposed application to allow for manual error correction. Through conceptual designs the thesis shows how the proposed system could interface with a process execution engine. There is also a proposal for what the architecture of an integrated system could look like. Integrating this system would allow for the generation of fine-grained event logs for process mining purposes. The data from these logs have a possibility of leading to future improvements in pathology lab workflow.

Contents

Scientific environment i

Acknowledgements ii

Abstract iii

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Problem Statement . . . 4
- 1.3 Objectives . . . 5
  - 1.3.1 Research questions . . . 5
- 1.4 Contribution . . . 6

2 Background 9
- 2.1 Pathology . . . 9
  - 2.1.1 What is pathology? . . . 9
  - 2.1.2 Histological sample processing pipeline . . . 10
  - 2.1.3 The cassettes . . . 11
  - 2.1.4 Data from the lab . . . 14
- 2.2 Technologies . . . 15
  - 2.2.1 Machine learning . . . 15
  - 2.2.2 Image processing and OCR . . . 16
  - 2.2.3 Python . . . 18
- 2.3 Previous Research . . . 19
  - 2.3.1 OCR research . . . 19
  - 2.3.2 Similar products . . . 21
  - 2.3.3 Research from the pathology lab . . . 23
  - 2.3.4 Relevant works in healthcare . . . 26

3 Methodology 29
- 3.1 Design Science . . . 29
  - 3.1.1 Design evaluation . . . 32
- 3.2 Proposed methods . . . 33

4 Methods 36
- 4.1 Process and procedures . . . 37

4.2 Implementation 37
4.3 Development challenges 39
4.3.1 Why not integrate with the LIS? 39
4.3.2 Pre-Processing challenges 40
4.3.3 Challenges with OCR in this domain 41
4.4 Software details and choices 44
4.4.1 Edge Detection 49
4.4.2 EAST 50
4.4.3 OCR 51
4.5 User interface 51
4.6 Suggested framework 56
4.7 Augmenting the logs 58
4.8 System architecture 60

5 Evaluation **63**
5.1 Results and Experiments 63
5.1.1 Data and analysis 67
5.2 Discussion 71
5.2.1 Performance 71
5.2.2 Specific cases 73
5.2.3 Barcode scanning times versus runtime 73
5.2.4 Comparing manual and automatic sorting 74
5.2.5 Possible use cases 76
5.2.6 Future challenges introduced by the system 77
5.2.7 Answering research questions 78

6 Conclusions and Future Work **80**

Appendix
.1 Code and file structure . . . . .
.2 Installation instructions and requirements to run the program . .
.3 How to use the program . . . . .

List of Figures

2.1 This petri net model from Stünkel et al. shows an abstract model of how a histology sample is processed in the pathology lab. Model from (*Hatlem et al.*, 2023, p. 3) updated from original model found in (*Stünkel et al.*, 2022, p. 3) 11
2.2 This illustration shows the data on the Cassettes 12

3.1 The seven guidelines for rigorous design science. (*Hevner et al.*, 2004, p. 83) 30
3.2 Evaluation methods for design science. (*Hevner et al.*, 2004, p. 86) 33

4.1 This mind map shows the different areas and corresponding sections this thesis touches upon 36
4.2 This figure shows the differences in the frontend for milestone 1 (left) and milestone 2 (left) 38
4.3 Shows a blurred cassette 42
4.4 Shows an obscured cassette 42
4.5 Shows a cassette with reflected text 43
4.6 Shows a with cassette with skewed text 43
4.7 This figure shows the pipeline for the rest API that processes images 47
4.8 This figure shows an image processed by edge detection 49
4.9 This figure shows how the algorithm merges the results from EAST to capture all the data on one cassette 50
4.10 This figure represents how the data of cassette location is converted and stored into plain-text 52
4.11 The start page on the application 53
4.12 This image shows the final user interface and the results from a scan. 54
4.13 The different colors used in the GUI 54
4.14 Shows the information that appears when clicking on a cassette 55
4.15 conceptual framework 56
4.16 Conceptual framework for enhanced process monitoring in pathology laboratories 57
4.17 Transformation of raw logs for coarse-grained and fine-grained process analysis 58

4.18 Shows how the proposed solution augments the workflow from figure 2.1 *Hatlem et al.* (2023) 59
4.19 The overall architecture of the system *Hatlem et al.* (2023) 60
4.20 BPMN model modified from *Platou* (2021) to show how the OCR tool would interface with the system 61

5.1 This graph show the variation in cassettes sorted in each 10 minute session 66
5.2 This graph shows an estimate of the sorting times saved and the amount of cassettes increasing each year 70
5.3 This image shows the difference in the cut-outs of ED and EAST 72

Chapter 1

Introduction

1.1 Motivation

The pathology lab at Haukeland university hospital perform analyses on different test specimens. An example of one type of specimens are tissue samples from human organs used for diagnosing the existence of cancers and tumours. This is a complicated process and involves the specimen getting split into multiple samples which are then analysed and processed by different machines. To maximize efficiency a lot of samples are parallel processed, this leads to the samples being split into processes at different physical locations with varying times to finish the sample analysis. The final step, which is a microscope study requires all the samples to be gathered back together again. As all the samples are stored in different locations, locating and sorting all the samples from one specimen can in some cases be very time consuming. This thesis will look at one possible solution to help the lab technicians at the pathology lab locate, sort, and archive their samples faster.

Healthcare is an important part of our daily lives. It is also important for government and society as a whole. However, healthcare systems are facing shortages of trained professionals and there is a growing demand for medical services especially considering the ageing population in Norway. Currently the healthcare industry is facing a digital revolution and work is being done at multiple stages to enhance and improve processes with new digital tools. This leads to less waiting time and faster diagnoses

which in certain cases have the potential to save lives. Appointments for pathology can have very long wait times. A patient can experience waiting for 6 months before they have a sample taken by their doctor. After the sample is taken the patient will have to wait again for the sample the get processed and analysed by a pathologist. If this process takes too long the patient can in some cases become sicker during the period of waiting. Especially if the patient is fighting a disease like cancer that progresses in different stages over time. Streamlining this process has the potential to not only help save lives, but will also contribute other benefits such as saving time and money by allowing pathologists to focus on more important tasks.

According to one of the doctors of the pathology lab at Haukeland university hospital, Friedemann Leh the pathology lab has a yearly increase of roughly 5 - 10% in test samples which needs to be analysed each year. This increase has led to queues forming at various stages of the sample analyses. In this project image processing and optical character recognition (OCR) will be utilized to automate certain parts of the process. Specifically, this thesis will be looking at the final step where currently a lab technician must locate all samples of a specimen by reading the labels of multiple samples stored in a container. Then proceed to sort all samples in numerical order by their ID, before finally checking that all samples are present. This process should be possible to automate using tools such as Python, openCV and PyTessaract.

Specimens in the pathology lab are stored in small cassettes after it is sliced into samples. Each cassette contains some information. This includes a unique id for each specimen, a batch number for each sample, a label indicating the sample is from Haukeland, the year the sample is from as well as an indicator to whether the sample is from a biopsy, an autopsy or cytology. In addition to these semantic identifiers there is also a data matrix code on the cassette. This code identifies the sample in the laboratory information system (LIS). However, the solution proposed in this thesis will not have access to the data in the LIS. The information contained on the data matrix code is therefore irrelevant as this project will be testing an OCR based approach to extract the information found on the text contained on the cassettes.

There are two main problems with the current process of manually sorting the cassettes. Firstly, it is a very time-consuming process. Second the lack of any trails if

a cassette is misplaced and sorted wrongly. Both of these have consequences that can escalate into larger issues for the pathology department. According to the lab technicians cassettes are sorted into the wrong slot on an almost daily basis.

There are multiple technologies that could be adapted to recognize the cassettes. Some of these include adding additional information to the cassette used for automatic scanning. However, the cassettes are small and the limited space for printing on them has already been used, in addition multiple data matrix codes could cause conflicts with the data matrix codes used by the LIS mentioned earlier. Cassettes could be identified by adapting a low frequency technology by using a tag such as radio-frequency identification (RFID) or near-field communication (NFC) tags. Using RFID to keep track of the cassettes compared to barcodes would allow multiple cassettes to be scanned simultaneously, increasing speed and efficiency. However, RFID could also lead to individual cassettes being harder to locate and RFID tags are significantly more expensive compared to barcodes. (*White et al.*, 2007, p. 122).

Implementing a new type of cassette to track would require a change in the production of cassettes which could be costly and time consuming. Using OCR allows the text data on cassettes to be extracted and the technology can be built into the existing process without changing anything about the cassettes themselves. The proposed solution would only require a camera able to capture images of the cassettes and the software created in this project. The result of this study has the potential to lead to a reduction of queued samples in the pathology lab, thereby increasing efficiency and sample analyses rate. The application developed also required a user interface which were developed with feedback from the pathology lab.

Another benefit that would help the pathology lab is that the system proposed by this thesis would allow cassettes routes to be traceable. This is not only good for the workers of the pathology lab in case they need to track down a specific cassette, but traceability is also important for patient safety. Cassettes that are ”lost” by being sorted into the wrong space can be very difficult to find as they can be almost anywhere in the archive. This can lead to issues and in worst case scenario lead to a wrong diagnosis of a patient if a cassette is mixed up at the wrong process.

Traceability of cassettes is extremely important to the lab. Not only to locate specific cassettes, but also for patient privacy. The current LIS system does not fully track cassettes in the lab. This is due to the long times required to scan barcodes. Scanning these at every step would introduce significant overhead. With an automatic solution, the scanning process should be significantly faster. Traceability will also allow the pathology lab to ensure that a good workflow is in place though empirical means. With the proposed system in place cassettes in the wrong location will be less of a problem as traceability will allow the pathology lab to locate when and where a cassette was put in the wrong location. Haukeland is not the only pathology lab dealing with issues of cassettes being placed in the wrong locations, as other pathology labs are dealing with similar issues as seen in the following quote. “When you’re looking at hundreds of blocks with a lot of numbers on them, eventually you’re going to pull the wrong block. It’s extremely time consuming and extremely wasteful.” -Dr.Sue Paturzo, pathology lab supervisor at Thomas Jefferson university hospital. (, Epredia, 2020, p. 2)

Traceability will also help when pulling cassettes from the archive, and activity which is done on a daily basis and one of the reasons the cassettes have to be sorted in the first place. At the moment there is no way to know if a cassette is currently sitting in its supposed location in the archive or if someone has pulled out an old cassette and are running tests on it. The reason cassettes are pulled from the archive daily could be to see if a disease is hereditary. Knowing if a patient is dealing with the same disease as their ancestors can give valuable insights on which treatments and cures have been applied in the past, and which of these treatments were successful.

1.2 Problem Statement

The problem this thesis attempts to address is the build-up of queues at various stages in the pathology lab at Haukeland university hospital. These queues are formed due to challenges related to a human’s speed of identifying and sorting samples. The existing queues and the addition of a considerate yearly increase in samples that have to be analysed each year, leads to a further build-up of queues and slows down the pathology lab considerably. The pathology lab also suffers from a lack of traceability in certain stages of processing. Both of these issues can be addressed with one solution.

To address these issue directly there is a need to propose a solution and figure out how to implement this solution without causing disruptions in the existing workflow.

1.3 Objectives

The author of this thesis hypothesises that a human's ability to identify cassettes will increase in speed when assisted by digital tools. The augmented identification will increase a user's ability to find specific cassettes and lead to an increase in cassettes sorted within a set period of time. Thus, making if faster for cassettes to be sent further into the lab instead of being stuck in queues and improving the speed of the overall process of a cassette traveling through the lab.

1.3.1 Research questions

There is one main research question which this thesis aims to answer. The main question can be split into three sub-questions which are also directly answered when resolving the main research question. All of these research questions directly relate to the main goal of implementing OCR to create an application that help identify cassettes containing specimen for the medical domain. To simplify the goals of the thesis into research questions the following research questions are proposed with RQ being the main question and RQ1, RQ2 and R3 being sub-questions:

RQ: How can the pathology workflow be streamlined by implementing automatic sorting mechanics for cassettes?

RQ1: How can the implementation of a tracking solution using optical character recognition (OCR) improve the efficiency of sample processing at a pathology lab?

RQ2: How does the accuracy of cassette detection vary when using different pre-trained image processing libraries, specifically edge detection and the EAST neural network?

RQ3: What effect does the proposed digital solution have on the time it takes to sort

cassettes compared to the current manual sorting method?

1.4 Contribution

The main contribution will be a demo of an application/artifact that aims to increase the productivity and speed of cassette sorting at Haukeland university hospital's pathology lab. This application and research around it should answers the question stated in RQ. The application developed should be tailored to the user's (pathology lab) specifications and easy to use without much prior knowledge. The process and evaluation around this artifact will answer RQ1, RQ2 and RQ3. The application should help resolve the problem of queues building up, thereby enhancing the efficiency of the lab. The application should be faster at sorting cassettes compared to the current method of manually sorting cassettes, which leads to a large build-up.

Essentially the goal for the project is to create an artifact that extracts the textual information from the cassettes and present it to a user in a way that makes it easier for the user to find specific cassettes. The program should store the location of all cassettes found in a format allowing for easy retrieval of this information.

The broader goal of the artifact is to propose a solution that allows more fine-grained tracking of a cassette's physical location inside the lab. The logs of such an application can help tell how long a cassette have been at a certain processing step and allows for backtracking to see which steps a specific cassette has been through. This is a continuous system running in parallel with the lab and will be constantly supplying data at a faster rate than what is currently done manually and with the current LIS. This data will allow the staff at the lab to locate individual cassettes by searching for them and getting an output of the cassettes previous location. It will also show the length of different processing steps and can be used to identify bottlenecks and figure out where in the most time-consuming areas to allocate additional resources. Knowing how long a cassette has been at a certain processing step and backtracking to see which steps a certain cassette has been through can provide an advantage when trying to optimize the processing pipeline by performing process mining.

The thesis includes a comparison of two different methods of identifying pathology

cassettes from images. One of these methods is edge detection, the other method is EAST a neural network created for text detection. These are evaluated on images captured in the pathology lab environment and the results are compared to identify which is more suitable for use in a pathology lab.

How to present the artifact to the user in the form of a graphical user interface is discussed in the thesis. Mainly in form of how to handle manual error correction and the interaction design choices that could make this process easier for the users of the artifact. This includes how to highlight specific cassettes with a search functionality and present an overview of all cassettes detected in an image.

The aim of this is to free up the time of the lab technicians for more important tasks. The work being done here is not removing jobs. In contrast this projects goals allow for more important work to be done faster by freeing up the lab technicians from a mundane task as the sorting process generally occupies the time of one lab technician each day.

The artifact will also contribute to the open-source community. All the work done in this thesis will be available for others to read and iterate upon further without the need for proprietary hardware or software.

The data produced by the proposed artifact could also be relevant for another current ongoing research project at the pathology lab. This project is focused on the workflows of the lab. The project aims to extract these using process mining. Certain events are not logged by the system currently in use. However, the data and event-logs generated by the artifact proposed in this thesis could be used to fill in the missing data and allow for more fine-grained process mining to take place. The data will increase the traceability of cassettes.

To show how the artifact produced can be incorporated into the workflow conceptual designs of how the artifact can interface with process execution engines are proposed. In addition to a proposal of how the overall architecture of the system could be setup if the pathology lab implemented an OCR based system for cassette tracking.

To compare the results of an automatic storage system to the current manual approach this thesis also includes data gathered from the pathology lab. This data has

been collected through a series of interviews, meetings, and natural observation of certain tasks within the pathology lab. This thesis will help make this data more openly available by open-sourcing data which will allow for others to help contribute to and use this data to further advance the field of pathology lab workflows.

Chapter 2

Background

This part of the thesis will present the background information required to understand the theory behind certain topics related to the thesis. Firstly, pathology concepts and workflows will be introduced. Then the different technologies utilized will be described. Prior research and current techniques related to this thesis will be discussed and an overview will be presented. Finally related works will be explored.

2.1 Pathology

This part of the thesis introduce the problem domain. Here pathology workflows and concepts required for understanding the work done in this thesis are presented.

2.1.1 What is pathology?

According to the oxford English dictionary pathology is “”The study of disease; the branch of science that deals with the causes and nature of diseases and abnormal anatomical and physiological conditions; (in later use) esp. the branch of medicine that deals with the laboratory examination of body tissues, cells, and fluids for diagnostic purposes. Frequently with distinguishing word.” (Oxford English dictionary, Pathology, 2022). This definition lines up with what Stünkel et al. points out when they say that the word pathology can be split into the Greek words “pathos” and “logos” which translates to “the study disease” or more narrowly “study of cause and ef-

fects of disease" (*Stünkel et al.*, 2022, p. 3). Pathologists do not usually interact with the patients themselves, but rather act as an advisor and consultant to other medical professionals. "The primary clinician takes a specimen from the patient, e.g., a tissue sample, and sends it to the pathologist, who examines the specimen and writes a report, most often with a conclusive diagnosis, which will help the clinician on deciding the further treatment, e.g. whether surgery or chemotherapy has to be scheduled" (*Stünkel et al.*, 2022, p. 3). The medical field is currently undergoing a large-scale digitization, and this also impacts the pathology department. This thesis will contribute to this digitization by looking at the potential use case of a new tool.

In this thesis samples from three different strands of pathology will be encountered which the information system must be able to handle. These are Histology/biopsy, cytology, and autopsy. Histology deals with the study of tissue abnormalities using gross and microscopic examination of biopsy samples. (*Othman*, 2019, p. 11). A cytology sample is a sample of a single cell specimen (*Stünkel et al.*, 2022, p. 3) and deals with the study of cellular changes. (*Othman*, 2019, p. 11). An autopsy sample deals with pathological examination of human cadaver after death(*Othman*, 2019, p. 11). Autopsy samples can be used to confirm or identify an illness and in cases where it is applicable an autopsy sample can be used to observe if treatment had any effect before the death of the patient. All of these three sample types are processed by different processing pipelines in the pathology lab and therefore have differences in processing times.

2.1.2 Histological sample processing pipeline

This thesis will focus on improving the processes in the laboratory. To give an example of what is done in the pathology lab a description of the processing of a histological sample will be provided. This process can be seen in the petri net in figure 2.1. "A petri net is a abstract, formal model of information flow." *Peterson* (1977). In this case it visualizes the workflow of a histological sample in the pathology lab.

The processing of a single histology sample at the pathology lab is a time-consuming process. After a sample has been extracted from its host it is placed in a container of fixative solution and put in cold storage until it is ready to start processing. Sam-

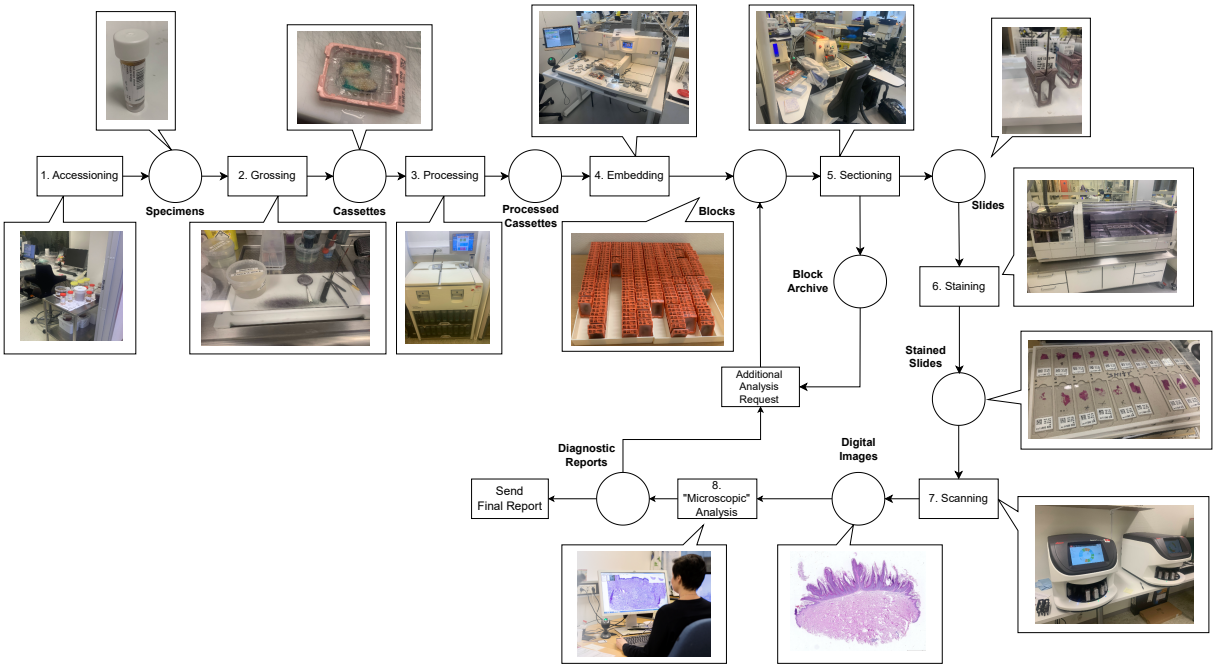


Figure 2.1: This petri net model from Stünkel et al. shows an abstract model of how a histology sample is processed in the pathology lab. Model from (Hatlem et al., 2023, p. 3) updated from original model found in (Stünkel et al., 2022, p. 3)

ples start out with grossing, here samples are sliced and examined by pathologists before being placed in cassettes. The samples are then split-up and put into separate machines for parallel processing. “This step is performed by a specialized machine that automates dehydration, clearing and infiltration of the tissue with paraffin wax” (*Stünkel et al.*, 2022, p. 4). Once the processing is finished the samples are embedded in paraffin. They are then sectioned into slices and stained before being studied under a microscope and once the study is finished the samples are put away for archiving. Between all these stages are large queues where the sample have to wait for an available spot in the processing machines. These queues of cassettes found around processing machines as well as before archive is where this project aim to make an impact.

2.1.3 The cassettes

The cassettes used at the pathology lag contain various information easily readable by humans, but the text provides challenging for a digital actor to interpret. The cassettes also often referred to as ”blocks” in the pathology lab is the plastic housing that holds the embedded sample. Before plastic cassettes were adopted as a method of storage

the samples used to be stored in embedded paraffin wax blocks hence the multiple names. In this paper the term cassette will be preferred, but in the pathology lab the two terms ”block” and ”cassette” are interchangeable. Cassettes contain a data matrix code which is used in the internal systems. These data matrix codes must be scanned one by one and cannot be scanned in batches. The OCR based system proposed in this thesis aims is able to handle multiple cassettes at the same time. The existing system records some data, but hand scanning takes too long therefore the cassettes are not completely tracked throughout the lab. This leads to certain events missing from the event-log generated by the existing LIS. The work done in this thesis aims to fill in the missing data by applying an automatic solution. In three out of four places in the pathology lab the OCR system could be used to scan cassettes in batches and fill in the data missing from the LIS. These three steps are coordinating which is a process that takes place before processing, during processing itself and archiving. It will also allow for an additional activity to be added to the archive. This would be a searching activity for finding cassettes when needed.

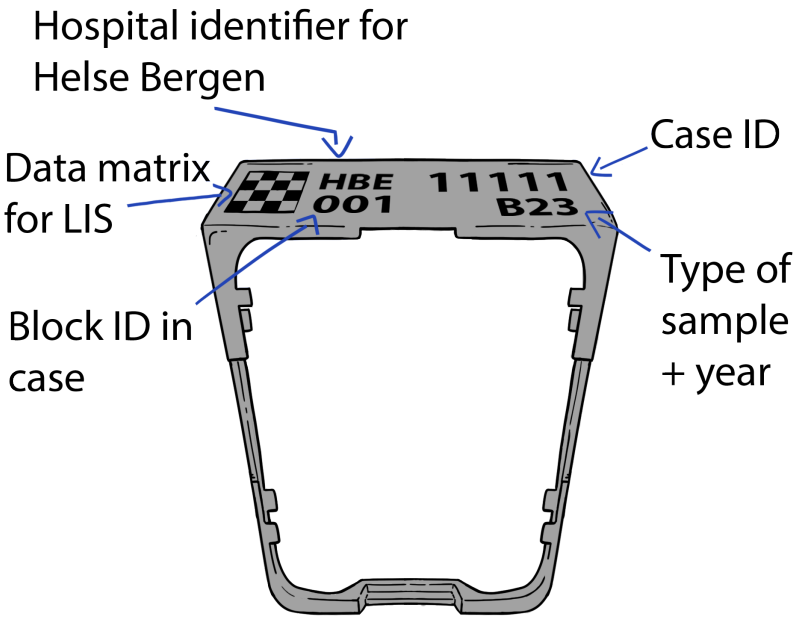


Figure 2.2: This illustration shows the data on the Cassettes

The top of the cassettes contains text data, this data is what the OCR system will extract. The cassettes also contain a data matrix code. This code contains a unique number used to identify specific cassettes in the LIS. The LIS system connects the

data on the cassette with patient data as well as other information about the sample. This data matrix will not be relevant to the OCR system. The data relevant to the archives and OCR system is:

- The hospital identifier - This is noted in figure 2.2 by HBE indicating that this cassette belongs to Helse Bergen. All other pathology labs within Helse vest use a distinct 3-letter identifier.
- The case identifier/(case ID) - This is very important and can be found in the top right of the cassette. This number indicates which case the cassettes belong to and is a continuous sequential order/requisition number within a year and specimen category.
- Block ID per case - Each case can have multiple blocks related to them. These are annotated by the Block ID which is the number seen in the lower left of the cassette. The samples per case can range from 001 to 999 and there are no holes in the numbering, this means that if the highest number detected is 98 the system should expect there to be 97 other cassettes related to the same case. The total amount of samples per case are unknown without access to the LIS which provides an additional challenge for the system as it does not have access to this information.
- Type of sample + year - The last data point on the cassette contains type of sample stating whether it is a biology (B), autopsy (O), or cytology (C) sample in addition to the last two numbers in the year which the sample was taken. This identifier is located in the bottom right of a cassette.

The case ID in combination with the block ID and Type of sample gives each cassette a unique identifier that would allow each cassette to be sorted individually. This unique ID would also not conflict with older cassettes from the archive.

The cassettes are color-coded. These colors signify different aspects to the staff such as which cassettes are a priority, if a sample is split into a lot of cases and if a certain sample has to be processed in a specific manner.

Some of the most common cassette colors and their significance:

- Red - Needs to be analysed now, the patient could be under an operation and the doctor is waiting for pathology results
- Yellow - High-priority
- Orange - Large, high-priority samples
- Pink - Small samples
- White - Very large, low-priority samples
- Purple - Intestine sample
- Blue - Autopsy sample

One of the largest challenges when sorting cassettes into the archive is to ensure there is enough space left open in the event of a case containing many cassettes need to be fitted between two other cases. Due to this the archive section need to keep empty space in each box when sorting. Storage space cannot be compressed until all cassettes are accounted for. This can take a long time as priority cassettes and parallel processing ensure that cassettes arrive in a different order from the numerical case ID which they are sorted by.

2.1.4 Data from the lab

Table 2.1: Cassettes sorted on average in different times

Year	Number of cases	Total number of samples/cassettes processed
2016	46276	155414
2017	48670	164179
2018	51110	169817
2019	52277	172879
2020	50625	171505
2021	54564	180323
2022	56637	185065

The pathology lab has data showing how many cases they process each year, as well as how many samples (cassettes) each case contains. Compiling this data allows us to see how many total cassettes are processed in the lab each year. All of these

cassettes also need to be sorted, this data will allow for estimates regarding how long this task takes. Comparing 2016 to 2022 shows us that in those 6 years they had a 22% increase in cases and a 19% in samples. The amount of samples per case can vary a lot depending on a case by case basis. If we look at the increases each year, we can see that every year besides 2020 has had a steady increase in samples. The reason for the slight fall in cases analysed in 2020 is most likely due to the COVID-19 epidemic, which also explains the significant rise seen in the gap between 2020 and 2021. The average increase in samples per year the last six years is 3.42% in cases and 3.39% in samples.

2.2 Technologies

This part will cover technologies that will be relevant in the project.

2.2.1 Machine learning

One of the tools utilized in this project is the Efficient and Accurate Scene Text detector (EAST detector) neural network as well as Tesseract. To have a better understanding of how these models work we first need to understand machine learning and subsequently neural networks. Machine learning (ML) is the process of training a computer to detect certain patterns by feeding it input and rating the output based on various factors. The English Oxford dictionary defines machine learning as ”the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data.” (Oxford English dictionary, Machine learning, 2022). ML has many applications including natural language processing, sentiment analysis, image processing and can also be used for classification tasks. Neural networks (NN) are a subset of machine learning. NN’s are inspired by how the human brain works and the neural structure, hence the name. There are a multitude of different network types such as the perceptron and feed forward neural networks. The different neurons (nodes) are linked together and send a signal through the network. The strength of this signal is calculated based on the weights and biases. More advanced networks also include support for back propagation. Back propagation means

the network has the ability to send data back to the previous node which allows it to learn from the earlier result and calculate weights more accurately in a way that reflects and considers earlier trials. LSTM stands for long-short term memory and utilize back propagation. Tesseract 4.0 has an OCR engine based on LSTM neural networks (Tesseract - GitHub.com, 2023). The other NN utilized in this thesis is EAST text detector which is a fully convolutional network (FCN) (*Zhou et al.*, 2017, p. 1). An FCN also known as a deep filter makes use of a nonlinear filter. This means that it is a neural network where all layers compute a nonlinear filter of average pooling or utilize matrix multiplication for convolution (*Long et al.*, 2015, p. 3). Deep learning-based OCR is an emerging technology that could be very relevant to this thesis, but will not be covered here as it would be too extensive.

EAST *Zhou et al.* (2017) introduces an efficient and accurate scene text detector (EAST). EAST is a fully convolutional neural network which aims to be as efficient and accurate as possible. In the paper the authors propose "the first fast and accurate scene text detection pipeline that has only two stages" *Zhou et al.* (2017). The paper outlines the pipeline and the decisions behind the neural network training design such as loss function, optimizers and number of batches required. Quantitative and qualitative experiments were done to compare with three public benchmarks datasets. These datasets were ICDAR 2015, COCO-Text and MSRA-TD500. The resulting finds were that "compared with existing methods, the proposed algorithm achieves significantly enhanced performance, while running much faster, according to the qualitative and quantitative experiments on standard benchmarks." *Zhou et al.* (2017)

2.2.2 Image processing and OCR

Image processing will be one of the topics discussed as it is part of the proposed solution for the challenges the lab is facing. Image processing is quite an extensive field and there has been done a lot of research around it. This section will cover the most essential parts that relates to this thesis. Essentially the point of image processing is to handle data contained in an image. This could be done by extracting data from the image, or by manipulating and changing certain aspects of the image. Image processing can also be split into subcategories such as text extraction, facial recognition and

vision systems for autonomous vehicles. Just to mention a few of the many use cases and subfields.

Optical character recognition (OCR), occasionally referred to as text extraction from images refers to the process of extracting text from images turning the ”text into analysable, editable and searchable data.” (*Memon et al.*, 2020, p.142,642). It is also one of the many sub-fields contained within the genre of “image processing”. To learn how to identify text from an image machine learning is often utilized. In this project the author will not be training a text extractor but rely upon a previously trained alternative created which will be described further later. The last few years ”OCR has gained increasing attention in both academic research and in industry” (*Chaudhuri et al.*, 2017, p. 9)

What advantages could the OCR tool lead to in the lab? Mainly there are three main use cases for the tool.

- 1. Finding lost cassettes
  The tool will be able to see when and where a cassette has been scanned. Using this data, one could track down the area a missing cassette was last seen in and get an idea of where to start looking for it.

- 2. Data extraction from tracking cassettes in the lab
  This is more of a general advantage. Cassettes being tracked and timestamped enables the lab to extract data such as timings and cassette paths through the lab. This data can be used to locate bottlenecks and help find areas which are in need of more resources or enable the lab to start working on solutions to currently unidentified problems.

- 3. Reduces time spent on sorting cassettes by hand for archiving
  Currently there is no digital archive system. All cassettes are sorted by hand and stored manually in numerical order by case number. This is an extremely large and time-consuming process which takes up the time of a trained lab technician every day. Being able to create a digital copy of where cassettes are stored in the archive with the OCR tool would allow the lab to just scan cassettes in batches and label where they are stored. Doing so will massively improve the speed as

it eliminates the need for a cassette sorting process and instead replace it with a scanning process which should be faster.

There are many OCR applications currently in user in different industries. Automatic number plate recognition (ANPR) is an example of a use case for an OCR application. For example, registering cars in a parking garage to keep track of how long it has been parked. ANPR is also a useful tool to ensure road safety. An example could be the Norwegian public roads administration (Statens Vegvesen) who use ANPR to scan vehicles for safety reasons such as ensuring road tax have been paid or that a vehicle does not have a deregistration claim (Statens Vegvesen, 2023).

Another example of OCR used in practice today is for automatic mail sorting of postal envelopes. Here a similar challenge to this project is presented. First the destination address block has to be identified, before OCR can take place *Radha and Aparna* (2013). Similar techniques could be of use when looking for the data on the cassettes. However, the images from the pathology lab contain multiple cassettes per image in contrast to the mail where each image only contains a single piece of mail with only one area of interest per image.

2.2.3 Python

The foundation of the project will be built in Python. Python is an interpreted, object oriented and high-level programming language. ”It incorporates modules, exceptions, dynamic typing, very high-level dynamic data types, and classes. It supports multiple programming paradigms beyond object-oriented programming, such as procedural and functional programming.” (Python, 2022). Python is generally considered easy to work with and very popular among the machine learning crowd. This means a large selection of machine learning related libraries are available to use. In addition to Python, multiple Python libraries will be utilized. The most important libraries related to this project are OpenCV and PyTesseract.

OpenCV (open computer vision) offers a real-time optimised computer vision library and tools. This makes it easier to process images in python as code does not need to be written from scratch. ”OpenCV was built to provide a common infrastructure

for computer vision applications and to accelerate the use of machine perception in commercial products" (OpenCV, 2022). It includes a plethora of algorithms that can be applied for a large variety of tasks. Anything from recognizing faces, tracking, and identifying 3d objects, image up-scaling and much more. For this thesis it will be employed as a way of processing and representing image data in a format the OCR tools recognize.

PyTessaract is a wrapper for Google's Tesseract-OCR Engine. This simplifies the process of reading text from images as it has already been trained for this purpose. This will allow us to extract the text from the cassettes. Tesseract rivals the performance and accuracy of commercial applications according to its release paper: "Tesseract is now behind the leading commercial engines in terms of its accuracy." (*Smith*, 2007, p. 633) This in addition to the fact that it is open-source and freely available to use makes it ideal for this project. Other alternatives include Vision AI by Google (Vision AI, 2022) and Azure computer vision by Microsoft (Azure Computer Vision, 2022). However, both of these tools requires to be paid per API call and neither of them are open source. They are thereby unideal for this project.

2.3 Previous Research

2.3.1 OCR research

Computer vision technology has been steadily progressing since the invention of computers all according to Moore's law which expresses that we'll see a "doubling in the numbers of transistors placed on an integrated circuit every two years" (Moore's law, 1965). The doubling of transistors should also in theory lead to a large increase in compute power every two years. The rise of cloud computing, neural networks, and the increase in compute power seen in the early 2000's all played a factor in computer vision becoming more and more usable in toady's systems. This allows features such as unlocking your phone with your face and the technology is also used in prototypes of driver-less cars. Together with all the other use cases for image processing OCR has also emerged and have been gaining more traction.

The author has not been able to find any cases where OCR has been used for the

exact same purpose as proposed in this thesis. OCR is not new to the medical field; Tesseract OCR is currently being implemented in a pipeline which aims to scan and process physical written medical reports from the UK's NHS to a digital PDF format. (*Karthikeyan et al.*, 2021, p. 2580)

In this pipeline the authors tried to enhance the process of digitizing patient journals, as OCR is not perfect and can cause errors. Especially when dealing with medical terminologies not commonly found in general language lexicons. This is also important to keep in mind for this thesis. However, instead of medical terminology The OCR engine in this thesis will be mostly dealing with digits and serials and ensuring that the OCR technology read them correctly. There is also a large difference between scanning large hand written reports and extracting printed data from digital images.

Outside of the medical domain in a project more similar to the one conducted in this thesis OCR is being implemented to track wines based on their serial numbers. (*Cakic et al.*, 2020, p. 1) As of currently writing a pilot study has been released. The authors tested and compared several OCR tools and their performance for the wine tracking. The OCR engines tested were: PyTesseract, Vision Ai by Google and Azure computer vision by Microsoft. In the end the authors decided upon using Tesseract OCR for their pilot study. However, their artifact was only tested under lab conditions on pre-captured images which had been cropped and audited manually. The research published so far does not contain any real word performance. Their implementation managed to correctly identify 87.5% of their labels. (*Cakic et al.*, 2020, p. 4) The additional challenge introduced in this thesis where the application will be tested in real world performance compared to lab trials are mostly factors such as lighting, camera angle and many more external factors which play a role in the quality of the image and thereby the resulting classification.

With the evidence presented so far, there is an impression that the current OCR technology still has a way to go until it is ”perfect” and can be utilized completely reliably on its own. Which is why the application created in this project will work in conjunction with a human user. As the state of the art OCR technology has an accuracy of roughly 90% in handwritten English with a ”character error rate(CER) and word error rate(WER) of 4.7%, 8.22%, 2.46%, 5.68% respectively” (*Memon et al.*, 2020, p.

142,655). Based on these results a claim can be made that current OCR technology is not suitable for solving sensitive tasks alone and still needs a manual overseer to confirm the results. Which is the implementation that will be tested in this project.

2.3.2 Similar products

There are also products available that servers a similar purpose in a pathology lab as this thesis aims to achieve. An example could be the product called PathTracker by SPOT Imaging which is a company that produce science imaging systems for pathology, bioresearch, and OEM applications (SPOT Imaging, 2023). However, instead of opting for OCR they aim to scan cassettes in batches by scanning the data matrix codes. They offer a specialized machine with software to do this operation and their way of scanning codes would require access to the LIS, which this project aim to avoid when testing. In contrast the proposed solution in this thesis would not require an expensive specialized machine, just a computer and a camera.

Another company that has developed a solution to this problem is Dreampath (Dreampath, 2023). They have developed a system called FINA currently being considered as a solution to the archive problem in the pathology department. This is a larger machine that scans the cassettes and archives their information. The drawback is that purchasing this machinery is expensive and FINA also requires proprietary storage solutions such as proprietary storage racks for cassettes, purchasing more of these could get expensive in the long run. Therefore, Dreampath's solution suffers from the same drawbacks as SPOT's solution.

Dreampath compares and advertises its product by referencing another product named Arcos by Epredia which seems to have the exact same functionality as FINA (Epredia, 2021). Epredia also provide some interesting research and case studies regarding how much money and time has been saved using their systems, these would be relevant here as the saved time would be interesting to compare to the system developed in this project.

An example of a hospital that adopted FINA is Thomas Jefferson university hospital (TJUH) in Philadelphia, USA. They process over 800 to 900 cassettes daily (, Epre-

dia, 2020, p. 2). Some of their problems where with the traceability of their cassettes. ”A block would be removed from the serially organized cabinet, and there would be a discontinuation of the audit trail, jeopardizing patient safety. Without proper documentation, there was no way to track who retrieved it, when it was retrieved, and for what purpose it was removed. A documentation process was difficult to enforce with this workflow” (, Epredia, 2020, p. 2). This was a large problem as according to the supervisor TJUH’s pathology lab Dr.Sue Paturzo ”we are constantly going back to our blocks, whether it be for additional testing or research for a clinical trial, and being able to locate them in a timely manner and track them is the most important thing for me.” (, Epredia, 2020, p.3)

Arcos helped the team at TJUH grain control of material management. This led to multiple gains such as ”The check-in/out controls provide visibility into who retrieved a block, why it was needed (reason), who requested the block (requestor) and when it is due to return (TAT reporting). Blocks that are checked out longer than anticipated populate a report that can be run as needed. As a result, the team no longer wastes time organizing blocks in numerical order and the increased traceability ensures patient safety is maintained.” (, Epredia, 2020, p. 4). In addition to all these TJUH also saved $51,168 due to the 2132 hours they saved by sorting, retrieving and refilling blocks using a digital system instead of manually. (Epredia, 2020)

The First Hospital of Jilin University in China have greatly benefited by adopting Arcos after testing it from 2019 - 2020 as it ”saves a lot of manpower and material resources, so that pathological data can be more traceable, reduce the risk of sample loss, ensure the safety of samples to the greatest extent, and improve the efficiency of pathological data management.” (*Wang et al.*, 2020, p. 2). During the year they saved 2356 hours on sorting an filing and they saved 203 hours on reworking. *Wang et al.* (2020). Before adopting Arcos they struggled with similar issues as Haukeland’s pathology lab related to ”a large number of sections and wax specimens that makes it difficult to locate specimens for filing, borrowing and returning” (*Wang et al.*, 2020, p. 2).

The process of manually storing is highly susceptible to human error. If a cassette is placed or sorted into the wrong place it can lead to an extremely time-consuming process to retrieve it. The Jilin University ”no longer suffer from issues related to ”loss”

of cassettes artificially created through filing errors (*Wang et al.*, 2020, p. 3) after using Arcos.

The data from these systems show that this project can be of value. If we had access to one of these machines it would be interesting to see the differences in results when comparing their method of data matrix scanning versus OCR based scanning proposed by this thesis.

2.3.3 Research from the pathology lab

Earlier research has also been done in the pathology lab at Hukeland, most notable for this case is the currently ongoing research by Stünkel et al. Where the co-supervisors of this thesis look at process mining the lab at Hukeland. In a paper recently published they cover some of their work done so far as well as some of their future plans and challenges. *Stünkel et al.* (2022). In this paper they give us a good overview of the aforementioned issues in the pathology lab and why they are important to solve. They propose to utilize process mining as a solution.

Process mining is the extraction of valuable, process-related information from event logs (*Van Der Aalst*, 2011, p. 3) (i.e., chronological records of activities). This data can give valuable insight into business processes. They have a plethora of use cases such as giving insight, starting structure discussions, verification procedures to find errors, performance analysis, animation models to test different scenarios, specification of models before creation and configuration of systems (*Van Der Aalst*, 2011, p. 6). Process mining allows an organization to identify activities and how long each activity takes based on the aforementioned event log. ”Unfortunately, often the 20% least frequent behavior may cause most of the compliance and performance problems. This is called organizational friction. Process mining aims to identify and remove such organizational friction” (*van der Aalst*, 2019, p. 4).

For the process mining done at Haukeland “the overarching objective of the project is to reduce the overall cycle time in the pathology department, i.e., the time from receiving a specimen to sending a diagnostic report back.” (*Stünkel et al.*, 2022, p. 5). Further on they describe a few of the challenges they have faced before and during the

process mining. Firstly, security clearance is an issue. The domain data in this case is healthcare data which is highly regulated and have “strict requirements concerning access to data: the project had to apply for exemption from the duty of confidentiality, to do a data protection impact assessment, to carry out a risk analysis and to establish a data management plan” (*Stünkel et al.*, 2022, p. 6). The project also faced technical issues once access to the LIS data had been granted. These issues ranged from badly structured data to lack of certain events (*Stünkel et al.*, 2022, p. 8). One of the larger challenges was that the pathology process is hard to represent using only atomic-token based workflow which means that “a case is represented as an atomic token that flows through a net structure representing the control flow” (*Stünkel et al.*, 2022, p. 9). The effect of this is that “existing process mining algorithms are not perfectly suited for the specimen preparation workflow in the pathology laboratory” (*Stünkel et al.*, 2022, p. 9). These are the only issues they have faced so far in their work, they also outline a few potential future challenges for later in the project. One of these being that the authors are unsure how they “eventually can transfer the analytical results to operational results. For instance, there are some physical limitations to what degree a “redesign” of the process is possible.” (*Stünkel et al.*, 2022, p. 10). The other future challenge they highlight is the potential future social ramifications of their research. One of the precautions they will be taking is to ensure that the workers of the lab will have their privacy insured by “hashing all usernames with a random and hidden salt” (*Stünkel et al.*, 2022, p. 10) thus obscuring all personal data related to specific users. Some of these challenges can be relevant to keep in mind when looking at what will be done in this project.

Another master thesis written by Platou at Haukeland pathology lab focuses on business process simulation in the pathology lab *Platou* (2021). Most of Platou’s work focuses on business process simulation, but there are some key points which will also be relevant for the project. One of the interesting things he points out is that the logs form the LIS does not include end and start points for specific tasks (*Platou*, 2021, p. 47). This is definitely information that would be possible to add into the system proposed by this thesis.

To better understand Platou’s work at the pathology lab we need a definition of Busi-

ness process management (BPM). “BPM is a comprehensive system for managing and transforming organizational operations” (*Hammer*, 2014, p. 3). Platou worked more specifically on business process simulation (BPS). “It is important to analyze processes not only before they are put into production (to find design flaws) but also while they are running (for diagnosis and decision support).” (*Van der Aalst et al.*, 2010, p. 314). This gives you an idea of where BPS comes in. As it allows you to simulate both future and current cases in a production environment to generate valuable data.

A process model includes processes (sometimes referred to as activities) as well as the flow between these processes. Business Process Modelling Notations (BMPN) have been adopted to represent these processes. A BMPN is supported by a process execution engine which allows processes to be executed in a predefined workflow. The pathology workflow can be shown using a BMPN execution engine as seen in the work done by Platou. *Platou* (2021). Here Platou used the state-of-the-art BMPN engine Camunda *Camunda* (2023) to represent the workflow of the pathology lab.

Platou did some experiments on the data exported from the LIS. As well as some pre-processing such as removal of rows with missing data (*Platou*, 2021, p. 51). Once data clean-up was complete, he created a business process model and ran four different experiments testing various constraints in the model he created.

The aforementioned studies that are done at Haukeland pathology department give insightful knowledge as to what goes on in the lab and what has been attempted before in terms of solving the queue problem and gathering data. They also help line out some potential challenges for this study as all these studies are essentially looking at different solutions to the same problem.

Both of these studies goes in-depth looking for digital solutions to the problems the lab is facing. However, to a certain degree they ignore the physical aspects of the lab and the tangible nature of the work done there. This paper will aim to bridge the physical and digital aspects of the pathology lab and create a software solution that encompasses and solves both real world and software challenges.

2.3.4 Relevant works in healthcare

In (*Dangott*, 2015, p. 43) The concept of a specialized laboratory information system is introduced. These systems are defined by the need to perform a “a limited number of functions extremely well rather than trying to serve the needs of an entire laboratory” *Dangott* (2015). The article outlines four pillars essential to consider when developing an in-house solution. These four pillars are scalability, building a design team, support costs and total cost of ownership versus long-term benefit. The proposed solution fulfils the requirements of a specialized laboratory information system and should therefore take these four pillars into account.

- Scalability and timeliness: These are very dependent on the existing LIS and the support that is available for it. Commercial solutions may already offer modules which already fit into the current LIS. These can often be integrated faster than an in-house solution. “If there is no acceptable product, then developing in house may be faster than waiting for a vendor solution” (*Dangott*, 2015, p. 45).
- Building a design team: It is important to consider that the design team must be familiar with the constraints. They also highlight the danger of such a system only being understood by few individuals and that “Retiring or departing staff who possess this intellectual capital can leave a knowledge gap that can be hard to fill” (*Dangott*, 2015, p. 45).
- Supports costs: Keeping in mind maintenance costs after the system is build and implemented. How many people are required to maintain the system and how often would it require maintenance?
- Total costs of ownership versus long-term benefit: Dangott presents eight questions to keep in mind in regards to the long term benefits and costs.
  ”
  - 1. Will the system deliver strategic advantages to the laboratory?
  - 2. Will the workflow be more efficient or allow more automation?
  - 3. Will the in-house system allow customized reports that are strategically

valuable to the client base?

- 4. Can integration with other systems and instrumentation be easily achieved?
- 5. Are Web-based reports a priority?
- 6. Are dollars that would be invested in LIS development better spent elsewhere?
- 7. What is the long-term impact of the LIS on the organization?
- 8. Does the in-house product allow better-quality management and improve operational efficiency?
  ” *Dangott* (2015).

These four pillars also applies if the pathology department wants to solidify the work done here and create their own in-house system for cassette management. This is also important for this project as the work done in this thesis could be built upon in the future.

Hanna and Pantanowitz outline the history of barcodes in pathology *Hanna and Pantanowitz* (2015). They outline the different types of codes used, including the 2-D data matrix codes utilized at Haukeland. They detail how Hospitals have improved their error rate by incorporating barcode scanning and they explain the shortcomings of how barcode scanning could fail. “A major advantage of implementing a bar coding and tracking system is the opportunity to eliminate labelling errors and achieve optimal patient safety, consequently reducing adverse events.” (*Hanna and Pantanowitz*, 2015, p. 17). These advantages would also apply to the system proposed in this paper.

In 2009 Buese looked at adapting lean workflow models to histology labs. He starts by presenting the history of workflow methods all the way back to the automotive industry, and how these management techniques evolved to include quality control and total quality management concepts. Before introducing lean manufacturing or just “lean” which “was coined to describe its fundamental characteristics of unitary production, minimum waste and customer “pulling” of the production process” (*Buesa*, 2009, p.324). Buese then goes on to show how these methods have been applied to

Histology labs. The article summarizes results of 25 histology facilities which implemented management tools and from this data extrapolates that laboratories handling more than 20 000 cases a year gain a greater benefit by incorporating these methods. However, all 25 hospitals saw an increase in performance. The article concludes by suggesting 13 changes to improve the workflow in any histology lab.

In *Zayas-Cabán et al.* (2021) 123 articles related to automation approaches implemented in different industries are reviewed to identify opportunities for workflow automation in healthcare. They identified characteristics that promotes automation are "manual data entry, high frequency or repetition, clearly defined independent and dependent variables for modelling, clear roles and responsibilities." (*Zayas-Cabán et al.*, 2021, p. 689). The paper found that different tasks can be automated to different stages such as low-, semi- and fully automated tasks. The level of automation is highly reliant on how well the task is defined. These findings support the argument to automate the sorting task in the pathology lab and the task should in theory be fully automatable.

Chapter 3

Methodology

For this thesis design science will be employed as a research methodology. The main goal of the project has been to develop a solution to the sorting problem taking advantage of the emergence of new technologies such as image processing and text extraction. The development of the application followed guidelines and design patterns specified in the design science approach as detailed below.

3.1 Design Science

Information systems and the tasks they perform are complex. To help us design and implement such complex systems we require rules and guidelines of which we need to abide. Design science is one such research paradigm, it has seven core rules which needs to be followed to ensure the smooth creation of an artifact. ”IT artifacts are broadly defined as constructs (vocabulary and symbols), models (abstractions and representations), methods (algorithms and practices), and instantiations (implemented and prototype systems)” (*Hevner et al.*, 2004, p, 77) ”Behavioural science encourages research by developing and justifying theories that explain or predict phenomena. Whereas design science aims to encourage research through building and evaluation artifacts. The goal of behavioural science is truth. The goal of design science is utility. Hevner et al claim that that truth and utility are inseparable. Truth informs design and utility informs theory.” (*Hevner et al.*, 2004, p, 80). According to Hevner the essence of Design science is inherently a problem-solving process.

There is an ongoing discussion regarding the final artifact of design science. Hevner argues that it should produce an artifact as described above, whilst other researchers like Jones and Gregor claim design science should result in a design-science theory. A theory which can generalize a product architecture or generalize a method (*Jones and Gregor*, 2007, p. 322). No matter which of these arguments are supported ”The design-science research paradigm is still evolving.” (*Williamson and Johanson*, 2017, p. 268) and any artifact which follow the rules stated by either claims are considered a piece of design science.

Here are the seven guidelines that design science implores to evaluate the artifact. These guidelines exists to ”to assist researchers, reviewers, editors, and readers to understand the requirements for effective design-science research” (*Hevner et al.*, 2004, p, 82). For a project to be a successful piece of design science each of these guidelines should be addressed at some point.

Table 1. Design-Science Research Guidelines

Guideline	Description
Guideline 1: Design as an Artifact	Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation.
Guideline 2: Problem Relevance	The objective of design-science research is to develop technology-based solutions to important and relevant business problems.
Guideline 3: Design Evaluation	The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods.
Guideline 4: Research Contributions	Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.
Guideline 5: Research Rigor	Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.
Guideline 6: Design as a Search Process	The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
Guideline 7: Communication of Research	Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences.

Figure 3.1: The seven guidelines for rigorous design science. (Hevner et al., 2004, p. 83)

Below is described how each guideline in the aforementioned guidelines seen in fig-

ure 3.1 were used to apply design science research (DSR) to this thesis. How these guidelines are followed will be seen more in depth in the proposed methods section.

1. An application have been produced through an iterative trial-and-error search process. It has the capability to identify specific cassettes using OCR to extract the data from cassettes labels. It has been tested to help sort the cassettes in the pathology lab and with the GUI it is able to tell the user if any cassettes are missing and the GUI allows the user to input the missing data.

2. This exact problem has yet to be solved using OCR methods. However, other methods such as process mining and Business process simulation has been attempted. There are also machines available that achieves similar results, but these are proprietary and the results are achieved through other methods than OCR. In regard to the more specific OCR challenge, OCR is mostly a cutting-edge technology which leads to it being considered too risky to be applied in a lot of cases. Since this system will be used in conjunction with a user who can verify results, applying OCR to this problem is more applicable. And second, this problem is a relatively recent problem. The problem occurred when the pathology lab got too many samples to analyse, and the samples started queuing up in the sorting zones. A few years ago, the amount of sample the lab had to analyse was low enough that a lab technician could reliably sort the samples manually. Now with the growing number of samples they need additional tools to help manage the workload.

3. The application have been tested on images from the pathology lab to ensure it fulfils the applications requirements.

4. This artifact will test the process of using OCR to see if it is a viable approach to solve challenges that previously required manual labour. To show that OCR can be an important tool in the future for workflow improvements in the pathology lab.

5.The application will be freely available as well as the code used for creating it, so all experiments should be replicable.

6. The artifact has been created with background research looking at earlier work in similar domains.

7. The main research outcome will be this thesis which should be approachable by people unfamiliar with both image processing technologies and pathology workflows.

3.1.1 Design evaluation

Once an artifact has been developed it has to be evaluated by certain criteria as evaluations are crucial for the research process. ”IT artifacts can be evaluated in terms of functionality, completeness, consistency, accuracy, performance, reliability, usability, fit with the organization, and other relevant quality attributes” (*Hevner et al.*, 2004, p. 85). Table 3.2 shows the evaluation methods proposed by Hevner et al.

Table 2. Design Evaluation Methods	
1. Observational	Case Study: Study artifact in depth in business environment
	Field Study: Monitor use of artifact in multiple projects
2. Analytical	Static Analysis: Examine structure of artifact for static qualities (e.g., complexity)
	Architecture Analysis: Study fit of artifact into technical IS architecture
	Optimization: Demonstrate inherent optimal properties of artifact or provide optimality bounds on artifact behavior
	Dynamic Analysis: Study artifact in use for dynamic qualities (e.g., performance)
3. Experimental	Controlled Experiment: Study artifact in controlled environment for qualities (e.g., usability)
	Simulation – Execute artifact with artificial data
4. Testing	Functional (Black Box) Testing: Execute artifact interfaces to discover failures and identify defects
	Structural (White Box) Testing: Perform coverage testing of some metric (e.g., execution paths) in the artifact implementation
5. Descriptive	Informed Argument: Use information from the knowledge base (e.g., relevant research) to build a convincing argument for the artifact's utility
	Scenarios: Construct detailed scenarios around the artifact to demonstrate its utility

Figure 3.2: Evaluation methods for design science. (Hevner et al., 2004, p. 86)

3.2 Proposed methods

This project primarily relies on the design science methods. Iterative design and software development allows progress to be evaluated underway. This allows time for changes to address concerns and requirements from the pathology department as they surface. This helps to ensure the end result is as cohesive as possible.

For evaluation multiple processes were utilized in this project according to the criteria seen in figure 3.2. Some of these evaluations are done multiple times during the iterative phases of the project such as interface and feature evaluations. These evaluations often take the form of discussion and specific feature testing of a component. Other types of evaluations are done at the end of the project such as performance evaluations, these evaluations were done through experiments both in the pathology lab and simulations of the lab environment.

Real world performance metrics also had to be gathered to allow for direct comparisons of performance between the artifact and the current solution. To gather this data natural observation using unobtrusive methods have been employed to gather empirical data. For more quantitative data a series of interviews and discussions were performed to understand the processes and challenges faced in the pathology lab.

There have been multiple sessions to gather feedback for evaluation through the course of this project. Certain parts of this theses have been submitted as a scientific paper to the 3rd International Health Data Workshop (HEDA) 2023. The HEDA submitted paper titled ”Intelligent Tracing and Process Improvement of Pathology workflows using Character Recognition” was also written by the supervisor and both co-supervisors of this thesis. Feedback and peer review of the HEDA submitted paper will not be public at the time of the thesis submission.

To gather data and better understand the processes at the pathology lab 7 meetings either digital or in person have taken place with representatives of the pathology lab over the course of this thesis. Some of these meetings included a tour of the pathology laboratory facilities, interviews, in lab trials, progress reports on the state of the project and feedback for the project. In addition to the feedback gathered from the pathology lab the contents of the thesis have been presented at a meeting for the Intelligent Information Systems (I2S) research group at Bergen University as well as a poster presentation for upcoming master students and staff at UiB. All of these occasions were used as opportunities to gather feedback and evaluations from domain experts.

The contribution of this thesis is based on the main artifact created which is the prototype of an image processing application capable of specimen tracking by locating, identifying, extracting, and storing data from cassettes. This artifact consists of two smaller artifacts. Mainly a REST API artifact which is responsible for processing an image. This API locates and extracts a cassette location and the text on the cassette. The API is tested using two different methods of cassette localisation these two methods are edge detection and EAST. The results of these two different methods are compared and evaluated in this thesis.

The REST API artifact is not capable of solving the problem of cassette localisation on

its own as it does not recognize 100% of all cassettes and data. To solve this manual error correction is required. The output from the REST API is passed on to another artifact, the graphical user interface (GUI) artifact. The GUI artifact gives the user an overview of the output of the REST API artifact and allows for manual error correction for the API output. This ensure the data is correct before being saved.

The combination of the REST API artifact and GUI artifact results in the prototype artifact evaluated in this thesis. To compare the results from the prototype to the current state in the pathology lab, data had to be gathered from the pathology lab. This was done through an onsite test gathering data by using unobtrusive natural observation techniques. To generate data from the prototype artifact an offline test was setup where pre-captured images were processed with the API and then manual error correction took place in the GUI. The data generated in the offline test is then compared with the onsite data from the natural observation to evaluate the efficiency of the prototype artifact compared to the current solution used in the pathology lab.

In addition to the main artifact capable of specimen tracking the thesis also covers how this artifact could be implemented in the pathology lab environment by proposing a conceptual design. This design allows the pathology lab to see how the main artifact would interface with process execution engines to create fine grained event logs. These logs can then be used for process mining and simulations to further improve the workflow processes in the lab. An overall architectural design is proposed which explains how these processes would work together to achieve these results.

Chapter 4

Methods

This part of the thesis will cover what was actually done, how it was done as well as some of the choices made during creation, implementation and evaluation of the application. The end of this chapter will show how the artifact can be implemented into to current pathology lab structure.

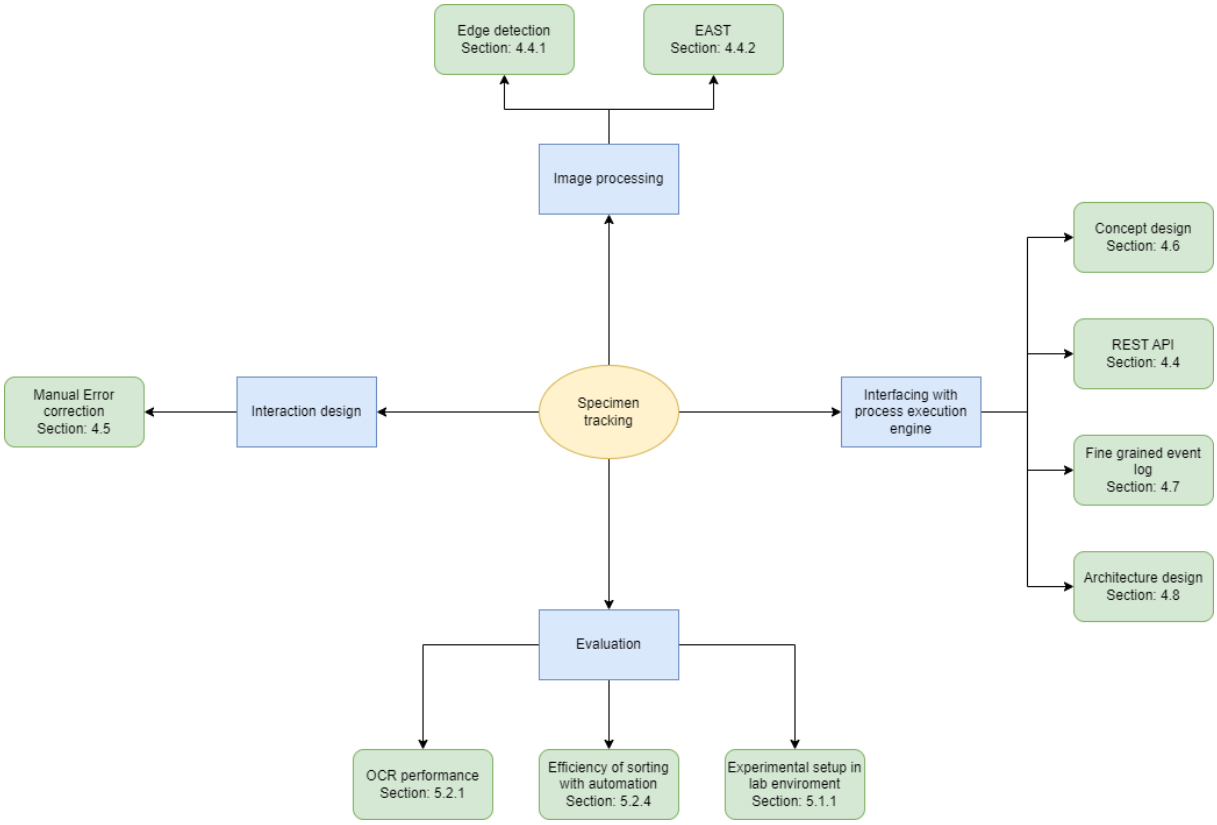


Figure 4.1: This mind map shows the different areas and corresponding sections this thesis touches upon

The contribution of this thesis spans a wide variety of topics but can be divided into four main categories. As seen in the mind map in figure 4.1 these four main categories are image processing, interaction design, interfacing with process execution engine, and evaluation. All of these four categories can also be split into subsections. The mind map shows which sections of the thesis the relevant subsections are covered in.

4.1 Process and procedures

The workload for the artifact development was divided and organized according to agile methods. Firstly, there was a research phase. Here the work was focused on previous research. What had been done before both in the medical domain as well as other domains which potentially utilized similar applications were considered. Goals were set based on what was considered feasible to achieve. After this phase, two milestones were set for the project. The first milestone was be to create a workable solution, the minimal viable product (MVP). This MVP included a preview of the user interface and some basic functionality. The goal of this prototype was to gather feedback for the second iteration which would be presented in the second milestone. The second milestone is where the goal was to incorporate domain knowledge, work on the models, and adjust performance. This is also the phase where most of the actual research was done. The final step after completing both milestones was the evaluation of the artifact.

4.2 Implementation

Initially the work on the artifact was started by creating a prototype of the user interface, this was done by pen and paper. As it is quick to create and gave an idea of what to work towards. Once a good design along with a list of features which needed to be included was created the project moved on to the design and programming of the actual application. The creation of a preliminary model signified the end of the first milestone. This model worked more like a proof of concept. It had the ability to read the text on some cassettes, but the OCR engine took an entire image as an input thereby all the noise between the cassettes were included. The user interface had the minimum viable features. At this point a user could only see cassettes, their data and

the text. The user also had the ability to “draw” on new cassettes. This prototype was then showcased in a meeting with the pathology department for feedback.

After showing the progress in the pathology lab and gathering some feedback for features and getting a general sentiment work started towards the second milestone. The second revision came with improved features both in terms of the backend and frontend. For the backend there were one major change. The image was now being split up into multiple smaller images of each cassette, instead of featuring all cassettes in the same image. The goal was to not feed tesseract with one image containing 40 cassettes, but rather feeding tesseract with 40 images containing 1 cassette each. To do so both edge detection and EAST were implemented. In addition to the inclusion of these methods, improvements were made to the image pre-processing pipeline. On the user interface side some new features were added. These include search, highlighting and a better display of how many cassettes were located. Additionally, a way to define where the cassettes are physically located or where in storage they were about to be placed had now been added. There were also a few minor tweaks and some visual changes. Most of the improvements to these features were inspired by the needs of the pathology lab after a meeting and improve based on the results. Once this was completed the second milestone had been reached and the solution had to be evaluated.

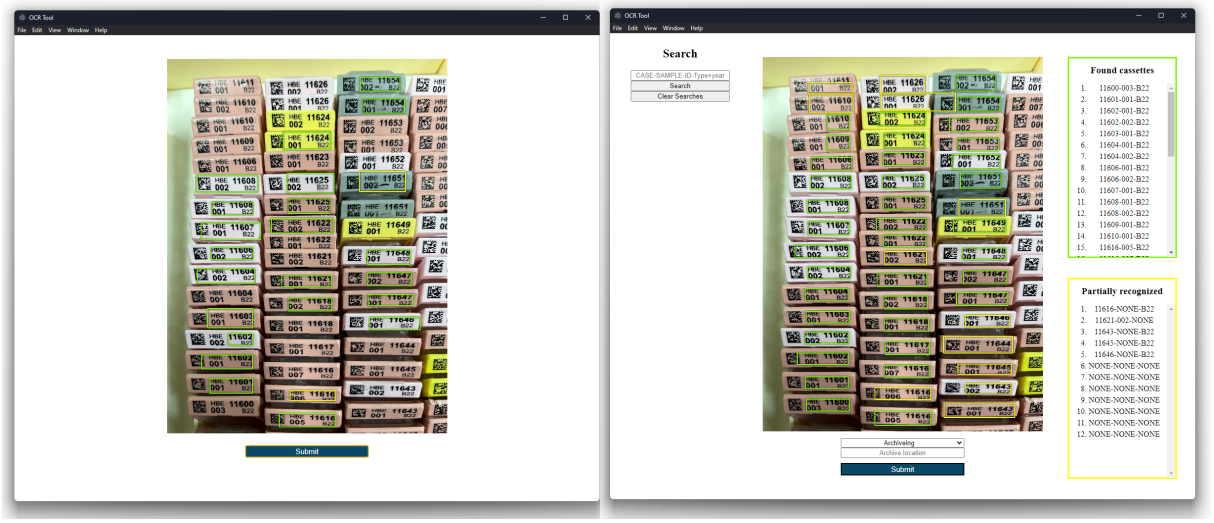


Figure 4.2: This figure shows the differences in the frontend for milestone 1 (left) and milestone 2 (left)

4.3 Development challenges

4.3.1 Why not integrate with the LIS?

Before the development of the system one of the decisions that had to be made was if the application should communicate with the laboratory information system (LIS). The LIS is what connects the patient data to the cassettes, as established the LIS also tracks cassettes through part of the lab. In *Henricks* (2015) Hendricks introduce the concepts of a LIS and goes though the architecture that supports the system. “An LIS is at its core a database” (*Henricks*, 2015, p. 4). Connecting this with the proposed system would give some advantages. Namely the proposed system would know how many cassettes to expect per case and additional data could be retrieved from the data matrix code. However, integrating with the LIS also comes at a cost. Firstly, gaining security clearance to it would introduce an additional challenge as clearance is highly regulated on a need to access basis as seen in the earlier work done by Stünkel et al. (*Stünkel et al.*, 2022, p. 6). If the LIS and the OCR system was linked the OCR system would now technically be responsible for customer data directly which would have put very strict regulations related to data storage and usage due to regulations such as GDPR. These regulations would be especially strict considering the data stored in the LIS is very sensitive healthcare data. Currently the data manged by the OCR system is of no value on their own as they are just arbitrary numbers. However, these arbitrary numbers once linked with the LIS would enable someone to tie a cassette to an individual patient which would compromise their healthcare data. With these restrictions in mind the author and representatives from the pathology lab decided to test a standalone system with no connections to the patient database. Essentially creating a new specialized LIS as detailed by Dangott.

Dangott presents the general characteristics of a specialized LIS which performs a critical function such as interfacing with equipment that a traditional LIS does not support. This leads to enhancements in operation such as improved turnaround time, specimen tracking, enhanced reports, diagnostic data representation, correlation with previous results, etc. *Dangott* (2015). These characteristics lines up with the solution proposed in this thesis. The current LIS has no way of interfacing with the cassettes

by reading the semantic text on top of the cassettes, and neither does it have a module to organize the archive and locate cassettes. The results of the system should improve turnaround time and reduce the amount of time spent on sorting cassettes manually.

The major drawback of not having access to the LIS is that the system cannot utilize the data matrix codes on the cassettes. Data matrix codes are also not flawless, and errors can occur. This can be seen in the research done by Hanna and Pantanowitz. Misread failures can lead to one barcode being read as another if a defaced coded string represents another valid barcode. This type of defacement is likely to occur in pathology labs where the barcodes are exposed to blades, heat, harsh chemicals, and microwaves *Hanna and Pantanowitz* (2015). The 2-D data matrix codes utilized at the pathology lab at Haukeland have some capacity of inbuilt redundancy, but it is not impossible for errors to occur. Departments at Haukeland university hospital have had issues with medicine mix-up in the past. As recently as 2018 the medicines kaliumklorid and calciumklorid were switched around due to a human read-error causing the death of a patient *Otterlei* (2019).

The LIS also has a problem with certain data points missing. This is due to the time it takes to scan all barcodes by hand. Scanning is a time-consuming process and cannot be done at all stages of sample processing due to the time required. The proposed solution would be able to fill inn these missing data points if the scanning is faster compared to barcode scanning. If the data generated by the OCR program is collected and merged with the data from the LIS, the system would be able to replace the current barcode scanning in three out of four distinct places. These three being the archive, processing, and coordination. Allowing for cassettes to pass though the lab faster as hand scanning them individually would be replaced by scanning trays of multiple cassettes in batches.

4.3.2 Pre-Processing challenges

There are two main challenges the proposed application faces. The first is individual cassette detection, the second part is the OCR itself. This thesis mostly focuses on cassette detection, as that case is unique to the pathology department. OCR has many

use cases in practice as seen in earlier examples. However, there is one large difference between the examples shown earlier and the solution attempted in the pathology lab. Both licence plate detection and text detection on mailed letters only feature one area-of-interest per image, the images from the pathology lab features multiple fields of text which needs to be segmented before it can be read by tesseract. If an image containing all cassettes as tested in milestone 1 performs badly as tesseract is not able to distinguish which text belongs together. The way the text is placed on the cassettes also creates some difficulty as each text field is in a separate corner and the text is spread out, this means text and paragraph detection algorithms would perform inaccurately due to the distance between the text that corresponds with each cassette. The real challenge with the OCR system for the pathology lab is to detect the cassettes individually first and segmenting each cassette into an individual image. Once that is done more traditional OCR methods and pre-processing can be applied to each cassette/image individually for greater results.

4.3.3 Challenges with OCR in this domain

When using OCR there are a few challenges and possible hurdles that need to be kept in mind. A sampled of 16 Pictures captured with a phone from various angles and camera distances in the pathology lab was provided. These images represented the conditions found in the lab. In an ideal situation the images should have been taken by mounted cameras. Making a camera mount would allow for a fixed angle and minimal motion blur being caused by unsteady hands which the images taken by mobile phones suffer from. Ideally a high-resolution camera would perform better we lacked equipment. After inspecting the images four specific elements were immediately observed as potential causes of issues. These four issues are text being too small to read (blur), obscured text, skewed text, and light reflections.

Figure 4.3: Shows a blurred cassette

If an image is taken too far away the text becomes too small and blurry for the OCR tools (and humans) to read. To fix this the image need to be taken closer to the trays.

Figure 4.4: Shows an obscured cassette

Cassettes sometime contains partially or in some cases completely obscured data from elements such as other cassettes, as can be seen image 4.4 is taken at an angle leading to the block ID and the type id being obscured. This can be fixed by positioning the camera further above the cassettes and would be easier with a mounted camera setup.

Figure 4.5: Shows a cassette with reflected text

Reflections from the lab lighting conditions can hide the cassette text and make it harder for the OCR tools to identify the text. This issue is caused by the lighting conditions in the lab. Finding an area with possibilities of controlling lighting and reducing sharp lights would alleviate this issue.

Figure 4.6: Shows a with cassette with skewed text

Text can become more and more skewed depending on the angle of cassettes and the angle the image is taken at. There seems to be two main causes of this issues. The first being the angle the image is taken at. As mentioned, having a custom camera mount would help with this. Second is the way cassettes are stacked in the lab. In the image above you can see that the cassette at the top is more readable than the cassettes in the lower part of the image. The cause of this is cassettes not being stacked properly together in the trays. This can be fixed by lab workers moving the cassettes a little closer together.

There are a few different measures that the pathology lab could implement to alleviate or remove some of these issues. The issues with obscured text could be fixed by packing cassettes closer together on a tray, making sure they are always standing the right way up and facing the same direction, this in addition to mounting a camera at the correct angle and distance would produce better quality images with less blur and less skewed text. The reflections could be lessened by having a designated area with stricter light control for image taking or an area with more constant light. Both of these measures are possible to implement in the lab. A camera with higher resolution than 3024x4032 which is the resolution of the images used in this thesis could also produce better results as the text would be easier for Tesseract to extract. A few measures could also be taken in software such as improvements to the pre-processing pipeline as well as specific training of a custom neural network, these will be discussed later.

4.4 Software details and choices

The software developed for the pathology lab had a few different goals. After talking to the staff at the lab some of these were established. Multiple options were considered, such as an application for phones, but after talking to the lab and taking their preferences into consideration and looking at OCR challenges with capturing images it was settled to develop the program as a windows application. The software needed the ability to take input images or capture images, count the number of cassettes, locate them, and read the text labels, with minimal user input required. When considering which programming language to use a few factors stood out. The author's previous experience, as choosing something the author was more familiar with would

save time. Speed and suitability for the task of image processing and OCR also had to be considered. As well as availability of tools such as libraries and documentation related to the task. After taking this into account the remaining languages for the image processing part consisted of either C++ or Python. Python was chosen as the author have previous experience with it for the sake of development speed. Python also has a high standard for images processing with libraries such as OpenCV. OpenCV is already extensively documented and easy to work with as it has ”more than 47 thousand people of user community and estimated number of downloads exceeding 18 million. The library is used extensively in companies, research groups and by governmental bodies.” (OpenCV, 2022). Python also have libraries available to allow us to easily connect it with an output engine such as pyTesseract, or API based OCR tools like Azure OCR.

When choosing OCR engine multiple factors came into consideration. As mentioned above there were multiple different choices to consider. In the end Tesseract was chosen for various reasons. Firstly, it has its own python library accessible which works natively and is therefore easy to setup. Tesseract is free to use and download, this means we can run it locally on our machine and not be worried about a Wi-Fi connection as well as dealing with time-outs and error messages that could be part of an online solution.

With python running in the backend, the application needed way to display information to the user and allow them to interact with the application in a more user-friendly manner than a terminal. Python have many libraries that allows for user-interface development, but from earlier experience with some of them none of these covered the baselines which the project aimed for. Instead, the choice was made to opt for using Electron. Electron ”is a framework for building desktop applications using JavaScript, HTML, and CSS. By embedding Chromium and Node.js into its binary, Electron allows you to maintain one JavaScript codebase and create cross-platform apps that work on Windows, macOS, and Linux” (Electron, 2023). This allowed development to be done using native JavaScript (JS), node JS and related JS libraries for the code related to the user experience. With this technology the structure of the the user interface could be created using html and CSS. These technologies are used in the

majority of the web with javascript being utilized for 98.5% of all websites as of March 1st 2023 (w3techs.com, 2023). This allowed for quick development due to the well documented frameworks and wealth of information easily accessible. Using Electron also lets the program run on a local machine eliminating requirement of running an online server and allowing for a local server instead, giving more options on how to setup the program depending on the pathology labs needs.

For python and Electron to communicate an internal API would allow for fast communication. For this flask was chosen as it is easy to use, fast to setup and meet all requirements for this project. As flask is a web framework written to be an extensible ”micro-framework”. It acts as a solid core for other services and allows to integrate with any database or components by using extensions. This lets us tailor our own networking stack. (*Grinberg*, 2018, p. 1). The local flask server starts and stops itself with the electron application and allows data to travel back and forth between the frontend and backend.

The API used for frontend and backend communication could also be hosted on a server as a REST API to allow the OCR tool to serve multiple clients at once without running locally. As of the current iteration the API set to handle images from local storage. However, for an implementation in the pathology lab the API can be possible to call by sending a link to an image over hypertext transfer protocol (HTTP) to a uniform resource identifier (URI) pointing to the API. The API returns a JSON object with the data generated by the program acting as a REST API. This works in the same way as the current backend to frontend communication, but instead of a local file path as a parameter the input parameter would be a link to the image that should be processed. This REST API would then be usable by multiple thin clients allowing for all the heavy processing to take place on one central server which communicates with the different clients. This would end up saving the pathology lab money in the amount of processing heavy clients required and make the API more easily accessible from different workstations.

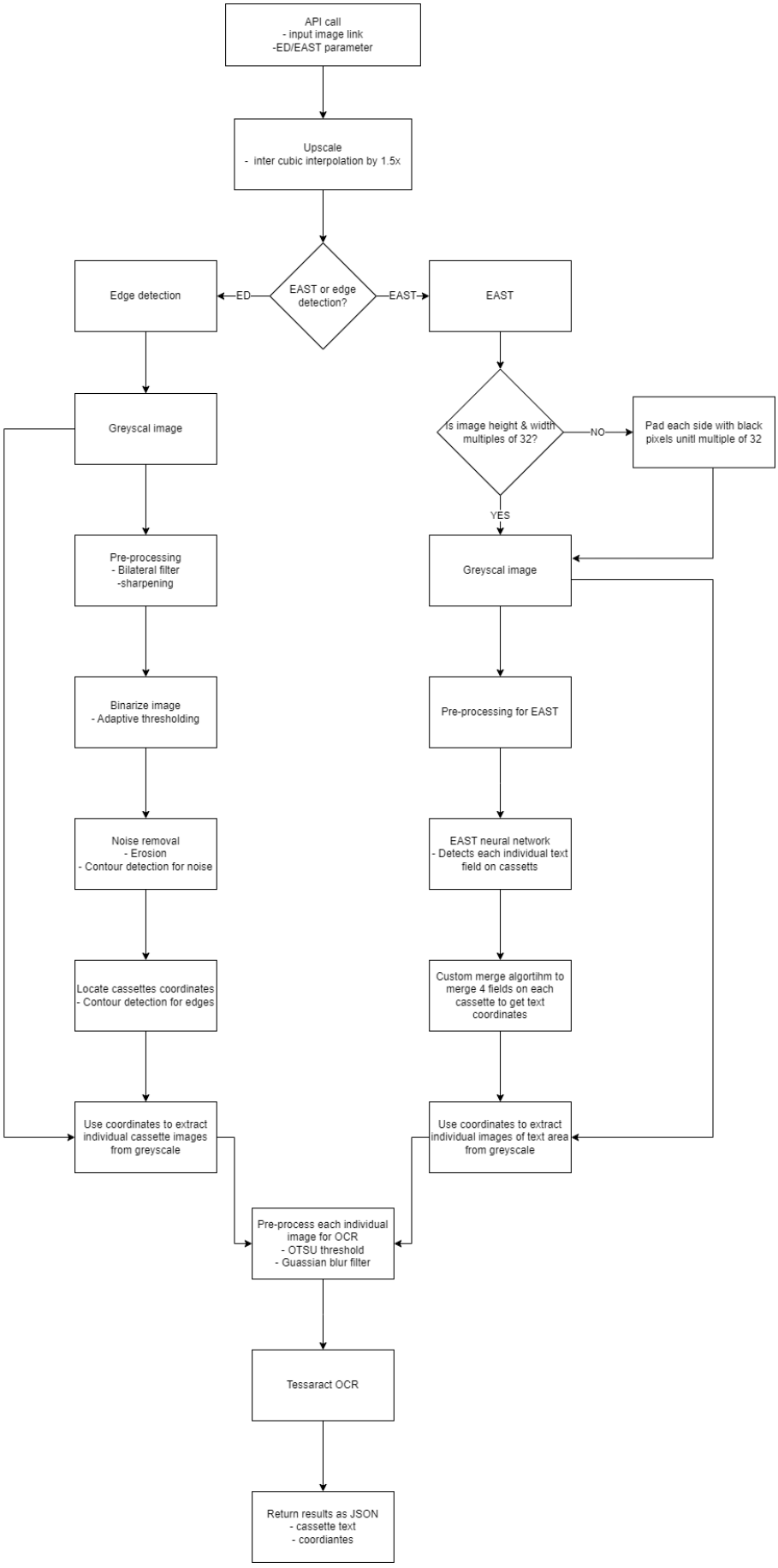


Figure 4.7: This figure shows the pipeline for the rest API that processes images

One of the largest challenges with the OCR engines is pre-processing the input. When doing so the program is attempting to do multiple things in one algorithm, these include noise removal, image correction and binarization (*Bieniecki et al.*, 2007, p. 1). Here different factors and image condition can yield different results in similar images. Two different options of pre-processing the images were utilized in this thesis. One more successful than the other, but the heavier processing algorithm also comes with higher computing requirements thus leading to a drastic increase in runtime.

Before doing any pre-processing, the program increases the resolution of the images, this would not be needed if the images were taken by high resolution cameras. However, as tesseract prefers an input of at least 300DPI (Tesseract - GitHub.com, 2023). The program upscale the images which are all 3024x4032 by 1,5 using inter cubic interpolation.

Tesseract works better when there is less text and noise in the image (Tesseract - GitHub.com, 2023), for noise removal two different methods were attempted. The first method tried to remove and filter as much as possible. The algorithm first attempts to use edge detection to detect and split each cassette into a separate image. The other alternative involved using a neural network called ”An Efficient and Accurate Scene Text Detector” or EAST detector for short. EAST attempts to locate the text on each cassette, a custom written algorithm for this project then merge the text on each individual cassette to locate them. Essentially, the program wants to take cassettes from an image containing 100 cassettes and instead have 100 images containing one cassette each. Edge detection allows us to see the results when attempting to capture an entire cassette, this is interesting to when comparing the results to only capturing area of interest, which in this case is the text on the cassettes. Edge detection is implemented through conventional pre-processing methods whereas text recognition is only possible with neural methods.

4.4.1 Edge Detection

Figure 4.8: This figure shows an image processed by edge detection

For the edge detection approach the image is first converted into grayscale, then various pre-processing effects such as a bilateral blur filter and sharpening is applied. The blurring smooths out some noise and the sharpening helps increase the outline of the edges the software is attempting to detect. The image is then thresholded using adaptive threshold, this binarizes the pixels in the image making it consist of only black and white pixels. Erosion is applied to remove noise and contour detection is applied, where the smallest contours are removed again to decrease the amount of noise in the image. Contour detection is then applied to the image to detect the remaining cassettes. The coordinates are then extracted, and each cassette is cut-out from the grey image.

The pre-processing done in edge detection needs to be tuned for specific distances and sizes of cassettes, creating a “catch-all” is incredibly difficult. Therefore, edge detection works significantly better on images with the same distance and cassette sizes as the images it is tuned for and performs much worse in images the more the

camera moves away from or closer to the cassettes.

4.4.2 EAST

For the EAST segmented cassettes, a slightly different approach is at play. Firstly, the images need to be padded as EAST only works on images with a resolution in a multiple of 32. To do so black pixels are added to the right and the bottom of an image until it fits the resolution requirements. The image is then prepared and input into EAST according to the pipeline the neural network expects *Zhou et al.* (2017). The output is parsed until the vertices surrounding each part of text is returned. An issue arises as each cassettes contain 4 pieces of text, and these four boxes needs to be merged for the text on each cassette to be joined in a segment as illustrated below.

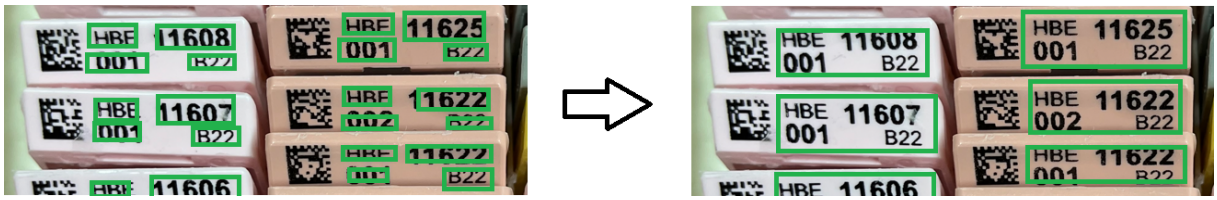


Figure 4.9: This figure shows how the algorithm merges the results from EAST to capture all the data on one cassette

To do so a custom algorithm was written, this algorithm loops over the boxes starting in the bottom left of the image and using a merge threshold it merges with the boxes to the right and above them as long as they are not outside of a certain distance or above certain size. The distance between boxes is decided by the average size divided by 3 or multiplied by 1.5 depending on if it is height or width respectively. The max size of a box is width and height multiples of 8 and 4 accordingly. This allows the merge algorithm to work on images with different sizes of boxes as the algorithm will scale the merging accordingly, and image resolution will be unaffected. The algorithm repeats itself until all boxes are either too large to be merged and none are overlapping. The coordinates surrounding the text is then extracted and cut out from a grayscale image. Processing images with EAST is considerably more compute heavy and takes more time than the edge detection approach, but it should give much better results.

4.4.3 OCR

Both the grayscale images from the EAST-text detector and the images cut-out using cassette edge detection are then processed for OCR. This process thresholds each individual image using OTSU thresholding which attempts to automatically split the greyscale images into foreground and background *Otsu* (1979). A mask is then created from the contours of the thresholded image. This mask is used to filter out noise. The last thing to be done to the image before OCR is a layer of Gaussian blur which is applied to smooth the image for the OCR process and remove speckle noise. When applying Gaussian blur ”the average value of the surrounding pixel or neighbouring pixels replaces the noisy pixel present in the image which is based on Gaussian distribution.” *Kumar and Nachamai* (2017).

Edge detection faces some additional challenges compared to EAST in this preprocessing step. This comes in the form of multiple cassettes being detected as one cassette if they are in a very close proximity. This effect can be seen the two first ”cassettes” on the second column in figure 4.8.

Each image is then individually fed to Tesseract for OCR, the coordinates and text is extracted and returned to the user interface Electron application. Here the information is presented to the user. The user can then confirm the information and help label the boxes that tesseract is not able to read. Before submitting the final results to the database.

4.5 User interface

This section will detail a few of the features in the graphical user interface (GUI) created for the application. As a user interface has also been developed for the application. This interface allows for searching and highlighting of individual cassettes, easily displays the information and allows a user to label cassettes not recognized by the OCR. The user interface also supports drawing bounding boxes around cassettes using the mouse cursor. These cassettes can be drawn onto the image in case the cassette was not detected by the edge detection or EAST. The user can save the data to a file or database which will allow for easy retrieval and searching. The data exported in-

cludes the unique key of each cassette, this consists of case ID, block ID and sample type. This information is also accessible in its own rows. Additional data generated by the program includes rough cassette location in image and tray, which is split into a 3x3 grid as seen in the figure below this allows quick location of specific cassettes without having to look through all of them. Additionally, local time, local date and an ISO timestamp is saved. And finally, cassette status, so whether it is in a machine or on the way to the archive if so which machine and exact location in the archive the cassettes are located in will be stored as well.

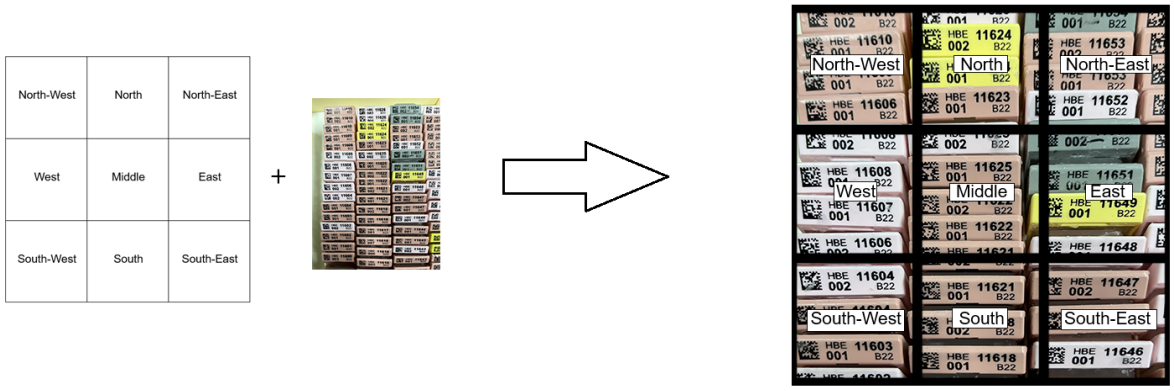


Figure 4.10: This figure represents how the data of cassette location is converted and stored into plain-text

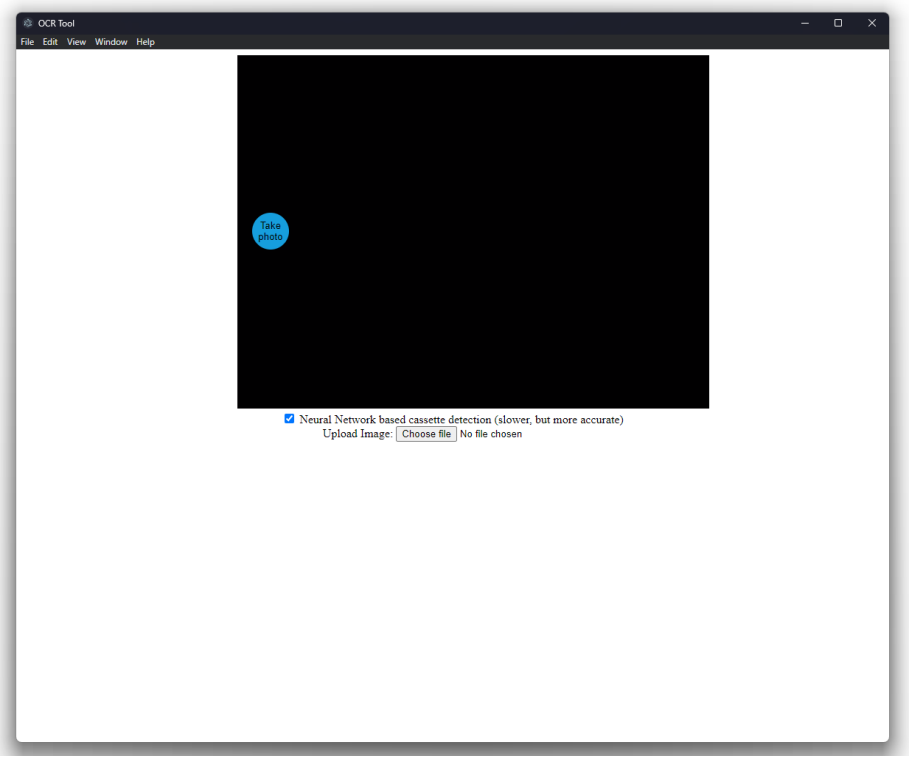


Figure 4.11: The start page on the application

When the application is launched users are greeted with a page that lets the user capture an image or upload an image file. The black box in figure 4.11 will show a live camera feed if a webcam is connected. When an image is uploaded or captured the user will be moved an animated loading screen whilst the cassette location and OCR is processed in the background. Once the process is done the user will be presented with an overview of all the cassettes as seen in figure 4.12

For this application 3 different colors were utilized as seen in figure 4.13. These colors were chosen based on a variety of factors. Green is often considered the ”ok” color and used in cases where all the information on a cassette is successfully extracted. The orange-red was chosen for the search functionality as it stands out and is easy to locate. The yellow color also stands out and is easy to locate among the cassettes, this color signifies that a cassette is missing certain information. For an actual implementation of the system other colors should be considered as the red and green colors could cause problems for a user suffering from deuteranomaly which is a type of color blindness affecting a user’s ability too differ those two colors. The colors used can also in certain cases blend with the cassettes in the image. Especially the orange-red

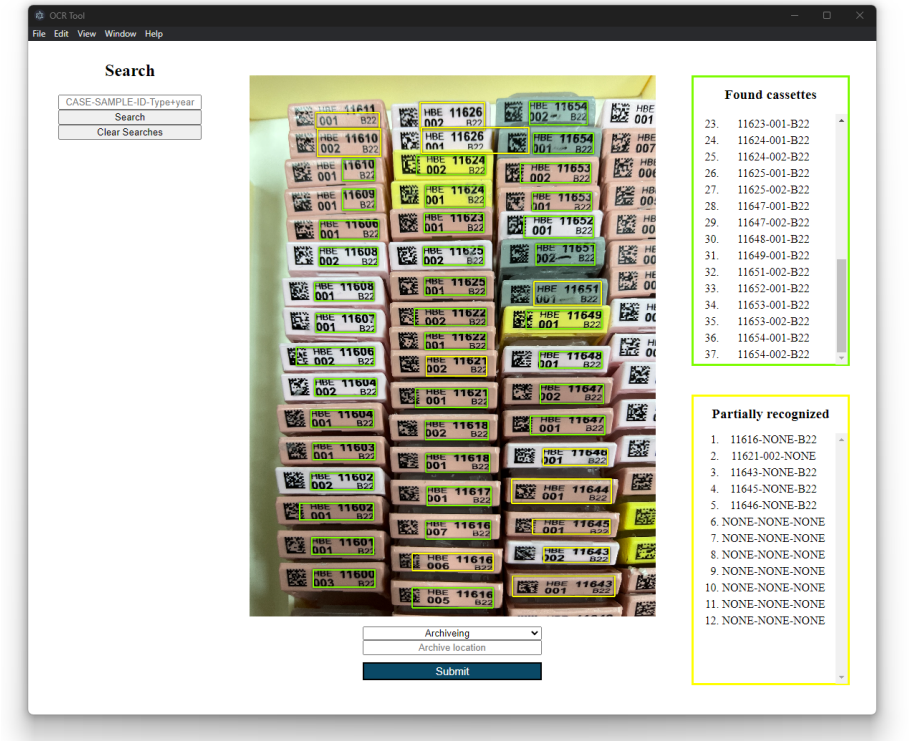


Figure 4.12: This image shows the final user interface and the results from a scan.

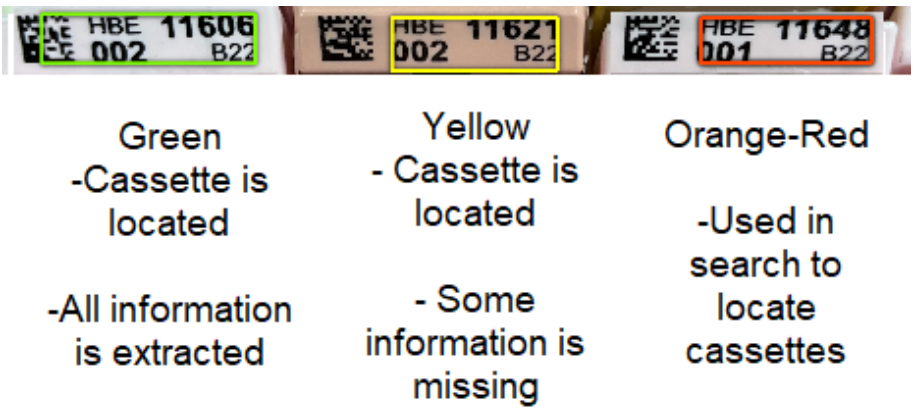


Figure 4.13: The different colors used in the GUI

and the orange cassettes. Research could be done to identify colors that are easy to spot mixed in with the colors already existing on the cassettes. There could also be potentials to find an alternative method to highlight the cassettes. However, this type of interface would not be needed in a fully automatic setting.

The green and yellow colors are also used in the sidebar located on the right side as seen in figure 4.12. The green box at the top lists all cassettes found with information successfully extracted by the OCR process. The yellow box list all cassettes located and in need of some manual data entry. Both boxes are sorted in numerical order and have a counter displaying the amount of cassettes. The boxes are updated dynamically whenever a user makes changes to any cassette data.

On the top left in figure 4.12 the search function is located. This allows the user to highlight specific cassettes by either searching for the text data the cassette contains or by clicking on a cassette in the green or yellow preview box on the right side of the image. To identify all cassettes missing data the user can search for ”none”.

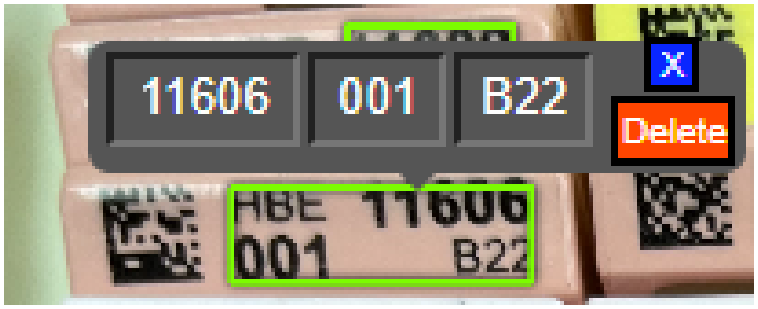


Figure 4.14: Shows the information that appears when clicking on a cassette

To change or add information to a cassette the user can click on the colored boxes seen in figure 4.13. This action will bring up a menu as seen in figure 4.14. Here the user has the ability to change the Case ID, Block ID and the type and year identifier. The data edited will not be accepted unless in a valid format. Saving the wrong cassette data will highlight the text filed containing wrong data in red. The boxes automatically turn from yellow to green once the issues are resolved. Colored boxes can also be delete in case they are in wrong location. To draw new colored boxes the user can click and hold on the image. The user can then surround an area containing a cassette and manually enter the information an empty text box, the manual entered data is also checked for validity.

Once a user is done, they can submit the data. The submit button is not available to press as long as there are still cassettes that require corrections. Once there are no more yellow cassettes the user needs to select which activity the cassettes are at. In the case of the archive the user must manually enter a storage location into a text field. Once these conditions are fulfilled the user can submit the data. This saves the results and bring the user back to the start page seen in figure 4.11 to scan another image of cassettes.

The code and installation instructions for the program including processing pipelines and user interface can be found in the appendix, see chapter 6.

4.6 Suggested framework

The cassette tracking process should ideally not interfere with the current workflow in the lab. Therefore, a framework must strive to be as non-intrusive as possible to avoid disrupting the current workflow. Ideally it would be light-weigh and should not include too many steps which adds to the process. The light-weight framework suggested here takes all these factors into consideration whilst aiming to enhance the traceability of cassettes.

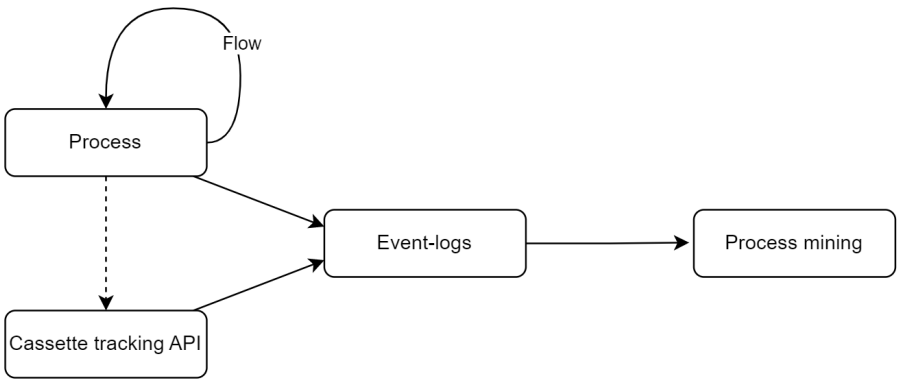


Figure 4.15: conceptual framework

The pathology lab requires access to cassette data on an individual level. The current system in place skips tracking certain processes due to the time constraints of scanning. Each cassette is not tracked in each individual activity leading to a loss of data when applying process mining techniques. The data generated by the proposed system would allow the pathology lab to use process mining to achieve a better overview

of the workflows in the pathology lab. As the new system will allow individual cassette tracking though all activities where tracking is currently missing. This is a requirement for process mining techniques and will allow for process mining on a granular level.

The concept framework seen in figure 4.15 is proposed for the pathology lab. It allows the cassette tracking to work in tandem with the processes in the lab, but at the same time both processes work independent from each other. Both the lab process and the individual cassette tracking will output event-logs. Using a combination of these will allow for fine-grained processing of the workflow. This could potentially lead to discoveries being made by the team working on process mining the pathology lab and further help gain insight into the processes and workflows.

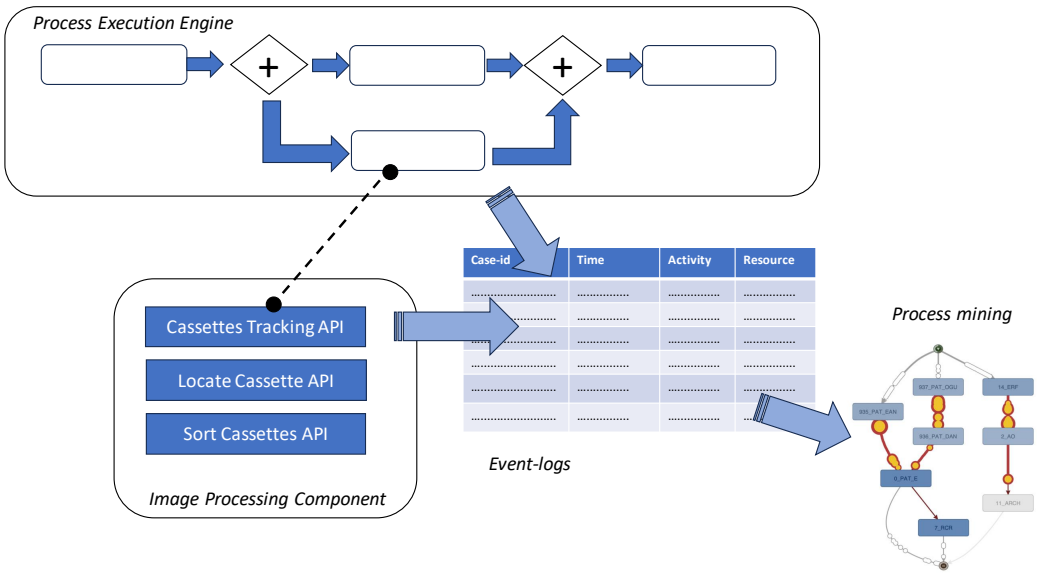


Figure 4.16: Conceptual framework for enhanced process monitoring in pathology laboratories

Figure 4.16 shows how the process execution in the pathology lab could be carried out. This will lead to enhanced process monitoring which will add cassette location and information to the event logs.

4.7 Augmenting the logs

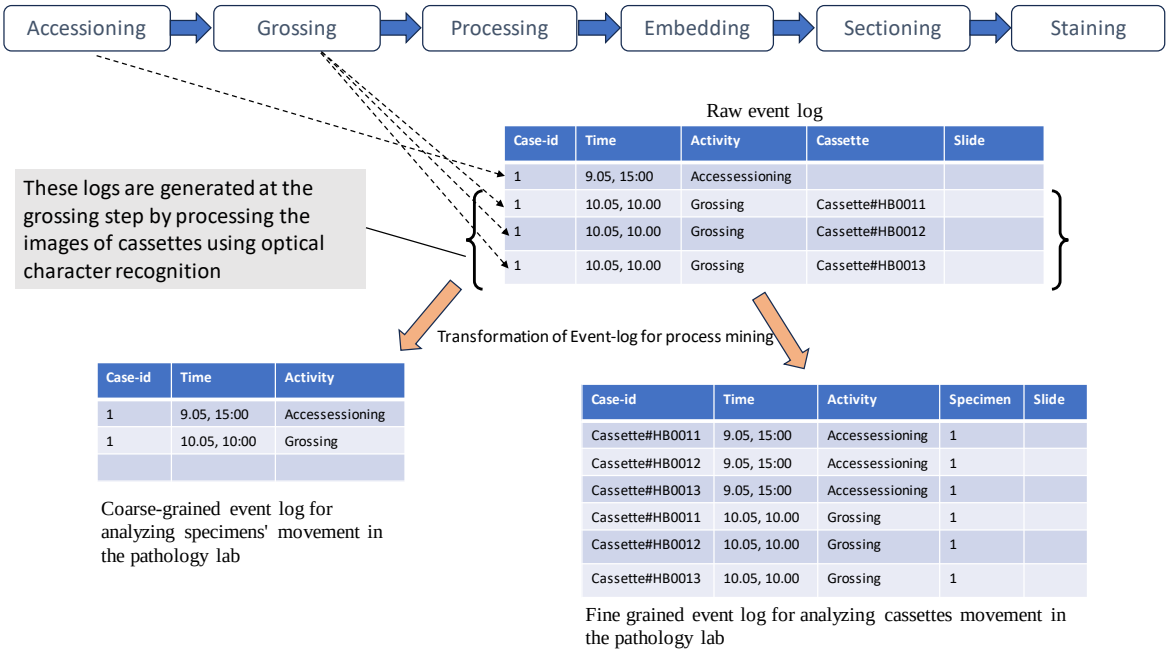
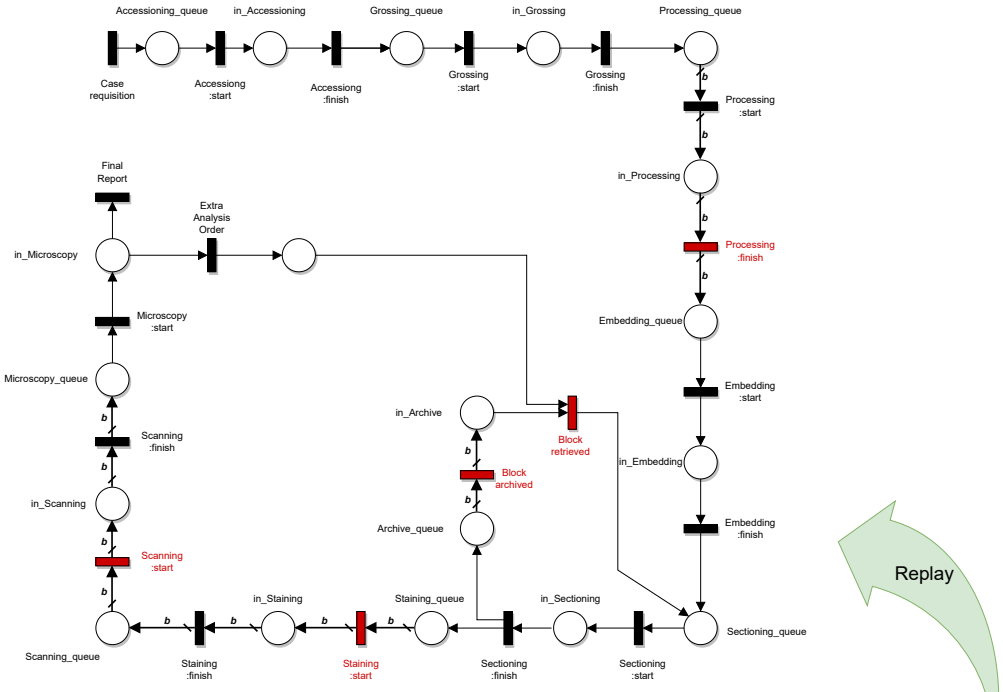


Figure 4.17: Transformation of raw logs for coarse-grained and fine-grained process analysis

Figure 4.17 is an example of what data output from the application could look like in conjunction with the data from the LIS. The figure shows how the data can be merged. With this merged data the pathology lab would be able to utilize process mining to gain a better overview of the workflows in the lab than what is available today using only the LIS data.



Case	Timestamp	Event	Token	
...	...	...		
B23 12345	2023-05-15 09:15:18	Grossing:finished	B23 12345	
B23 12345	2023-05-15 12:01:23	Processing:start	B23 12345 01	
B23 12345	2023-05-15 12:01:28	Processing:start	B23 12345 02	2023-05-16 01:00:00; Processing:end; B23 12345 01 2023-05-16 01:00:00; Processing:end; B23 12345 02
B23 12345	2023-05-16 08:00:51	Embedding:start	B23 12345 01	
B23 12345	2023-05-16 08:03:10	Embedding:finish	B23 12345 01	
B23 12345	2023-05-16 08:03:10	Embedding:start	B23 12345 02	
B23 12345	2023-05-16 08:05:42	Embedding:finish	B23 12345 02	
B23 12345	2023-05-16 14:11:11	Sectioning:start	B23 12345 01	
B23 12345	2023-05-16 14:18:01	Sectioning:finish	B23 12345 01	
B23 12345	2023-05-16 14:18:03	Sectioning:start	B23 12345 02	
B23 12345	2023-05-16 14:23:42	Sectioning:finish	B23 12345 02	2023-05-16 15:03:17; Block archived; B23 12345 01 2023-05-16 15:15:01; Block archived; B23 12345 02 2023-05-16 16:06:01; Staining:start; B23 12345 01 01 2023-05-16 16:08:01; Staining:start; B23 12345 02 01
B23 12345	2023-05-16 18:07:01	Staining:finish	B23 12345 01 01	
B23 12345	2023-05-16 18:07:03	Staining:finish	B23 12345 02 01	2023-05-16 10:49:01; Scanning:start; B23 12345 01 01 2023-05-16 10:52:01;Scanning:start; B23 12345 02 01
B23 12345	2023-05-17 10:52:00	Scanning:finish	B23 12345 01 01	
B23 12345	2023-05-17 10:55:00	Scanning:finish	B23 12345 02 01	
B23 12345	2023-05-18 13:01:33	Microscopy:start	B23 12345	
B23 12345	2023-05-18 16:08:02	ExtraAnalysisOrder	B23 12345	2023-05-19 07:45:52; Block retrieved; B23 12345 01
B23 12345	2023-05-19 07:55:10	Sectioning:start	B23 12345 01	
B23 12345	2023-05-19 07:58:55	Sectioning:finish	B23 12345 01	
B23 12345	2023-05-19 12:45:44	Staining:finish	B23 12345 01 02	
...	...	...	...	

Figure 4.18: Shows how the proposed solution augments the workflow from figure 2.1 Hatlem et al. (2023)

The top part of figure 4.18 shows how the solution proposed here is able to augment the current workflow from figure 2.1 and fill in the data currently missing in the LIS, which is displayed in red in the model. The lower part of figure 4.18 shows how the results of combining the LIS logs with the data this system allow us to generate. This will result in a more comprehensive log able to display more of the events happening the pathology lab. This data can be of use when process-mining.

4.8 System architecture

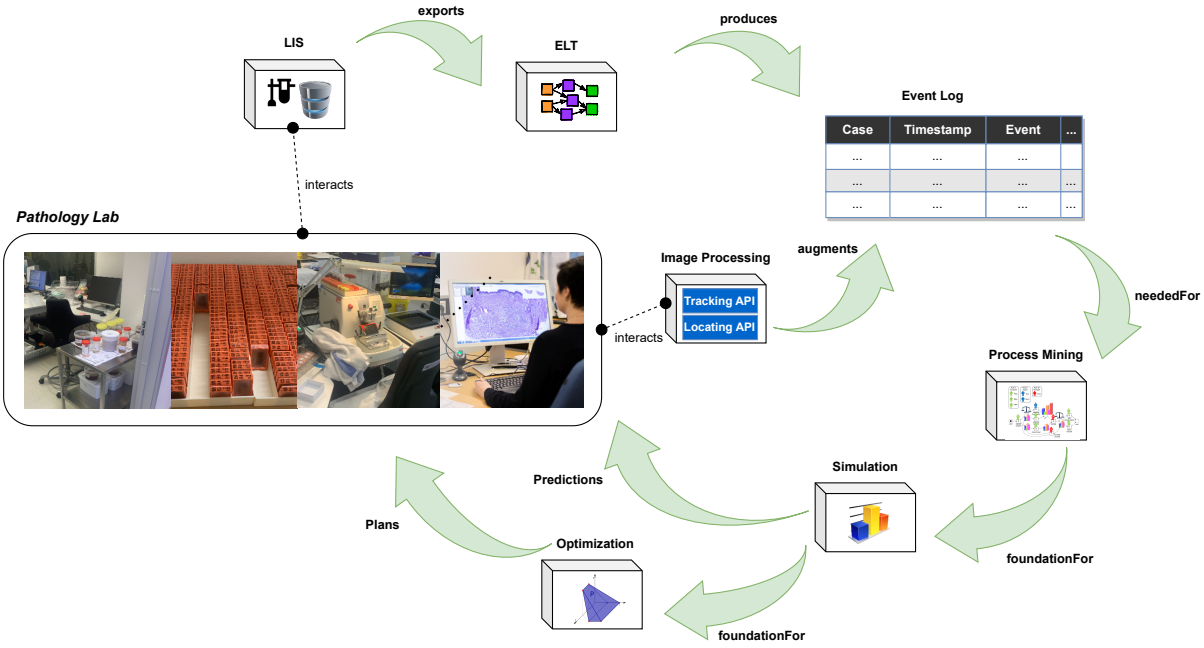


Figure 4.19: The overall architecture of the system Hatlem et al. (2023)

The overall architecture of the system would be as seen in figure 4.19. Data is gathered from the pathology lab by both the LIS and image processing system proposed here. The logs from the LIS are exported and turned into an event log. This log is augmented with the data from the image processing system. This comprehensive event log is then used for process mining. This is fundamental for allowing simulations and predictions which can be used to better understand and optimize the workflows in the lab. The logs from the LIS data and image processing system can then verify these workflow improvements and locate new areas suitable to improve. Creating a self-perpetuating loop.

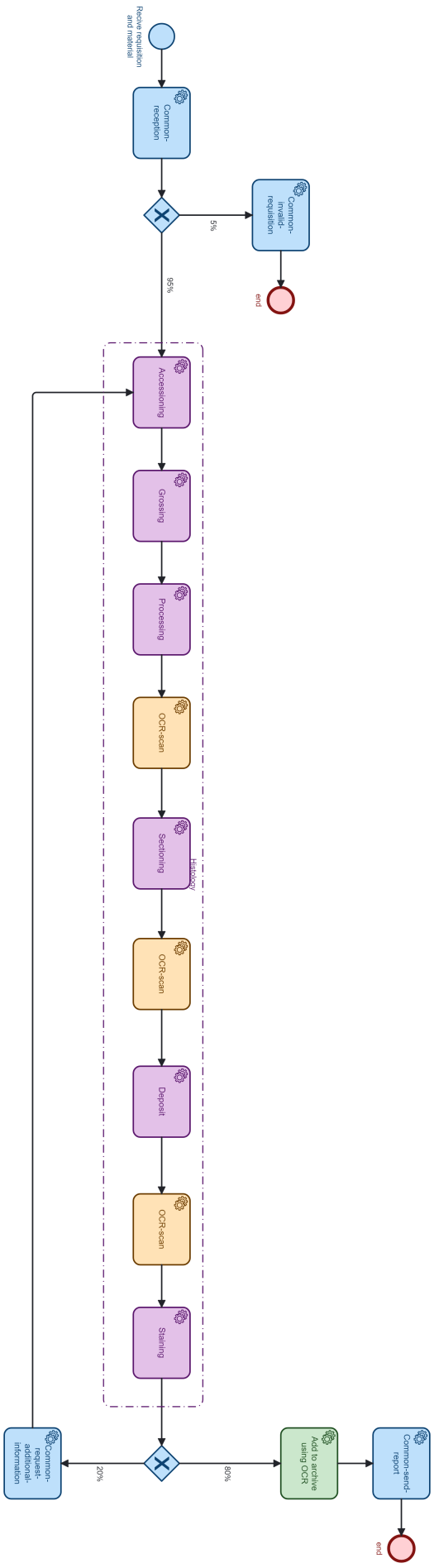


Figure 4.20: BPMN model modified from Platou (2021) to show how the OCR tool would interface with the system

The BMPN model in figure 4.20 is a modified version of the model created by Platou in *Platou* (2021). The model has been slightly simplified by removing the cytology section. The histology section has been modified with new activities added by the OCR-scanning tool. These are color coded by the orange boxes. Activities that are altered are displayed in green, this is only the archive section as instead of a sorting process the manual sorting activity is now replaced by a scanning activity.

Chapter 5

Evaluation

This chapter will show how evaluation was done, summarize the findings, and give an analysis of the results from this study. In addition, an in-depth discussion will be looking at the how sorting assisted by the application compares to doing it manually.

The performance of the OCR algorithm will also be reviewed. Is it able to read numbers consistently? Will it struggle under the conditions of the pathology lab? How many cassettes does it misidentify with the wrong number? Is there a notable difference in time between edge detection and EAST for locating cassettes?

5.1 Results and Experiments

For the experiment data from 4 different images will be compared. These images were chosen because they have as few of the challenging components mentioned earlier i.e. they have little, skew, reflection, and blur. The images are all taken at consistent lengths using an iPhone 13. The images have not been altered in any way except for a small amount of zoom to make cassettes roughly the same size.

table 5.1 and 5.2 show the results from the edge detection and East when locating and identifying cassettes. The same four images were used for testing both methods. The different columns in the table display the following information:

- Fully recognized are all cassettes that are found and all the text is completely

read. An example of a fully recognized cassette would be 11111-001-b23.

- Partially recognized are cassettes that are found and the text is unable to be read or the text is partially read, but not enough to get the full unique ID of the cassette. A partially recognized cassette could be either (NONE representing missing data) 11111-NONE-b2, NONE-NONE-NONE or any combination where a part is unidentified.
- False detections are cases where an area not containing a cassette is highlighted by the system. This can be caused by noise or by false detections.
- OCR errors tells us how many fully recognized cassettes contained errors. An example of this could be a number being read wrong such as 11111 being seen as another similar number like 71111. As it still is a valid number the system does not flag it as incorrect and it needs to be discovered by the user.
- Runtime is displayed in seconds. Runtime shows how long it took the program to run, one thing to keep in mind here is that runtime includes the time it took to both locate and to perform OCR on the cassettes. Hardware does of affect the speed of the runtime, in the case of this program runtime is mostly CPU bound. In the testing done here an AMD Ryzen 3800x with 8-cores running at a 4.5Ghz boost was used.

Table 5.1: Results from edge detection

Image name	Total Cassettes in image	Total Recognized	Fully Recognized	Partially recognized	False detections	OCR Errors	Runtime (seconds)
IMG 0892	37	22	16	6	1	0	6.90
IMG 0894	49	37	19	18	0	1	9.02
IMG 0895	37	32	9	23	3	0	8.48
IMG 0896	44	39	3	36	0	1	8.80

In addition to data gathered from the program real world data was gathered in the pathology lab for comparison. This was gathered through a series of interviews and meetings. to get an idea of how much time and resources is spent sorting cassettes natural observation techniques were used to gather data. The author spent over one hour with a lab technician working in the archive section on a randomly selected day.

Table 5.2: Results from EAST detection neural network

Image name	Total Cassettes in image	Total Recognized	Fully Recognized	Partially recognized	False detections	OCR errors	Runtime (seconds)
IMG 0892	37	35	27	8	1	0	31.44
IMG 0894	49	49	37	12	0	0	35.23
IMG 0895	37	34	18	16	1	1	27.42
IMG 0896	44	38	25	13	0	1	27.15

This design setup meant that the lab technicians had no time to prepare for the experiment in any way. This allowed the observation of the sorting activity to be as natural as possible, with as few external factors as possible manipulating the results. Had the staff been alerted to the fact that sorting would have been observed at a specific time, it is possible they would have prepared themselves to seem as productive as possible. The selected approach minimized the chances of an artificial performance.

The lab technician observed had over 17 years of experience in the pathology lab and can therefore give a good idea of the sorting speed of a more experienced worker. For one hour the amount of cassettes that were moved from the pile of unsorted cassettes and sorted into the archive was counted. Results were written down every 10 minutes to get an overview of how much variation could occur in short intervals of sorting.

The table below shows the sorting times measured in the lab. The first column shows the time, the second column displays how many cassettes have been sorted in the last 10-minute interval and the final column shows the accumulated amount of cassettes sorted since the start of the observation.

Table 5.3: Results from hand sorting in the lab

Time	Sorted last 10 minutes	Total sorted
09:57	0	0
10:07	109	109
10:17	84	193
10:27	44	237
10:37	115	352
10:47	122	474
10:57	173	647

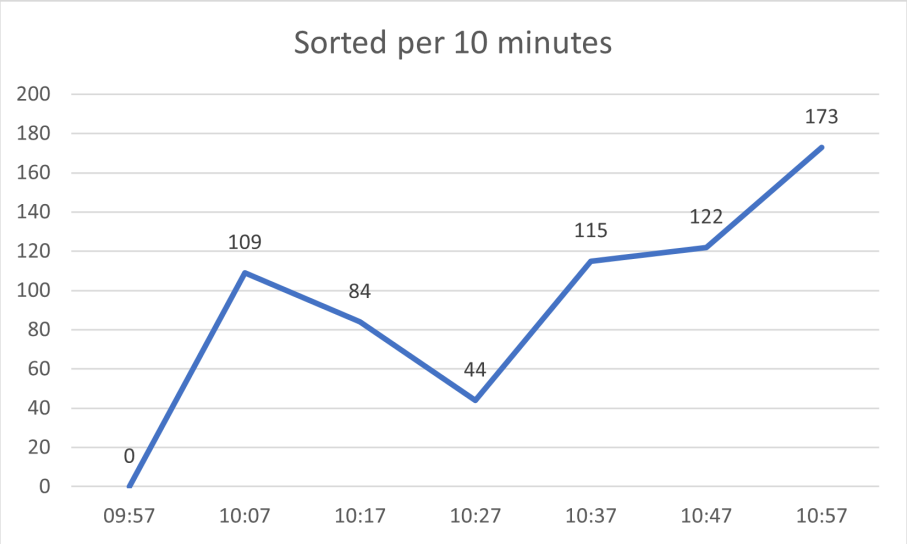


Figure 5.1: This graph show the variation in cassettes sorted in each 10 minute session

Sorting times were also gathered and measured from the application. The same images as the earlier trial were reused as they were a great baseline and reusing them allows for some interesting comparisons and statistics in the discussion.

- Total cassettes in the image shows how many cassettes each image contains.
- Corrected cassettes tells us how many cassettes had to be edited to get the information right.
- OCR runtime is the runtime of the OCR and EAST/ ED algorithms. OCR runtimes are slightly different from last trial, this is to be expected as it can variate slightly based on system performance and background processes.
- Error correction time is how long it took to edit and fix the errors using the software.
- Total time shows how long this process took in total with the accumulated times for runtime and error correction.

The last row shows the accumulated times for all images processed. All times are written as minutes:seconds.

Table 5.4: Results for using EAST for image sorting

image id	Total cassettes in image	Corrected cassettes	OCR runtime (minutes)	Error correction time (minutes)	Total time (minutes)
IMG 0892	37	11	00:30	01:41	02:11
IMG 0894	49	12	00:34	01:24	01:58
IMG 0895	37	21	00:26	02:26	02:52
IMG 0896	44	20	00:26	02:34	03:00
Total	167	64	01:56	08:05	10:01

Table 5.5: Results for using ED for image sorting

image id	Total cassettes in image	Corrected cassettes	OCR runtime (minutes)	Error correction time (minutes)	Total time (minutes)
IMG 0892	37	22	00:06	02:51	02:57
IMG 0894	49	31	00:08	03:45	03:53
IMG 0895	37	31	00:07	03:38	03:45
IMG 0896	44	42	00:08	04:46	04:54
Total	167	126	00:29	15:00	15:29

5.1.1 Data and analysis

The ED model has a total accuracy of 77.84% (130/167). Compared to using EAST which achieves an average accuracy of 93.41% (156/167). This accuracy increase comes at a cost of an average runtime increase of 22.01 seconds more as the average runtime of edge detection is 8.3 seconds whereas EAST has averages 30.31 seconds per image processed. In both cases the runtime does not seem to have a strong correlation with the amount of cassettes detected or the amount of cassettes in an image as seen IMG 0892 and IMG 0896 compared to the other two.

After looking at cassette detection, we see that OCR performance also vary from edge detection to EAST. Out of the 130 cassettes detected by ED only 47 is completely recognized which gives ED an accuracy of 36.1% in cassette detection. EAST performs a little better completely recognizing 97 out of 156 for an accuracy score of 62.1%.

The system occasionally detects cassettes where there are none. Edge detection seems to have a slightly higher likelihood to detect false cassettes at 2.39% (4/167) false cassette per cassette in an image compared to EAST which detects 1.19%

(2/167) false cassettes per image. None of the false cassettes contain any text and they are all labelled as ”none-none-none”. In addition to these some of the cassettes which were fully recognized contained incorrect information which have to be identified and corrected by the user. Out of the 47 recognized by ED 2 cases contained incorrect information giving a failure rate of 4.2% thus 95.8% of the text recognized by the ED based methods are correct. EAST also had 2 errors out of the 97 cassettes giving it an error rate of 2.06% and it means that EAST is able to correctly read 97.94% of the text it is able to recognize. One interesting fact to note is that one of the cassettes which were incorrectly read by the OCR program were the same cassette in both ED and EAST.

Looking at the cassettes sorted in the lab we can see the lab technician were able to sort 647 cassettes in one hour. This gives us an average of 107.8333 cassettes in a 10-minute interval and 10.7833 cassettes per minute. The amount of cassettes sorted in 10-minutes intervals vary by a large amount. The lowest amount moved from the unsorted pile into the archive in a single 10-minute session was 44. Whereas the largest amount was 173. The reason for the drop of cassettes sorted in the third interval is due to a sample containing too many blocks to fit in the space that was left. This meant that a large amount of already sorted blocks had to be moved and restructured to make space for the new blocks. Usually, they make space for a few large samples, but very often situations like this occur where a large section of cassettes has to be moved in order to make space. Moving and transferring cassettes like this is very time consuming. In addition, items are pulled from archive daily and must often be re-sorted into it. When sorting they have to take this into account and make space for cassettes by checking numbers before and after to ensure that there is always space for new cassettes. This is very time-consuming, and it also stops lab technicians from being able to compress space and put as many cassettes as possible in a single box. As they never know how many cassettes might appear to fill an empty space to keep the archive in numerical order. Restructuring can take time if a large order needs to fit into a small space. Other challenges also appeared in the hour of observation the technician worked. One of the boxes of cassettes that came in had been processed by a student. Unfortunately, this student had made the mistake of placing the cassettes in the wrong order. Thus, the lab technician had to re-structure the entire box be-

fore placing the cassettes into the archive. Leading to further slowdowns. In the final 10 minutes a box of cassettes arrived where the cassettes were already in numerical order. This greatly sped up the process, but this does not happen often. During sorting two cassettes placed wrong were located and fixed, human errors such as these are constantly taking place and slip though. When mistakes like this happen it makes those wrongly placed cassettes impossible to locate in an archive where the section for 2022 contained a total of 185065 cassettes as per table 2.1

To get some more useful data out of the table 5.5 and 5.4 we first need to convert the data into a format easier to compare. To figure out how many cassettes can be sorted each minute using the application you can take the total of 167 cassettes and multiply with the minutes and seconds it took to sort. Each second needs to be converted into a numerical scale that goes to 1. This is done by dividing 1 by 60 resulting in 0.0166. EAST spent 10 minutes and 1 second in total on runtime and error correction giving us this calculation: 167/(10+0.0166) = 16.67221 cassettes per minute. Multiplying by 10 gives 166.7221 cassettes per 10 minute and multiplying this number by 6 gives a total of 1000.333 cassettes sorted per hour. For ED we can do a similar calculation to find that ED had a longer total time with a shorter runtime, but much more time spent on error correction for a total of 15 minutes and 29 seconds. To calculate cassettes sorted per minute we do: 167/(15+(29*0.0166)) giving us a total of 10.7857 cassettes sorted per minute. 107.8579 in 10 minutes and 647.1475 cassettes sorted per hour.

Table 5.6: Cassettes sorted on average in different times

	Manual sorting	EAST	ED
Per minute	10.78333	16.67221	10.78579
Per 10 minutes	107.8333	166.7221	107.8579
Per hour	647	1000.333	647.1475

Looking at how many cassettes can be sorted on average it is obvious that EAST outperform ED by quite a bit. As ED barley outperform manual sorting by a negligible amount of time. Looking at EAST the difference between the proposed solution and the manual solution of today is quite noticeable. In total EAST gives a sorting speed increase of 54% compared to manual sorting. Whereas ED gives a performance increase of 0.0028% essentially being akin to the current manual sorting speed.

Table 5.7: Estimated amount of hours it takes to sort in lab and automatic and time saved

Year	Cassettes	Manual sorting time	EAST sorting time	Time saved with EAST
2016	155414	240.21	155.36	84.84
2017	164179	253.75	164.12	89.63
2018	169817	262.47	169.76	92.71
2019	172879	267.20	172.82	94.38
2020	171505	265.08	171.45	93.63
2021	180323	278.71	180.26	98.44
2022	185065	286.04	185.00	101.03

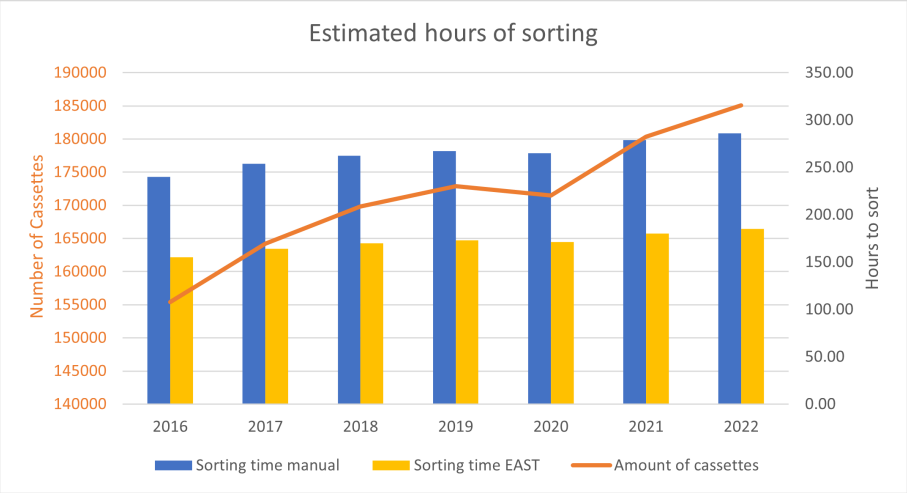


Figure 5.2: This graph shows an estimate of the sorting times saved and the amount of cassettes increasing each year

Table 5.7 shows the estimated times of the current manual sorting technique compares to using EAST in the current state it is in. In 2022 101 hours of manual labour could have been saved in sorting activities if a digital tool with the performance of the proposed EAST sorting tool was available to assist. This means that a lab technicians would have an addition 101 hours of time to focus on more important tasks. As seen in figure 5.2 more time is saved each year as the time to sort increases due to the amount of cassettes need to be sorted which also keeps increasing. As can be seen in figure 5.2 the pathology is estimate to be able to save 35.32% of the time spent on sorting activities by using the proposed solution using EAST.

5.2 Discussion

This part will be an in dept-discussion regarding the use of the application and if it is suitable for deployment in the pathology department. The data from results will be analysed. does the application provide enough of an increase in production and is it accurate enough to possibly utilize? If not, what, and how can be improved?

5.2.1 Performance

At the moment the application is able to locate almost all cassettes in an image with a high consistency using both edge detection and EAST. However, EAST performs slightly better with 15.57% higher average accuracy, but also comes with a trade-off of much higher average runtimes with an additional 22.01 seconds. There might be ways to improve edge detection and achieve scores more in-line with EAST at a faster time. The easiest would probably be to test a few different ways of pre-processing the images and tuning the algorithm accordingly.

As you can see the system occasionally ”find” cassettes that does not exist. This is easily rectified by the user when using the application as the user would spot that the system highlighted an area without cassettes. As mentioned, the system requires all cassettes to contain data before allowing for data submission, none of the false detections contain any text and can easily be found and deleted.

Once the cassettes are detected both edge detection and EAST struggle to read the text. This could be due to some of the challenges related to image recognition mentioned earlier, but also due to limitations with OTSU thresholding and Gaussian blurring being the only pre-processing steps before OCR. Different methods of thresholding can be attempted as well as more intensive pre-processing, It would also be interesting to see compare how different OCR solutions would perform. It is possible that tesseract is under performing.

The almost 30% difference in OCR performance in EAST and ED is very interesting. As once the cassette detection has detected the cassette the OCR pre-processing is the same for both cases. The only difference between the two methods of findings

cassettes is the size of the cut-out as can be seen in Figure 5.3. ED keeps a larger part of each cassette as it keeps the entire cassette in the cut-out, whereas EAST just keeps the text part, ignoring the data matrix code and the area between. This could be why the OCR performance between the two images is so different. The data matrix code could also be the cause of the large gap as it could be recognized as text and obscure the results.

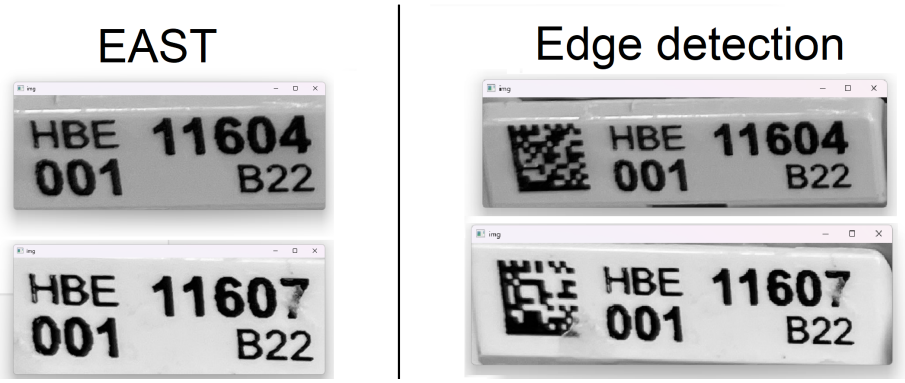


Figure 5.3: This image shows the difference in the cut-outs of ED and EAST

The reason OCR performance varies in EAST and ED could be due to the size of the detected area. Edge detection finds the entire cassette whereas EAST only finds the area containing text. This means images from ED contains more noise, specifically the data matrix code could be causing issues with the tesseract. It should be possible to remove it from the image and see if that gives any better results. Larger images also contain more noise and shadows these factors would make the OTSU threshold perform differently depending on the content and size of the image. A more advanced processing pipeline should be able to produce better results for ED either by shrinking the detected area a little or removing the data matrix code. A more advanced pre-processing line in general would be ideal to achieve an accuracy higher than both ED and EAST as they are currently suboptimal.

The system developed is not fully automatic and the user is able to correct some mistakes through the GUI. If the system were to be fully automatic, which is a long-time goal we would need some way of error handling to ensure all cassettes are detected. As mentioned earlier losing even one cassette can be detrimental in changing the course of a patient's life. The current goal of creating a user overseen system.

Manually typing in the text on cassettes does introduce the possibilities of human er-

rors. There are a few methods to counter this. Firstly, a method to ensure there are no cassettes containing the same information is one. Human error is one of the largest challenges and cassettes are placed wrongly every day. Making them almost unfindable in the sorted pile. According to the technician that were sorting when observation occurred.

5.2.2 Specific cases

Looking at the images on a case-by-case basis it is apparent that some images perform very differently when the images are similar in nature. One difference can be seen in image 0896 which has a lot of shadows. It also features darker cassettes which in turn causes the shadows to be perceived as even darker. This could be causing issues with the OCR.

Image 0892 also does not work well with edge detection. This specifically suffers from an issue where due to the edges not being too visible the cassettes are captured and perceived as being too large and therefor disregarded as noise. Image 0892 also suffers slightly from obscured cassettes due to the image angle.

5.2.3 Barcode scanning times versus runtime

Looking at runtime, both methods proposed would be faster than barcode scanning the cassettes one by one. According to SPOT imaging which claim scanning 150 cassettes takes 450 seconds to scan by hand (SPOT Imaging, 2023). This translates to an average of 3 seconds of scanning time per cassette. To hand scan 35 cassettes would already take 2 minutes which is much longer both solutions for identifying cassettes proposed and tested here. In addition, hand scanning would also require manually sorting cassettes which would add a lot of additional time to the total not required by the proposed tracking solution. According to the pathology lab SPOT's timings are not entirely correct, cassette scanning time can vary between 1 second per cassette to 3 seconds. It depends on the training of the person scanning and how many times a cassette has to be re-scanned if scanning fails. They estimated that scanning 150 cassettes would take somewhere around one to two minutes for a trained employee in contrasts with SPOT's estimation of seven and a half minute. Both of these are still

longer than the runtime for the application proposed here.

To verify these claims, we did a quick in lab trail by scanning twp boxes of cassettes. The first box was not sorted, and cassettes occasionally had to be moved to perform a scan. It took 3 minutes and 43 seconds to scan 136 cassettes. Giving a time of 1.5 second per cassette. We also tried scanning a newer box with spacers to see if there was a noticeable difference. Using the spacer, we were able to scan a box of 226 cassettes in 2 minutes and 11 seconds, which is 0.5 seconds per cassette. However, the box with the spacers are currently in limited supply and part of a new project. Meaning that currently for most boxes we need to look at the first and slowest of these timings. Comparing these measurements to SPOT's we see that they are way lower and much closer to the estimates made by the lab workers.

5.2.4 Comparing manual and automatic sorting

Looking at the improvements gained compared to how long sorting takes in the lab we see some interesting results. Again, EAST has a much longer runtime compared to ED, but the amount of manual work that has to be done and labelled after ED's poor OCR performance leads to a lot of time lost manually typing in the cassette data. In turn ED ends up being almost as slow as manual sorting. However, even though ED is as slow as manual sorting it does come with some advantages. Cassettes with their data digitally stored should be easier to locate and can be directly searched and should be retrievable faster.

If OCR performance is increased by any means and the time spend on error correction could be lowered the amount of time saved could be immensely improved. In theory if we could achieve an accuracy of 100% and keep the same run times as EAST currently have we can achieve a cassette sorting speed of 86.38 cassettes per minute. These calculations are based on total runtime without error correction: 167/(1+(56*0.0166)). If this were feasible, we would achieve a sorting speed increase of 701% compared to manual sorting. This is however a theoretical maximum and is reliant on improvements to the OCR technologies to a point where they are completely automatic and able to perform without producing any errors.

One thing this study has not been able to time is the retrieval process. Right now, a pathologist or lab technician need to manually look through the cassettes to find a specific sample. If the sample is in the sorted pile finding it requires them to look through until they find the start and follow the numbering to locate the cassette. If the cassette however is in the unsorted queue, it can be extremely difficult to find. The digital solution would hopefully help trivialize both of these processes into a search interface which would allow for quick retrieval. The author did not have time to develop this interface as it was not a priority, but the data required is stored in searchable format with the current solution. Being able to quickly locate cassettes should also save time for the pathology lab.

Another thing to keep in mind when comparing the manual and automatic sorting comes regarding the software's runtime. During runtime the user does not need to be present. Knowing runtime takes around 30 seconds the lab could use more than one computer and a worker could possibly run two or three boxes through the system instead of waiting for a runtime to finish, further saving time by parallel processing boxes of cassettes into the archive.

This improvement in sorting time will not only lead to benefits for the pathology lab in allowing them to spend less time per sample, but it will also make each sample less expensive to analyse. From an economical perspective reducing sorting time will also help reduce the cost per sample as each sample is not manually handled for as long during automatic sorting as it is during manual sorting. Therefore, the pathologists and lab technicians can focus on other more important tasks and the administration does not have to pay an educated lab technician for spending a large amount of time doing a trivial task such as sorting.

In addition to the improvements in sorting speed the digital solution would also offer a new layer of traceability to the cassettes in the lab. This will ensure that cassettes are located where they are supposed to be. In an event where a cassette is lost and needs to be located traceability will allow the cassette to be found. If a cassette is not located in its designated position after manual sorting there is currently no easy way to locate it, a digital solution would solve this issue. The logs created by a traceable system will allow for fine grained process mining to be applied. Which can further be

used to increase overall productivity in the lab as seen in the proposed architecture in figure 4.19. Traceability in the archive will be an entirely new addition that can allow the workers in the lab to see if a specific cassette is in the archive or if it has been pulled out to be inspected. Cassettes from the archive are frequently retrieved for various reasons such as checking if a certain type of cancer could be hereditary.

The data reviewed in the results directly supports the initial thoughts that a human's ability to identify cassettes will increase in speed when assisted by digital tools. The augmented identification led to an increase in cassettes sorted within a set period of time. Thus, making if faster for cassettes to be sent further into the lab instead of being stuck in queues and improving the speed of the overall process of a cassette traveling through the lab.

5.2.5 Possible use cases

The proposed system has potential to be used in creating a training set. The data being handled by the system is exactly the type of data that would be needed for such a task. If all the images processed are saved after automatic labelling and user verification, the pathology lab would have all the data necessary to train a neural network. This would require the system to be used in the lab for a prolonged period of time and all data would need to be labelled correctly.

It is also possible to further enhance the system by incorporating additional features. One such feature which were considered during development was incorporating a digital twin. The Digital Twin concept consists of three distinct parts. Firstly, the physical object or the process and its physical environment, the digital representation of the object or process, and the communication channel (the digital thread) between the physical and virtual representations. The connections between the physical version and the digital version include information flows and data that includes physical sensor flows between the physical and virtual objects and environments.

If a digital twin was implemented as such, this digital twin would work in tandem with the OCR system. Incorporating a digital twin would make the system capable of more sophisticated handling and processing of the data output. If implemented in the en-

tire lab the cassettes will be scanned in batches by OCR based tech every time the cassettes are moved from one location to another. The cassette data, location, time and more will then be logged in the system. The digital twin is updated and can be accessed in real time as it is continuously mirroring the physical state of the lab. Utilizing digital twin technologies would allow for a better connection to bridge the physical and the digital gap between cassettes in the lab.

5.2.6 Future challenges introduced by the system

The system presented here has to be rigorously tested and iterated upon if it were to be implemented in the lab. The demo system for this thesis was created to gather data and see what is feasible in the pathology lab. There is a high probability that the current solution contains bugs and issues that can lead to errors down the line if implemented in its current state. As an example, if the system is used by the pathology lab for a set amount of time such as 5 years. An issue could occur rendering the data storage solution inaccessible. This would mean that all the cassettes in the archive would be considered ”lost” as they would all be stored in random orders in random places. Finding a cassette in such conditions would require too much time to be feasible.

If the pathology lab is interested in developing an in-house solution it would require more work than what one master student is capable in producing in one year. Not to mention that the pathology lab would require a solution to maintaining the software over longer periods of time. As mentioned by Dangott there is inherent risk in having the system rely on the knowledge of only a few individuals. *Dangott* (2015). Maintenance and future development will also be required. One such support case could be that the current way of labelling the cassettes would be unfeasible for the system proposed here after the year 2100, due to how the system creates unique IDs. As seen in figure 2.2 the semantic information on cassettes only consists of the last two numbers in the current year. This will lead to conflicts in creating a unique ID for each cassette once it starts to loop over and the numbers start to repeat. To solve this the pathology department could display the full year instead of just the last two digits. However, this introduces another support case to consider. If the pathology lab makes any changes to the semantic information on the cassettes, parts of the application would have to

be reworked to account for such changes. Without developers to support the application this will create certain limitations for the pathology department if they want to keep using the system but make changes to the semantic information contained on cassettes.

5.2.7 Answering research questions

RQ: How can the pathology workflow be streamlined by implementing automatic sorting mechanics for cassettes?

It is possible to streamline a pathology workflow by integrating automatic sorting mechanisms for cassettes. The solution proposed and presented in this thesis implemented automatic sorting mechanisms by utilizing OCR and pre-trained libraries for image processing. This allows a pathology lab to implement workflows as seen in figure 4.18. This enhances the speed of cassettes being sorted in a set amount of time drastically compared to sorting these cassettes by hand.

RQ1: How can the implementation of a tracking solution using optical character recognition (OCR) improve the efficiency of sample processing at a pathology lab?

Implementation of a tracking solution using OCR is able to improve pathology workflow by reducing the time spent on trivial tasks such as cassette sorting. In addition, the software allows for traceability and tracking of cassettes which generate fine-grained data for process mining when appended to the data from the LIS. The results of this process mining can be utilized for simulations which can lead to further improvements of pathology lab workflows.

RQ2: How does the accuracy of cassette detection vary when using different pre-trained image processing libraries, specifically edge detection and the EAST neural network?

According to the results presented in this thesis the EAST Neural Networks outperform the image processing technique of edge detection. EAST has an accuracy in

cassette detection of 93% and edge detection is able to locate 77% of all cassettes in an image. The accuracy of OCR performance also varies between the two different models with EAST achieving 62% and edge detection performs significantly worse at 36%. There are improvements that can be made to both of these. For instance, one could label images and train a custom neural network instead of relying on off the shelf solutions. Other image processing approaches could also be attempted to see if they achieve better results by some significance. Improvements need to be made in the OCR section before a fully automatic solution can be attempted. As was covered briefly in the background section there are other OCR tools available some of which could potentially produce better results in terms of OCR accuracy than tesseract. These OCR models were not tested here due to their financial models and proprietary closed-source approach.

RQ3: What effect does the proposed digital solution have on the time it takes to sort cassettes compared to the current manual sorting method?

This thesis has proven that even a low performing OCR solution is able to reduce the time spent sorting cassettes in certain workflows by 35% with a 54% increase in the amount of cassettes sorted in a set amount of time. In total this decreases the time it takes to process each individual cassette. The result of the saved time frees up the pathologists and lab technicians to focus on more important tasks than sorting. It will also reduce the time spent looking for cassettes as a digital archive will most likely speed up the process of cassette retrieving.

Chapter 6

Conclusions and Future Work

This Chapter concludes the thesis by summarizing the findings from the study, the contributions, and possible limitations of the approach. It can also identify issues that were not solved, or new problems that came up during the work, and suggests possible directions going forward.

In this paper a new approach to locate, sort and track cassettes in the pathology lab at Haukeland university hospital have been proposed. The proposed system is currently able to give an average of 54% increase in cassettes sorted in a set amount of time and reduce the time spend on sorting activities by 35% compared to the current solutions in place. This increase would be very beneficial as it would allow the pathology lab to reduce a large amount of the time spent on manual sorting activities in the pathology lab.

In addition to increased sorting time and sorting performance the system will also gather more data than currently available in the LIS, enhancing the traceability of the cassettes in the pathology lab. This data will allow for more fine-grained process mining and can potentially lead to further increases in productivity and future reductions in manual labor. This data will also allow for cassettes to be easier and faster to locate in specific areas of the lab where they currently have to be searched for manually.

Two different methods to locate and sort cassettes have been attempted. These are edge detection and EAST, a neural network for text extraction. Out of these two EAST

performed significantly better compared to edge detection.

The system proposed still have a few challenges. Firstly, the OCR performance could and should be improved before a system like this is put in place in the pathology lab. EAST has an OCR accuracy of 62%. This should be improved before a system can be put into place, even though the current accuracy allows for faster sorting than the current manual method used in the archive. Testing another OCR engine instead of tesseract could yield interesting results.

Currently the cassette detection system is at an acceptable level with an accuracy of 93.41%, but there are small improvements to be made here. It must be considered that this system will operate in the healthcare sector and thus needs to be as error-proof as possible. There is an argument to be made that a fully automatic solution should be completely accurate due to the nature of the healthcare data. Even a semi accurate solution should strive for as few errors as possible, as the current system in place already suffer from human-errors. Interesting future work can be done looking at how to handle possible errors both human made and OCR based. The system developed here is a proof-of-concept to see what can be done with limited resources, but it shows that there are significant improvements to be made and gives an idea of which techniques to pursue.

A GUI have also been developed to give an idea of how manual error correction would work in a semi-automatic solution. With proposals for how a user might interact with the system to fix incorrect data extracted from the cassettes.

The thesis presents models to show how the pathology lab could incorporate a tool for cassette tracking. The advantages of the system would allow the pathology lab to generate fine-grained data, which would allow for process mining to find methods which could further improve the pathology labs workflows.

With enough time and resources, one could process images and export the data from the software to create a training set and use this to adjust tesseract or create a new OCR-tool from scratch more accustomed to the environment in the pathology lab. In the long run this could also remove the need for using EAST for text detection. However, doing so would require a lot of cassette images and manual labelling to

create.

Multiple computer vision-based systems already exist, many of these already utilize text extraction. However, none of the systems that exists would fit the requirements needed at Haukeland. The system create would also have the possibility to be augmented further in the future. It should be possible to utilize the cassette detection in synergy with a mechanical arm and use it to fully automate the process of sorting cassettes. The technology created in this thesis could also lay the groundwork for such an alternative solution to this problem. Instead of creating a digital log of cassettes for the archive the cassette location technology and OCR tool could be combined with robotics to sort the cassettes in the correct numerical order automatically. Such a system would bring even more benefits as you would have a digital catalogue in addition to the cassettes being sorted and findable without the need for a look-up tool to search for their specific locations.

The digital tools developed in this thesis might also be applicable in for domains facing similar issues. The artifact is freely available as an open-source project at GitHub.

Bibliography

(Azure Computer Vision, 2022) (Online; Accessed 27-03-2022), Azure products - computer vision, available at: https://azure.microsoft.com/en-in/services/cognitive-services/computer-vision. 2.2.3

Bieniecki, W., S. Grabowski, and W. Rozenberg (2007), Image preprocessing for improving ocr accuracy, in *2007 international conference on perspective technologies and methods in MEMS design*, pp. 75–80, IEEE. 4.4

Buesa, R. J. (2009), Adapting lean to histology laboratories, *Annals of diagnostic pathology*, *13*(5), 322–333. 2.3.4

Cakic, S., T. Popovic, S. Sandi, S. Krco, and A. Gazivoda (2020), The use of tesseract ocr number recognition for food tracking and tracing, in *2020 24th International Conference on Information Technology (IT)*, pp. 1–4, IEEE. 2.3.1

Camunda (2023), Camunda platform: The universal process orchestrator, https://camunda.com/, accessed: 2023-15-05. 2.3.3

Chaudhuri, A., K. Mandaviya, P. Badelia, and S. K. Ghosh (2017), Optical character recognition systems, in *Optical Character Recognition Systems for Different Languages with Soft Computing*, vol. 352, pp. 9–41, Springer, doi:10.1007/978-3-319-50252-6. 2.2.2

Dangott, B. (2015), Specialized laboratory information systems, *Surgical Pathology Clinics*, *8*(2), 145–152. 2.3.4, 4.3.1, 5.2.6

(Dreampath, 2023) (Online; Accessed 17-04-2023), Solutions - dreampath, available at: https://www.dreampathdx.net/solutions/#fina. 2.3.2

(Electron, 2023) (Online; Accessed 02-04-2023), Docs - electron, available at: https://www.electronjs.org/docs/latest/. 4.4

(Epredia, 2020) (Online; Accessed 17-04-2023), Overcoming archiving inefficiencies with arcos - thomas jefferson university hospital, philadelphia, pa, available at: https://epredia.com/pdf/arcos/tjuh-case-study/. 1.1, 2.3.2

(Epredia, 2021) (Online; Accessed 17-04-2023), Tissue sample security made simple - epredia, available at: https://epredia.com/pdf/arcos/brochure/. 2.3.2

Grinberg, M. (2018), *Flask web development: developing web applications with python*, " O'Reilly Media, Inc.". 4.4

Hammer, M. (2014), What is business process management?, in *Handbook on business process management 1: Introduction, methods, and information systems*, pp. 3–16, Springer. 2.3.3

Hanna, M. G., and L. Pantanowitz (2015), Bar coding and tracking in pathology, *Surgical pathology clinics*, *8*(2), 123–135. 2.3.4, 4.3.1

Hatlem, M., F. Rabbi, P. Stünkel, and F. Leh (2023), Intelligent tracing and process improvement of pathology workflows using character recognition, *Submitted to HEDA 2023: 3rd International Health Data Workshop*, pp. 1 – 14. (document), 2.1, 4.18, 4.19

Henricks, W. H. (2015), Laboratory information systems, *Surgical pathology clinics*, *8*(2), 101–108. 4.3.1

Hevner, A. R., S. T. March, J. Park, and S. Ram (2004), Design science in information systems research, *MIS quarterly*, pp. 75–105. (document), 3.1, 3.1, 3.1.1, 3.2

Jones, D., and S. Gregor (2007), The anatomy of a design theory, *Journal of the Association for Information Systems*, *8*(5), 312–335. 3.1

Karthikeyan, S., A. G. S. de Herrera, F. Doctor, and A. Mirza (2021), An ocr post-correction approach using deep learning for processing medical reports, *IEEE Transactions on Circuits and Systems for Video Technology*. 2.3.1

Kumar, N., and M. Nachamai (2017), Noise removal and filtering techniques used in medical images, *Orient. J. Comput. Sci. Technol*, *10*(1), 103–113. 4.4.3

Long, J., E. Shelhamer, and T. Darrell (2015), Fully convolutional networks for semantic segmentation, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440. 2.2.1

Memon, J., M. Sami, R. A. Khan, and M. Uddin (2020), Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr), *IEEE Access*, *8*, 142,642–142,668. 2.2.2, 2.3.1

(Moore's law, 1965) (Online; Accessed 05-05-2022), Oxford reference, available at: https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100208256. 2.3.1

(OpenCV, 2022) (Online; Accessed 13-04-2022), About - opencv, available at: https://opencv.org/about. 2.2.3, 4.4

Othman, H. (2019), Pathology an introduction, doi:10.13140/RG.2.2.25459.68647. 2.1.1

Otsu, N. (1979), A threshold selection method from gray-level histograms, *IEEE transactions on systems, man, and cybernetics*, *9*(1), 62–66. 4.4.3

Otterlei, S. S. (2019), 14 pasienter har dødd etter sykehusfeil som kunne vært unngått, *NRK*, available at: https://www.nrk.no/vestland/14-pasienter-har-dodd-etter-sykehusfeil-som-kunne-vaert-unngatt-1.14440507. 4.3.1

(Oxford English dictionary, Machine learning, 2022) (Online; Accessed 09-04-2022), Oxford english dictionary : the definitive record of the english language, available at: https://www.oed.com/view/Entry/111850?redirectedFrom=machine+learning#eid38479194. 2.2.1

(Oxford English dictionary, Pathology, 2022) (Online; Accessed 09-04-2022), Oxford english dictionary : the definitive record of the english language, available at: https://www.oed.com/view/Entry/138805?redirectedFrom=pathology#eid. 2.1.1

Peterson, J. L. (1977), Petri nets, *ACM Computing Surveys (CSUR)*, *9*(3), 223–252. 2.1.2

Platou, H. S. (2021), Business process simulation as a service, Master's thesis, University of Bergen. (document), 2.3.3, 4.20, 4.8

(Python, 2022) (Online; Accessed 13-04-2022), General python faq - what is python?, available at: https://docs.python.org/3/faq/general.html#what-is-python. 2.2.3

Radha, R., and R. Aparna (2013), Review of ocr techniques used in automatic mail sorting of postal envelopes, *Signal & Image Processing*, *4*(5), 45. 2.2.2

Smith, R. (2007), An overview of the tesseract ocr engine, in *Ninth international conference on document analysis and recognition (ICDAR 2007)*, vol. 2, pp. 629–633, IEEE. 2.2.3

(SPOT Imaging, 2023) (Online; Accessed 26-03-2023), Pathtracker - spot imaging, available at: https://www.spotimaging.com/pathtracker/. 2.3.2, 5.2.3

(Statens Vegvesen, 2023) (Online; Accessed 12-04-2024), Automatic number plate recognition (anpr), available at: https://www.vegvesen.no/en/fag/fokusomrader/trafikksikkerhet/automatic-number-plate-recognition-anpr/. 2.2.2

Stünkel, P., S. Leh, and F. Leh (2022), Process data science for workflow optimization in digital pathology: A status report, *HEDA 2022*. (document), 2.1.1, 2.1, 2.1.2, 2.3.3, 4.3.1

(Tesseract - GitHub.com, 2023) (Online; Accessed 26-03-2023a), Tesseract-ocr/tessdoc, available at: https://github.com/tesseract-ocr/tessdoc. 2.2.1

(Tesseract - GitHub.com, 2023) (Online; Accessed 26-03-2023b), Improving the quality of the output, available at: https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html. 4.4

Van Der Aalst, W. (2011), *Process mining: discovery, conformance and enhancement of business processes*, vol. 2, Springer. 2.3.3

van der Aalst, W. M. (2019), Object-centric process mining: Dealing with divergence and convergence in event data, in *Software Engineering and Formal Methods: 17th International Conference, SEFM 2019, Oslo, Norway, September 18–20, 2019, Proceedings 17*, pp. 3–25, Springer. 2.3.3

Van der Aalst, W. M., J. Nakatumba, A. Rozinat, and N. Russell (2010), Business process simulation, *Handbook on Business Process Management 1: Introduction, Methods, and Information Systems*, pp. 313–338. 2.3.3

(Vision AI, 2022) (Online; Accessed 27-03-2022), Google - vision ai, available at: https://cloud.google.com/vision/. 2.2.3

(w3techs.com, 2023) (Online; Accessed 04-03-2023), Usage statistics of javascript as client-side programming language on websites, available at: https://w3techs.com/technologies/details/cp-javascript/. 4.4

Wang, Y., H. Xu, C. He, and L. Hong (2020), New concept of filing pathological tissue samples, *World Journal of Pharmaceutical Research*. 2.3.2

White, G. R., G. Gardiner, G. Prabhakar, and A. A. Razak (2007), A comparison of barcoding and rfid technologies in practice., *Journal of Information, Information Technology & Organizations*, *2*. 1.1

Williamson, K., and G. Johanson (2017), *Research Methods: Information, Systems, and Contexts*, Elsevier Science & Technology, San Diego. 3.1

Zayas-Cabán, T., S. N. Haque, and N. Kemper (2021), Identifying opportunities for workflow automation in health care: lessons learned from other industries, *Applied Clinical Informatics*, *12*(03), 686–697. 2.3.4

Zhou, X., C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang (2017), East: an efficient and accurate scene text detector, in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 5551–5560. 2.2.1, 4.4.2

Appendix

.1 Code and file structure

The code developed for this thesis is freely available at GitHub and can be found at the following link: https://github.com/M-Hatlem/ocr-mastersthesis

- The assetes folder includes icons and animations utilized in the user interface. The current iteration only includes a loading icon for when cassettes are being processed.
- The Output folder contains an empty JSON file. When the program is run the output will be appended to this file.
- The python folder includes the python scripts used for the image processing pipelines. When follwing the installation instructions below the user will need to add a python environment folder and the EAST files to this folder.
- README.MD includes information and install instructions.
- functions.js is a JavaScript file containing functions for the GUI.
- index.html is the HTML page for the GUI
- index.js is the electron index file and contains the code to start the user interface and the image processing python scripts.
- package-lock.json and package.json includes the node packages.
- style.ccs includes the CSS styling for the GUI.

.2 Installation instructions and requirements to run the program

These instructions cover installation for the OCR application

- 1. First install Git and clone the repository from GitHub with the commnad:

```
git clone https://github.com/M-Hatlem/ocr-masterthesis
```

- 2. Download node (recommended is latest LTS) (16.18.0 used for development)
- 3. Install electron (Developed using version 21.0.1) by navigating to the Ocr-Master directory and running the command:

```
npm install electron --save
```

- 4. Download Python (Developed using Python 3.10) and setup a Venv with flask, tessaract and CV2 by running:

```
pip install flask tesseract cv2 numpy
```

- 5. Move the Python Venv into Ocr-Masterthesis/Python/ –> should look like Ocr-Masterthesis/Python/Venv....
- 6. Download EAST text detector and place the file in Ocr-Masterthesis/Python/
- 7. Install Tessaract and add it to the console/Path
- 8. Run the app using ”npm start” from Ocr-Masterthesis in command line

```
npm start
```

.3 How to use the program

The program should run after following the instructions above. The program is able to capture images using a connected webcam as prompted on the initial start page. There is also support for uploading images from local storage and run the images through the programs cassette location and OCR pipelines. At the request of the pathology lab images of pathology cassettes from Haukeland are not included with the GitHub repository. However, permission was granted to use images of cassettes within the thesis.

Once an image is captured or uploaded to the application it will run the images through the application pipeline seen in figure 4.7. The program will use EAST or edge detection based on the EAST checkbox being ticked or not. This process could take a while depending on the specifications of the machine it is running on and methods selected. Edge detection should perform significantly faster than EAST.

Once the process is finished the user interface should look like figure 4.12. Here cassettes can be clicked on, searched and highlighted using the boxes on the right or search bar on the left. To draw a cassette on the image, move the cursor to the location the rectangle should start. click and hold the left mouse button and pull until the box surrounds the desired area. Boxes can be deleted and the information on the changed by clicking inside the colored rectangles. Once all cassettes are labeled and fixed as well as the archive field being filled out the results can be submitted. This will save the output as a JSON object in the file located in: Ocr-Masterthesis/Output/OutputData.JSON