Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Engineering
Department of Mechanical and Industrial Engineering

Tina Kazemitalachi

Data-driven approaches for identifying risk factors in indoor environment during a surgery

Master's thesis in Reliability, Availability, Maintainability, and Safety
Supervisor: Yiliu Liu

June 2022

NTNU
Norwegian University of
Science and Technology

Tina Kazemitalachi

Data-driven approaches for identifying risk factors in indoor environment during a surgery

Master's thesis in Reliability, Availability, Maintainability, and Safety
Supervisor: Yiliu Liu
June 2022

Norwegian University of Science and Technology
Faculty of Engineering
Department of Mechanical and Industrial Engineering

NTNU
Norwegian University of
Science and Technology

Abstract

Surgical site infection (SSI) is one of the common complications after surgeries. It results in various consequences, from additional treatment to morbidity and mortality. Therefore, it increases the probability and duration of hospitalization. In Norway, SSI is counted for one-quarter of all healthcare-associated infections. Aside from physical impacts, it threatens patients' emotional well-being. It also has negative effects on hospitals, for instance, more work pressure on the hospital personnel. Furthermore, SSI puts a substantial financial burden on the patient and hospital. It is estimated that the cost of treating SSI can vary from 400$ to 30,000$ per patient. Therefore, SSI is a potential threat to patient safety.

This thesis aims to develop a data-driven model based on different factors for predicting SSI, considering infection as an undesired result of surgery that can negatively affect a patient's safety. This study suggests how different models can be selected in different predictions. Many factors lead to infection. Patient-related reasons, like the well-being of the patient and factors related to each patient. Environmental factors include factors that are presented in the operating room or from the people presented in the surgery. Surgical-related factors which happen during the surgery or because of the preparation for the surgery. Lastly, heating and ventilation factors that influence the air quality. In this study, the focus is on environmental factors. The study begins with discussions about risk analysis and SSI and how SSI can be investigated in risk analysis studies. For performing any data-driven study, data is required. As a result, an experiment to collect indoor data was performed. The experiment was done on a pig in an acute operating room in St.Olav's hospital. Due to the novelty of the work, dataset faces some problems, like not enough number of observations; therefore a part of this thesis is dedicated to solve these issues before beginning the risk analysis part. In the end, different measures were used to evaluate the efficiency of different models in different settings. Finally, when coming up with data-driven models and calculating their efficiency in prediction, suggested models can differ depending on what healthcare personnel require from prediction.

Preface

This master's thesis is the mandatory part of the two-year international master's degree program in Reliability, Availability, Maintainability, and Safety (RAMS) Engineering at the Norwegian University of Science and Engineering (NTNU). The thesis is carried out in collaboration with Operating Room Of The Future (FOR), located in St.Olav's hospital during the spring semester of 2022.

Acknowledgment

I would like to extend my gratitude towards my supervisor Yiliu Liu for his invaluable support and directions. I would like to say thanks to professor Cao Guangyu and Liv-Inger Stenstad for providing the opportunity to use FOR in St.olav's hospital. I would also like to thank Yang Bi and Tomáš Fečer for their help and useful inputs throughout the thesis.
At the end, I would like to send my greatest gratitude to my family who supported me during my master studies and without whom I could not fullfill my dreams.

Abbreviation

KRI Key Risk Indicator

RIF Risk Influencing Factor

SSI Surgical Site Infection

CFU Colony Forming Unit

BMI Body Mass Index

AUC Area Under Curve

ANN Artificial Neural Network

TP True Positive

TN True Negative

FP False Positive

FN False Negative

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Almost 313 million surgeries are performed each year around the world.[1] The percentage of surgical site infection (SSI) per 100 surgical procedures varies from 0.6% to 9.5% depending on the type of procedure.[2] SSI results in a negative impact on patients' physical and mental health. It brings 2–11 times higher risk of death to patients with SSI compared with patients without an SSI. Besides, it prolongs the hospitalization period and imposes a burden on hospital, patients, insurance companies and etc.

Different types of factors are suspected to influence SSI. These include environmental, patient-related, surgical-related, and ventilation factors. Environmental factors of operating theatre such as level of fungi, bacterial contamination, temperature, humidity level. Currently, SSI surveillance is being done by direct and indirect methods like a review of medical records, visiting the ICU and wards and talking to primary care staff, surgeon, and patient surveys by mail or telephone. However, these surveillance methods have deficiencies for predicting SSI. They just address SSI when it is too late to take any precautions since SSI has has already happened. Therefore, other methods should be developed to predict the happening of infection.

To overcome challenges in time and predictability, this master thesis aims at developing a data-driven model for assessing SSI risk in an operating room using environmental data gathered from an acute operation room in st. Olav's hospital.

1.2 Motivation and objectives

In this thesis, the main objective is to investigate the value and limitations of machine learning techniques to predict SSI occurrence. In other words, can a predictive model be used as a dynamic model to make decisions regarding existing surgery? To achieve this goal, first, some sub-objectives have to be achieved, which are listed below:

- To understand how data can achieved and get prepared for a data-driven work.
- To understand the behavior of the dataset.
- To investigate methods to solve the problem in a dataset.
- To implement machine learning on the dataset and discuss the results and measures for selecting the optimal model.

1.3 Project scope and limitations

This master thesis is limited by certain boundaries, narrowing the scope and its coverage. These limitations are the following:

- The result of the thesis is based on the limited data and information provided by the colleague and hospital during the whole duration of the study.
- The focus of the master thesis is limited only to one case study.
- This master thesis uses the data only from one small procedure on a pig under surgical conditions, which was tried best to be ideally close to real human surgery. However, some defects and interruptions were present before and during the surgery. Also, not having permission to collect data from a real human surgery was another limitation.
- As said, there was no prior experience in performing such experiments, and when an experiment is new, more trial and error gives better results. However, the limitations of budget, rules in the healthcare section in Norway, and time schedule of the hospital for reservation of an operating room made data collection challenges. (It is worth mentioning that data collection was done in an acute operating room in the biggest and busiest hospital in Trondheim. Therefore, reservation of an acute operation room cannot be made often.)

- This master thesis is an interdisciplinary work, bonding healthcare, data science and risk studies together. Therefore, consultation with healthcare personnel could have been helpful, which was not available.
- Aside from investigating machine learning logarithms, this thesis addresses challenges and limitations within this topic. Therefore, the whole project can be regarded as a road map rather than a determined work.
- The master thesis is time-bound with a limited duration within the Spring 2022 semester.

1.4 Outline

The master thesis is organized with the following structure:

- Chapter 1 states an introduction of SSI, its impact on people and healthcare section, why is it important to be mitigated, and what limitations are present.
- Chapter 2 discusses risk, risk analysis, surgical site infection, and causes related to SSI. It also connects risk analysis with SSI.
- Chapter 3 summarises the experiment from which the data is retrieved. This chapter discussed the surgery plan and the people and equipment that were involved.
- Chapter 4 is about which data will be used and how to handle data to prepare for a data analysis study.
- Chapter 5 visualize the data using graphs. This chapter guides perceiving the data and the first step in analyzing the data.
- Chapter 6 includes an implementation of algorithms on the initial dataset to find the problems and discusses potential methods for overcoming the problems.
- Chapter 7 discusses methods to solve the problems found in the previous chapter, including a literature study around the raised problems.
- Chapter 8 is about machine learning implementation on the final dataset and discusses the performance of evaluation metrics.
- Chapter 9 presents the conclusion and results of the work and addresses limitations and challenges in the project. Finally, it brings recommendations for improving the work for later stages.

Chapter 2

Risk information of surgical site infection

In this chapter surgical site infection and its causes are introduced. Moreover, risk and risk analysis are being discussed. At the end, how SSI as a threat to patient's safety can be studied in risk analysis is explained.

2.1 Surgical site infection

Surgical site infection (SSI) is defined as infections occurring up to 30 days after surgery (or up to one year after surgery in patients receiving implants) and affecting either the incision or deep tissue at the operation site. Despite improvements in prevention, SSI remains a significant clinical problem as it is associated with substantial mortality and morbidity and imposes severe demands on healthcare resources [3].

2.1.1 Causes of surgical site infection

The risk of SSI is strongly correlated with the number of airborne bacteria in the operating room and the surgical field. The source of these bacteria is the surgical team itself, as we all emit thousands of bacteria-carrying skin particles every minute into the air. Although the risk of airborne contamination has decreased over the years, thanks to modern surgical clothing and advanced operating room ventilation, airborne bacteria are still detected and cause SSI [4]. Also, the concentration of bacterial load or Colony forming unites (CFU) is blamed for SSI. CFU is related to relative humidity, temperature, and the particulate matter concentration (PM2.5 and PM10)[5]. It has been presented that the happening and severity of SSI, includes

patient susceptibility to infection, surgical staff practices, operating room cleanliness, and the Heating, Ventilation and Air-Conditioning (HVAC) system. In some studies for developing prediction model for SSI, different factors related to patients were also considered, such as BMI, the history of alcohol consumption, history of heart disease, and other health problems. [6],[7]

2.2 Risk and risk analysis

In book "risk assessment" Marvin Rausand explains, risk is the combined answer to the three questions: (1) What can go wrong? (2) What is the likelihood of that happening? and (3) What are the consequences? The three questions may be explained as followes:

(1) what can go wrong?
For answering this question, it needs to identified the possible accident scenarios that may damage assets (people, animals, the environment, buildings, technical installations, infrastructure, information, data) that we want to keep safe. An accident scenario is a consequence of events, starting with an initiating event and ending with a state that affects and causes harm to the asset.
(2) What is the likelihood of that happening?
The answer to this question can be a qualitative statement (too high, too low) or quantitatively like probabilities or frequencies.
(3) What are the consequences?
For each accident scenario, potential consequences and harms must be identified.

Risk analysis is carried out to provide answers to the three questions. Therefore, risk analysis can be defined as "A systematic study to identify and describe what can go wrong and what the causes, the likelihoods, and the consequences might be".

2.2.1 Key risk indicator and risk influencing factor

For risk analysis, risk level should be compared to a criteria to define how much the risk is low or above the tolerance level. This criteria is called indicator in general, and combined with risk, is called key risk indicator (KRI). In Rausand's book indicator is "A measurable/operational variable that can be used to describe the condition of a broader phenomenon or aspect of reality". On the other hand, there is an other metric called risk influencing factor (RIF). Risk influencing factor is "A relatively stable condition that influences the risk ". In healthcare, RIF is something that

increases the chance of developing a disease and KRI can be diseases itself or levels that if a patient passes them, he shows signs of disease or health complications.

2.3 Attaining key risk indicator and risk influencing factors

Referring to previous sections, KRI is a metric that predicts potential risks and threatens safety. In this study, the final objective is to identify risk factors resulting in a high infection rate. Therefore, in this study, KRI is the high infection level traced by the CFU level. In addition, RIF is any factor causing a high rate of infection. RIF is listed as the indoor factor in the operating room during surgery in this work. RIF can have multiple natures like a patient's background health condition, type of surgery, medications, etc. However, this thesis decided that only indoor factors be investigated.

For conducting a study to identify the risk of SSI based on indoor factors of the operating room, data should be collected. According to the literature review of SSI causes, different variables like humidity, temperature, and the number of particles should be collected. Data collection is conducted with the collaboration of St.Olav's hospital and the energy department in NTNU. This data collection is in the form of a simple procedure on a pig. The details of the surgery are explained in the next chapter.

2.4 Summary

This chapter provided information about risk and risk studies. First, it started by discussing surgical site infection and its causes. Second, it stated that risk influencing factors and key risk indicators have to be understood for performing a risk study. Therefore, RIF and KRI for performing a data-driven study to determine indoor risk factors during surgery were defined in the final part.

Chapter 3

Animal experiment

In this chapter, the procedure for data collection is going to be described. The objectives of this part is to elaborate required preparation for experiment, procedures in conducting the experiment like description of equipment, type of gathered data, and finally discussion of the results.

3.1 Experiment preparation

Based on the discussion with other students involved in the project and supervisors, it has been decided to set up two types of experiments: tentative and real animal tests. The experiments were held in collaboration with the energy and process department of NTNU and St.Olav's university hospital. In total, two animal experiments were done, one in 26th November and one on 10th December in an acute operating room with a mixing ventilation system, which is also called the operating room of the future. In this thesis, the data from the later surgery is used. In all the experiments a set of equipment are used are the same. In the following part, target of the experiment, used equipment and surgery plans will be discussed.

3.1.1 Target

The aim of this data collection is to gather data from operation's room environment like temperature, humidity, and particle results and use this data to investigate if infection can accrue. The reason is that according to literature studies environmental-related factors as well as patient-related factors, can have adverse effect on the occurrence of SSI.[8],[9] Ideally, data has to be collected from a real surgery, but due to limitations, mock-up surgeries have to done instead. Therefore,

data collation has to be done during a surgery where the conditions are close to a real surgery.

3.1.2 List of equipment used in the experiment

Here the used equipment are listed. Pictures of all the equipment can be founded in Appendix A. 1. Aerotrak 9306
The Aerotrak 9306 is a handheld device for monitoring the number of particles ranging from 0.3 to 10 micron in diameter.

2. Tiny Tag
The Gemini tiny tag plus 2-channel self-contained temperature and relative humidity data logger. This unit features a coated RH sensor that has good resistance to moisture and condensation, ensuring measurement reliability.

3. Thermal Camera
Bosch PD1 is a thermal detector that measures surface and room temperatures and also humidity. PD1 is a thermal detector based on infrared technology that detects the surface temperature of the surgical incision. This device was used to collect information of temperature above the operating table.

4. Thermal Couple
A thermal couple is a wire attached to different surfaces to gather temperature data.
5. Data Logger for Thermal Couple, Hioki LR 8400
It is a portable data logger with 30 standard channels expandable to 60 channels. It was used with the thermal couples. The temperature received from thermal couples were logged into this device.

6. Flir E60
It displays infrared images with the surface temperatures of the assigned spots (here people around the operating table) and recorded thermal images and surface temperatures. Temperature measurements by this device have an accuracy of ± 2 °C for ambient temperatures between 10 and 35 °C Surface temperatures range from 20 to 120 °C, with a thermal sensitivity of 0.05 °C at 30 °C.

7. Air thing
This device is a portable and advanced air quality monitor, with 7 sensors including radon, particulate matter (PM2.5) and CO_2.

8. Blood Agar
A petri dish is used for sampling bacteria. Blood agars are used both for passive and active sampling. Both types of sampling were used in the experiments. However,

for active sampling air sampler must be used as well.

9. Air Sampler
This device is used to estimate the level of CFU in the air. In this regard, by analyzing a fixed volume of air, colonies are expressed as CFU per cubic metre of air. The used device in the tests measured CFU level in 1000 L air. (Note that the uploaded picture of air sampler is not the same used in the experiment. The purpose of this picture is just for showing an example of air sampler.[10])

3.2 Experiment description

3.2.1 Tentative test

Two tentative test were conducted which aimed at getting familiar with hospital set-up, installation of equipment, people's responsibilities during the experiment. First tentative test was more about getting an idea whether the placement of equipment according to the first design is proper. In this test human being was used as the patient. In the second tentative experiment a piece of pork meat was used. Furthermore, in the second test, the final placement of people and instruments were decided.

3.2.2 Real test

In the real test, two live pigs as the patients were used. However, only the data from one experiment, performed on 10 December will be used. The timeline of the experiments is explained in Figure 3.1 in which total 3 surgeries were conducted in each day. The difference in each surgery was the temperature. The surgery started from 21 Celsius and ended up with 25 Celsius. In addition, the set up of both experiments were the same which is shown in Figure 3.2. In this Figure, placement and number of different equipment are shown. Also, Figure 3.2 contain information in where each personnel should stand during the surgeries. In Table 3.1, responsibilities of each person in the operating room are elaborated.

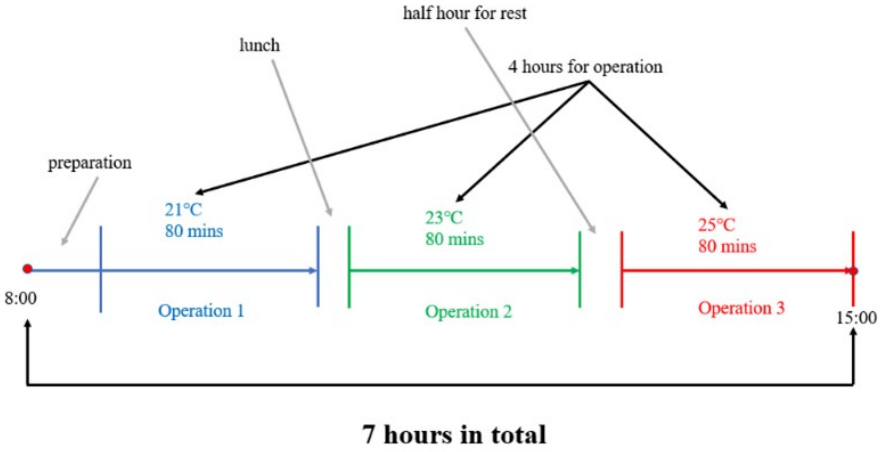


Figure 3.1: Schedule of the real experiment

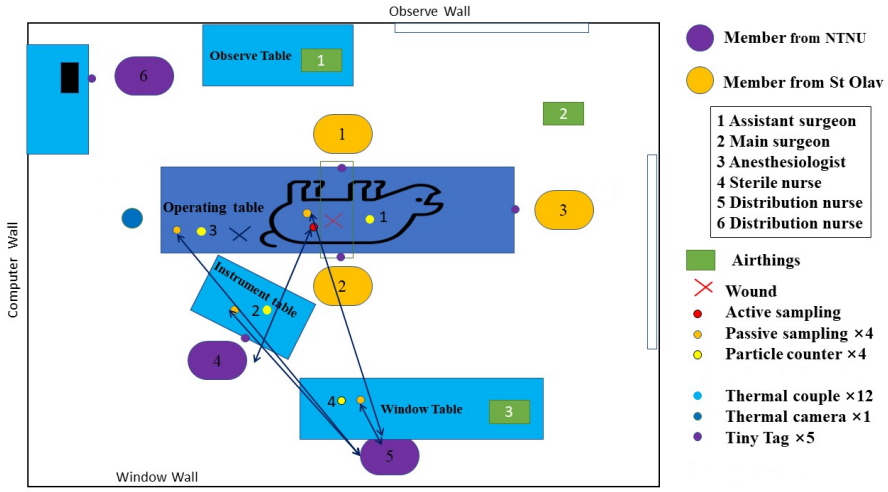


Figure 3.2: Personnel position and measuring points in actual animal experiment

Table 3.1: Personnel's responsibilities during the experiment

No of person	Responsibility
person 1	Help person 2 to cut two wounds and mount the thermal couple during stage 2
person 2	Cut 2 wounds and mount the thermal couple during stage 2
person 3	Anesthetize and monitor the pig's vital signs
person 4	Replace the active sampler from stage 1 to stage 3
person 5	Place the passive sampler and retrieve it one hour later

The experiment started at 8 in the morning, but before starting the surgery on the animal, active sampling was done twice for having the default setting. The detailed setup of the operating room is as followed:

- 3 aerotrak 9306 were located in the operating room: one on a table near the window wall, one on the instrument table, and one at the end of the operating bed. The description of the used items are the same as the second tentative experiment, except for the items listed below:
- 1 thermal camera was used to measure the temperature of 4 walls, floor, and ceiling every 10 minutes.
- 3 air thing were located at three different spots: on the window table, on a table near the observer wall and main door wall.
- 5 tiny tags were located on the computer table, instrument table, on the ceiling right above the surgical lamps, and two at each end of the operating table, near the surgeons.
- 10 thermal couples were located in the following positions: Four at each side of the wound, three inside the wound in different depths (one at three cm and one at one cm from the surface, and one on the surface of the wound). Finally, three thermal couples were attached on the floor, near the main surgeon, assistant surgeon, and anaesthesia. Since the operation was going to be performed in three different temperatures, one 10-cm cut was done on the pig at each temperature.
- Passive sampling: blood agars were placed in 4 spots according to figure 22, and were replaced every hour. Therefore passive sampling was done once during each operation
- Active sampling: this was done using an air sampler and the blood agars; the air sampler was located near the wound and was changed every 20 minutes.

In the meantime, the pig was closely observed, and the vital health signs (core temperature and heartbeat) were written down every 3-5 minutes from the health monitor in the operating room. During the breaks, surgical lamps were turned off to reduce the effect of cumulative heat. Moreover, in section 2.2 each of these equipment are elaborated thoroughly. Using the inforamtion from chapter 2 about KRI and RIF and the data, Table 3.2. is a conclusion of the work.

Table 3.2: Brief description of the gathered data from animal experiment

Type of data	Where have the data been gathered from?	Is it a input or output data?
Temperature	in 10 different places: 4 around the pig, 2 in the wound, 4 around the people near the animal	input
Relative humidity	Around the people who are near to the animal and air exhausts	input
Number of particles	In four different positions: near and far from animal	input
Thermal pictures	From the animal and the environment around it	input
Core body temperature	Hospital screens	input
Colony Forming Unit-Active(CFU)	On the operating table, near to the animal	output
Colony Forming Unit-Passive(CFU)	In four positions: near the pig and surgical staffs	output

3.3 Result

In this part, data gathered from a 6-hour surgery was explained and discussed. Using the equipment described earlier, all the data that are related to the operating room environment were collected. They include temperature of surfaces and areas near the pig, relative humidity in different places, and number of particles. Moreover, the level of colony-forming unite is also measured. Data has to be studied to understand if they are proper for further studies or not. This will be done in the next chapter

3.4 Summery

In this chapter, the experiment from which data will be used is elaborated by using the specialisation project. These surgeries serve multiple purposes. First, they help

in getting acquainted with the hospital and operating room's setting and how each piece of equipment can be installed properly. Second, they help gain knowledge on how each tool can be used. The main and most important purpose is to gather the data. The data gathered include the temperature of surfaces and areas near the pig, relative humidity in different places, and the number of particles. However, data gathering faced many limitations like budget and, more importantly, restrictions to gaining data from real human surgery. This section aims to provide information as a basis for future work about the experiment's setting. Therefore, using the information in this chapter, this experiment can be used as a basis for planning and adjusting experiments according to the needs.

Chapter 4

Theory and framework of data analysis

This chapter captures background knowledge and theory for thesis. It provides an introduction to data analysis, machine learning concepts, and evaluation metrics which are used throughout this thesis. It does not go in-depth, but it serves to give the reader the insight required to understand the concepts presented in this thesis.

4.1 Objective and challenges of data analysis

The total amount of data has grown extensively during the last decade. From 2010 to 2020, the amount of data in the world grew by 5,000%. However, this much raw data cannot be useful, and at best, it has a very limited value. Converting raw data into knowledge and useful information requires some steps which in science are known as data analysis. This field of science helps to get an insight into relations and correlations among the data. Many challenges are presented in analyzing the data. Many processes should happen in order to turn data into useful information. Like all industries, the healthcare industry produces huge data each year. However, it struggles to convert them into insights that improve patient outcomes and operational efficiencies. Data analysis in healthcare is intended to help providers overcome obstacles and spread the application of data-driven methods. This results in making healthcare data easier to share among colleagues and other partners, easier visualization for public consumption, providing accurate data-driven predictions in real-time to allow healthcare personnel to respond more quickly and accurately to changing healthcare markets and environments, and enhancing data among healthcare organizations to convert data into information.[11]

Meanwhile, data analysis in every industry can be challenging. These challenges that were presented in this data collection can be listed as below:
. In the previous chapter method to collect the data was introduced. According to that chapter, all the data expected for CFU level is collected manually. It starts by putting the blood agers into the air sampler, replacing the blood ager with a new one, labelling each blood ager, and sending them to the laboratory for further examinations. Along with all these processes, human error, and limitation of using blood ager, and time limitations for a lab to investigate them are clear. This resulted in fewer CFU samplers in each surgery. Another challenge was collecting data from a surgery that imitates a real operation. This poses a lot of difficulties.To imitate a real surgery, both the procedures done by surgeons and the behaviour of the healthcare personnel should be like a real surgery. During data collection, all the roles of nurses and surgeons were done by students and employers who were not medical students. Still, the lack of previous experience of how they behave in the surgery was problematic, for example, how much talking and moving is allowed? If talking is allowed, how loud can people talk? Another problem was due to the equipment that automatically logs the data. The ideal method is to log them in the shortest time intervals.
Nevertheless, this was not possible due to the different settings of the equipment. Data like humidity and particle counting were saved every 5 seconds, while thermal pictures were done every 10 minutes. Also, CFU, which was the output of the work, was logged every 10 minutes. This is clear that the inconsistency of data logging can lose a lot of valuable data. However, if CFU collection is done manually, it cannot be done every five seconds.
Moreover, some data were available in picture format, like thermal camera images. These images should be converted into readable numbers. Therefore, data collected from the surgery cannot be used immediately for analysis. First, the problems in the dataset should be solved to make it consistent. The theory on how these challenges can be solved as much as possible is discussed next.

4.2 Data analysis process

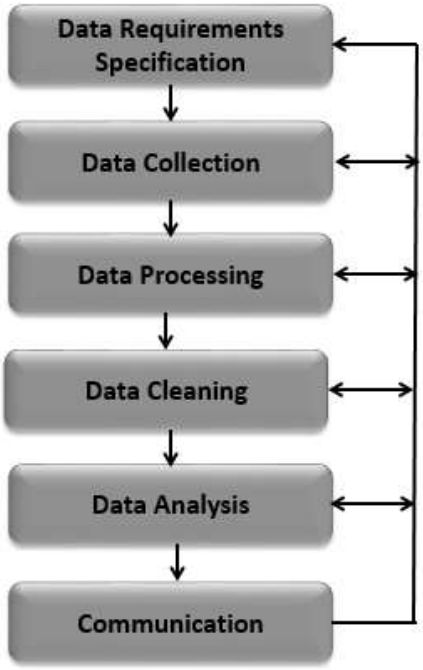


Figure 4.1: Data analysis process

Source: [12]

Based on the discussion in the previous chapter, data can have flaws. The data gathered for this work is not flawless as well. Therefore, Figure 4.1 is the road map of this study before implementing machine learning methods. First, each of these steps have to be discussed generally, and then define how this step is related to this study.

- Data Requirements Specification
  The question of the experiment defines the required data for analysis. Based on the requirements set by those who direct the analysis, data necessary as inputs to the analysis are identified.[12] This section was discussed in the specialization report by doing literature research on what environmental factors affect the occurrence of infection

- Data Collection
  Data Collection gathers information on desired variables defined in the previous step, data requirements specification. The emphasis is on ensuring accurate and precise collection of data. Data Collection ensures that data gathered is precisely such that the related decisions are valid.[12] For this study, data collection was also conducted as a part of the specialization project, where two

training surgery and two real animal experiments were conducted in St.olav's hospital.

- Data Processing
  The collected data must be organized for analysis. This includes structuring the data based on the requirements of the analysis tools. For example, the data might have to be placed into rows and columns in a table.

- Data Cleaning
  The processed data may be incomplete, contain duplicates, or contain errors. Data cleaning phase is common when two or more databases are merged. Data cleaning is the process of correcting these errors. For this study, since the data sheets were so widespread and different data were measured in different periods, the data cleaning phase was assigned to make different data compatible with each other.[12]

- Data Analysis
  Data which is processed, organized and cleaned is ready for analysis. Various data analysis methods are available to understand, interpret, and derive conclusions based on the requirements. One important and useful technique is data visualization. This method is used to examine the data in a graphical format to obtain additional insight regarding the messages within the data. Statistical data models such as regression can identify the relations among the data variables. These descriptive models help simplify results. The process might require additional data cleaning or additional data collection, making all these activities iterative.[12]

The first two step, data requirements specification and data collection are done in the previous part of this study, therefore, in the report, only data processing, data cleaning and data analysis will be discussed. Communication step is out of the scope of this study, therefore will be eliminated.

To start analyzing the gathered data, the available data has to go through processing and cleaning phase. As said, in these phases, data will be arrange in the required format. For this project, these phases included on how to extract required data from huge excel sheets and make them ready for study. Also, some of the data were not available in any readable format for computer (like in transferring data the from documentations and pictures). Therefore, these data had to be set and arranged. The items below are the suggested methods on how to convert data to a useful dataset.

- Data integration is combining data available in different sources and provide

users with a unified view of the data. It results in a coherent storage.

- Data transformation is a data migration technique that converts data from one format to another by normalization and aggregation.
- Data reduction is generating reduced views of the data with no impact on the analytic results. This technique includes data compression, clustering and dimension reduction.
- Data Cleansing refers to the process of searching, identifying, and correcting errors. It determines and identifies inaccurate, incomplete, or unreasonable data and then updates, repairs or deletes these data to improve quality [13]

In all data-based works, before any attempt for analysis, data needs to be investigated and prepared carefully. In hospital experiments, several types of data were gathered. In Table 3.2., there is a brief description of each data and The first column of the table is written based on results of the experiment, and the rest is based on the discussions with experts and literature review around this topic. All the data sets except one, are available in txt or csv format which can be converted to xlsx format for easier access and readability. The one data which was not presented in any of the above formats, is the data related to thermal camera pictures. In section 4.3 it is discussed how to extract the required information from a thermal image.

4.3 Thermal image data processing

One type of data needed for modeling, is the information from thermal camera. These thermal pictures were taken every 5 minutes using Bosch PD1. As said, this equipment is a thermal detector, based on infrared technology, measures surface and room temperature. This device was used to collect information of temperature above the operating table.

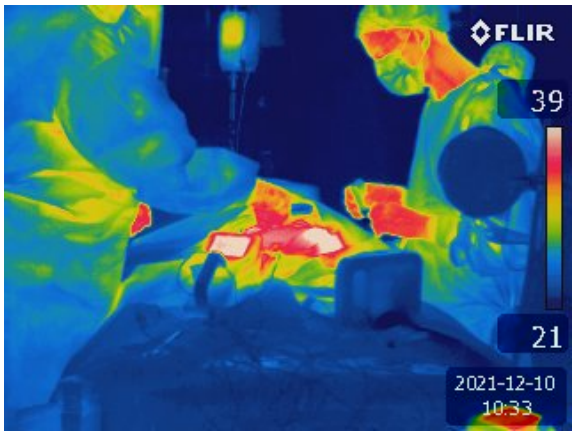


Figure 4.2: An example of thermal camera pictures

As seen it the picture, the range of temperature varies from the lowest, 21 celcuis to the highest 39 celcuis. The aim here is to extract temperature data. Based on the discussion with the expertise, there are two possible options for thermal pictures:

- If the heat source in the image is only the human body. The high-heat parts can be easily segmented by image threshold segmentation algorithm, and then the corresponding temperature can be matched. This case is simple because both the image threshold segmentation algorithm and color to temperature value mapping are easy to implement.
- If there's some other heat source in the image besides the body. It is necessary to identify parts of the human body first, which is a little difficult and may need deep learning method, such as U-Net and DeepLabv3, etc.

For deciding which method can be used to extract the thermal information, the following logic is considered: At first glance, case number 2 is relevant to this study, since there are non-human objects in the pictures. However,by reviewing the thermal pictures, it can be seen that the non-human objects do not vary in temperature that much over time, so by simplification, case 1 can be implemented.

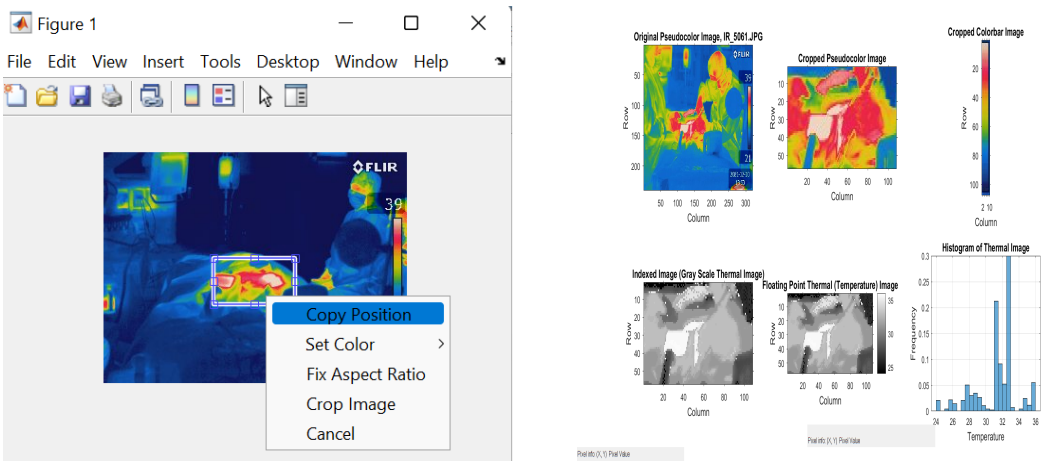
4.3.1 Converting thermal image into corresponding temperature

For this part of the work the latest version of MATLAB, MATLAB R2021b is used. By searching in MATLAB Answers, a prepared MATLAB code was retrieved from the website. The code was used for extracting thermal information from thermal pictures. However, some modification in the code was applied to fit the purpose.

This code can be founded in the appendix B. Aside from the main code, an auxiliary code for cropping the desired area in the thermal pictures was used. This code can be found as below: minted

```
I = imread('thermalPicture.png');
[J,rect] = imcrop(I);
```

The first line of the code opens the picture, and the second line opens a additional tab where one can move the cursor over the image, it changes to a cross-hairs. The crop image tool blocks the MATLAB command line until the operation is completed. After finding the position of the desired section, coordinates can be founded and replaced in the converter code. After replacing the coordinates of each picture like in Figure 4.3.a, the code in appendix B has to be ran. The result is like Figure 4.3.b. This value has to be calculated for all the pictures and in different positions. All the temperature values are shown in Table 4.1.



(a) An example of how the cropping code works

(b) Conversion of thermal image into temperature

Figure 4.3: Result of the code

Table 4.1: Temperature results from thermal images

Time	21 Celcius	Time	23 Celcius	Time	25 Celcius
10:08	28.52	11:44	30.78	13:26	31.19
10:13	29.09	11:49	31.24	13:27	31.45
10:18	28.9	11:57	31.03	13:33	31.59
10:23	28.48	12:00	30.67	13:38	31.47
10:28	29.95	12:05	30.8	13:43	32.01
10:33	30.68	12:10	31	13:48	31.01
10:38	30.17	12:15	30.78	13:53	31.34
10:44	30.01	12:20	30.79	13:58	31.14
10:49	30.11	12:25	30.32	14:03	31.89
10:54	30.59	12:30	30.3	14:08	31.67
10:59	30.72	12:35	31.21	14:13	31.67
		12:40	31.18	14:18	31.87
				14:23	31.51
				14:28	32.29

4.4 Machine learning

4.4.1 Need of machine learning

The simple answer to why we need machine learning is that machine learning uses data to help predict more accurate and precise. In the healthcare domain, these decisions can be the diagnosis of illness, choice of medication, treatment, predict population health risk by identifying patterns, surfacing high-risk markers, model disease progression, and more. When data is available, decisions can be made for the future by searching into the historical pattern.

This thesis implements machine learning since it is a cutting-edge technology with wide use cases. Using machine learning to find whether a high infection will threaten patients' safety has many advantages. First of all, since infection after surgeries develops a few days after the surgery, hospital care can be offered in advance to mitigate the risk of further danger. Moreover, by knowing what condition leads to infection, hospitals can avoid having this condition in future using different measures. This can save hospitals from spending resources and money on a health complication that could have been mitigated first. Knowing what advantageous machine learning brings for decision making in healthcare, the section discusses how to start practices in machine learning.

4.4.2 Starting with machine learning

When it comes to machine learning techniques, there are plenty of them that can be implemented on any data set. In Figure 4.4, branches of machine learning method is shown.

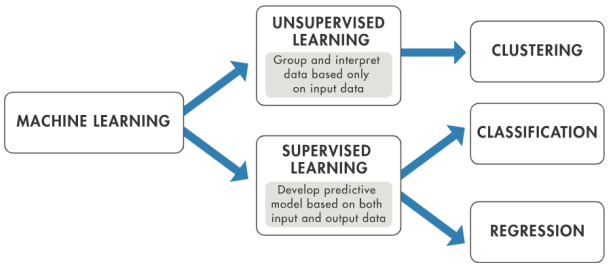


Figure 4.4: Summary of machine learning techniques

Source: [14]

For choosing whether clustering, classification and regression should be used, an other look to the data is necessary. Also, there is no best method or one size fits all. Finding the right algorithm is partly just trial and error. In some cases, even highly experienced data scientists cannot tell whether an algorithm will work without trying it out. But algorithm selection depends on the size and type of data, the required insights from data, and how those insights will be used. Considering the dataset and the fact that input and output are available, the question is to find the pattern that connects inputs to output(s). As a result, this problem is a supervised learning problem. "Supervised machine learning builds a model that makes predictions based on evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data. Use supervised learning if you have known data for the output you are trying to predict. Supervised learning uses classification or regression techniques to develop machine learning models."[14]
It is vital to know which kind of problem is our problem. Classification techniques predict discrete responses—for example, whether an email is genuine or spam, or whether a tumor is cancerous or benign. Classification models classify input data into categories. Typical applications include medical imaging, speech recognition, and credit scoring. On the other hand, regression techniques predict continuous responses. For example, changes in temperature or fluctuations in power demand.[14]
In this project, the aim is to understand under a specific condition, CFU can be dangerous for patients' health or not, so the aim is not to calculate CFU level. As a result, the output is a discrete YES or NO,making this problem a classification problem.

Classification learning can utilize many algorithms. Common algorithms for performing classification include support vector machine (SVM), boosted and bagged decision trees, k-nearest neighbor, Naïve Bayes, discriminant analysis, logistic regression, and neural networks.[14] This section first discusses classification learner app in Matlab. Using the initial dataset in the app results will be discussed. An the end measures will be suggested to improve the work.

4.4.3 Introduction of the classification learner app in MATLAB

Classification is a type of supervised machine learning in which an algorithm learns to classify new observations from examples of labeled data. To explore classification models interactively, classification learner app gives a great flexibility to pass predictors or feature data with corresponding responses or labels to an algorithm-fitting function in the command-line interface. Inside this app classification can be implemented using different algorithms.[15] The steps below states how the app works.[16]
1- Select data and validation
2- Choose classifier option
3- Train classifier
4- Assess classifier performances
5- Export classifier

4.4.4 Measures for evaluation of machine learning methods

This section describes measures to evaluate performance in binary classification, that is when there are only two classes, two possible outcomes. The machine learning models listed in the previous section have various characteristics and behaviour,Therefore, a set of tools applicable for assessing the performance of any machine learning model is important. To measure the performance of the models, precision and recall are calculated to gauge the number of false positives and false negatives, respectively. The precision and recall are also combined to calculate the F-measure and kappa statistic, to further help evaluate the performance of the models. The kappa statistic is an interesting metric that compares the accuracy of the model with the expected accuracy of a guess. These popular quality control measures are defined below.[17],[18]

Confusion matrix

Supervised machine learning classifiers have several evaluation metrics to choose from. Many of them come are connected to a confusion matrix which records correctly and incorrectly classified samples from both classes. The metrics following will use confusion matrix in the definitions.

Table 4.2: Confusion matrix

	Actual positive	Actual negative
Predicted positive	TP	FP
Predicted negative	FN	TN

Accuracy

Accuracy, as shown in equation 1, indicates how well a machine learning algorithm can label data points out of all the data points. In the scope of this study, it means whether the algorithm can distinguish between dangerous and safe levels of CFU.

Precision, recall, and F1

Precision is a measure of how well the algorithm can find true positives. In this project, it means if the algorithm is capable of finding out the dangerous level of CFU. Recall is a measure of how reliable can the classifier identify all true positive samples. For example, the recall of 30% means that only 30% of the dangerous levels were distinguished, remaining were unable to be identified by the algorithm. Ideally, a good classifier should maximize both precision and recall. However, in reality, there is often a trade-off between precision and recall. This trade-off is important when data experts want to choose among models. As shown in figure 4.5, increasing recall means decreases precision, and vice versa.[19] Equations 2,3,4 calculate precision, recall, and F1 respectively.

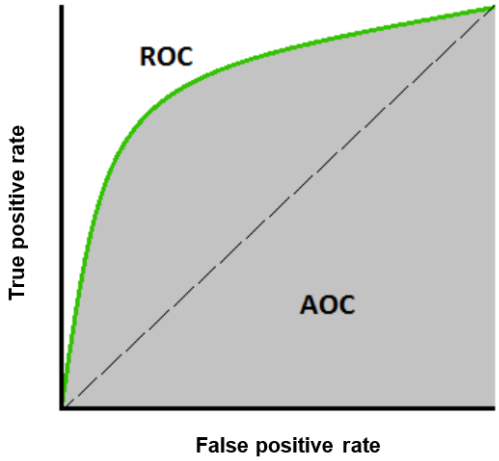


Figure 4.5: Relationship between precision and recall

Source: [19]

Each of these risk criteria have utilization in a specific situation. These metrics help to select the best machine learning algorithm based on the need of the industry. For this project, for example, recall can be used when any trespassing from CFU threshold can have severe consequences, for example, in head and throat surgery in a patient with many health complications. The next step is how these criteria can be calculated.

First, confusion matrix is required, and its components have to be introduced. The confusion matrix has four categories: True positives (TP) are examples correctly labelled as positives. False positives (FP) refer to negative examples incorrectly labelled as positive. True negatives (TN) correspond to negatives correctly labelled as negative. Finally, false negatives (FN) refer to positive examples incorrectly labelled as negative. F1 Score is needed when a balance between precision and recall is required. F1 Score perhaps is a better measure to use if a balance between precision and recall is needed and there is an uneven class distribution (a large number of actual negatives).[19]. Table 4.2. gives a view of where each of these components is located in a confusion matrix.

Each of these metrics can be calculated using formula below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{4}$$

4.5 Summery

In this section, a literature review of about data analysis process was performed. The goal was to clarify how the project should be started. Different methods in data processing and data cleaning were introduced. These methods are helpful since the number of variables is too many. The first step of data cleaning was started in this section, where the MATLAB code for finding temperature from a thermal image was performed. In the next section, the purpose is to understand the behavior of data.

Chapter 5

Data visualization

This chapter focuses on make plots of inputs and outputs. These plots show the changes of each variable against time. The aim is to get familiar with the behaviour of the dataset. The trend in each variable help to figure out if the data should be included in further studies or not.

5.1 Converting data into plots

When data is processed, organized and cleaned, then it would be ready for analysis. In the data analyzing phase, the answer to the question of how much and how input data (temperature of different settings, humidity, particles) affect the output data (CFU level) will be put in more clear vision. The first step in data analysis is to understand the behaviour of data. One easy way to do so is data visualization. Data visualization helps in grasping how a dataset behaves and how different data evolve individually and even whether they are connected. Data visualization represents data graphically by using visual elements like charts and graphs to provide an understandable way to see and understand trends, outliers, and patterns in data. Since the data is recorded in different time- series, all the data are plotted against time to understand their progress over time. In many cases, it helps to alter "unprocessed information" into useful information. Visualization is used to represent abstract data, like business data, while scientific visualization presents scientific data, which are usually physically based, like human body, the environment or the atmosphere. Both of these categories focus on transforming data into a visual form, to become understandable information for gaining insight and knowledge. This work is a data visualization in a scientific format.[20] By looking at the dataset in excel format, it is shown that 37 columns of different data are measured in every 3 temperatures. Due to the large number of columns, data visualization helps to figure out

whether all the columns of data are important for study. It has been discussed that data analyzing and finding out the relation between different variables can be done by machine learning techniques. Machine learning methods have different levels of complexity and effectiveness. However, for implementing them, it is wiser to start from the simplest methods and step by step add to the complexity to achieve the desired outcome.
In this project, data visualization is being done using Excel and is based on the data gathered from the experiment done on 10 December 2021. In the next part, plots will be discussed.

5.2 CFU level

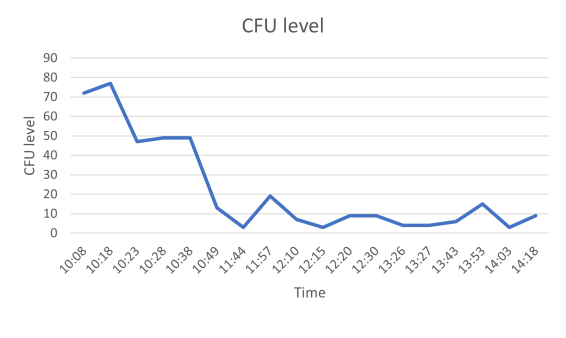


Figure 5.1: CFU level

Figure 5.1 shows how active level of CFU changes over time. Although the level of passive CFU was also measured, but after studies and advice from the experts from machine learning field, it has been decided that in this project passive level of CFU will be disregarded. The reason to do so, was that passive level of CFU was sampled only once in each surgery, making it in total three in the whole experiment.
On the other hand, this plot shows how CFU level is high at the beginning of the experiment and decreases in one hour and a half. After the noticeable decrease, the level of CFU stays the same till the end of experiment. The reason of this significant high level can be due to the preparations and people's movements before the experiment.

5.3 Temperature

5.3.1 Body temperature

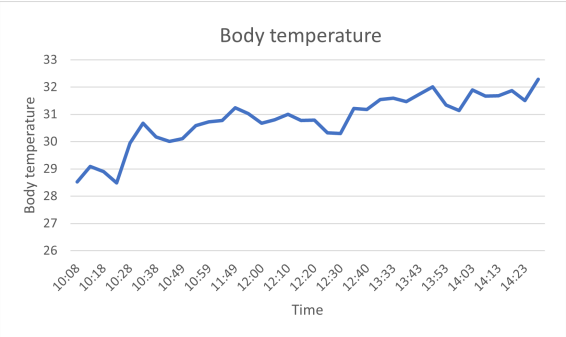
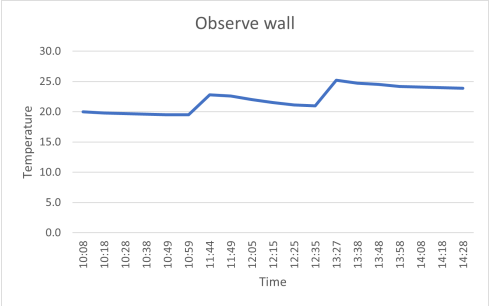


Figure 5.2: Body temperature

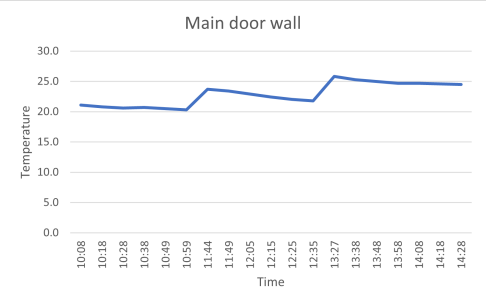
Figure 5.2 shows the trend of animal body temperature where the surgery was performed. This data is gathered from the thermal camera installed in the operating room. The details of this data gathering was discussed earlier in section 3.4.This data should not be misunderstood by core body temperature derived from the surgical screens. The whole trend is increasing which is due to the reason that animal is located under surgical lamps the whole time.

5.3.2 Surface temperature

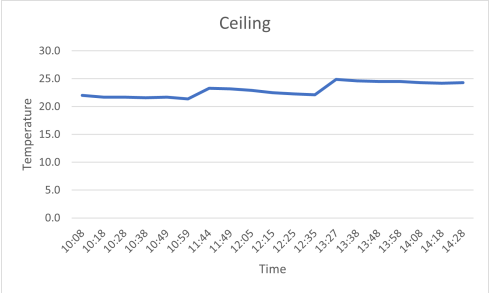
Temperature in different surfaces in the room, like wall, window, floor, and ceiling reflects the temperature set for the operating room by the hospital personnel. In Figure 4.4 temperature changes in different surfaces are shown.Surface temperature reflect the temperature set for the operating room. All surfaces start with a temperature around 20 degrees. Each small increase that happened two times in each plot, reflect the temperature increased by the hospital personnel. So, by the end of first surgery the temperature ended up to 23 degrees, and by the end of the experiment it was 25 degrees.



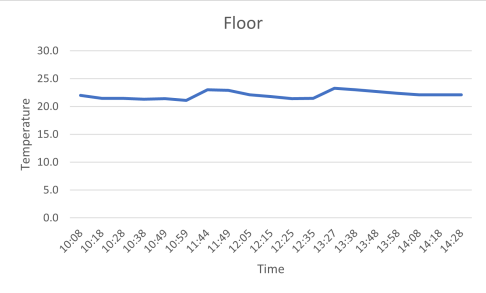
(a) Temperature of observe wall



(b) Temperature of main door wall

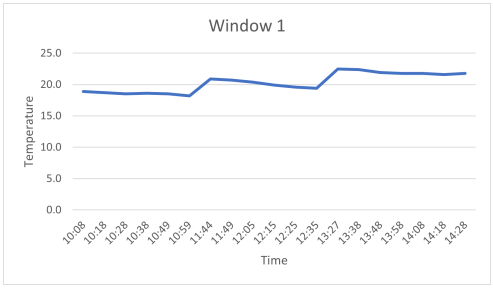


(c) Temperature of ceiling

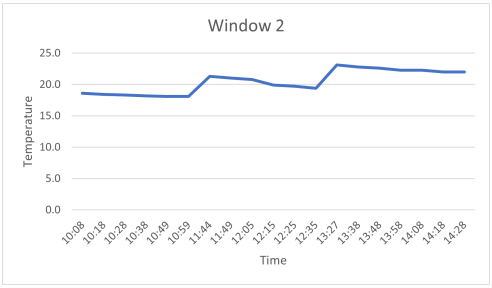


(d) Temperature of floor

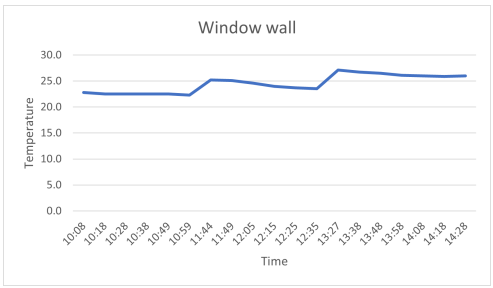
Figure 5.3: Temperature changes in two of the surfaces



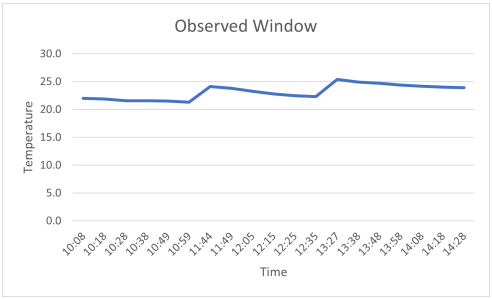
(a) Temperature of window 1



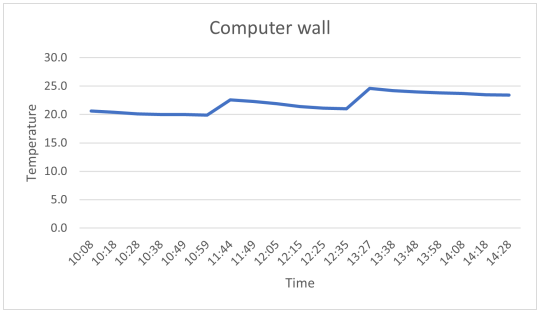
(b) Temperature of window 2



(c) Temperature of window wall



(d) Temperature of observe window

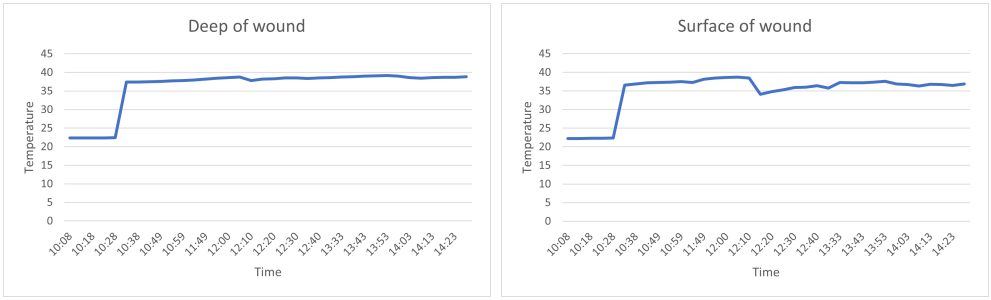


(e) Temperature of computer wall

Figure 5.4: Temperature changes in surfaces

5.3.3 Wound temperature

Temperature of the wound measured by thermometers in two different depths, shows how at the beginning of the surgery it increases rapidly but no significant increase can be detected afterwards. The temperature of the surface of the wound and deep of the wound remain almost the same after the huge jump in the beginning of the experiment. Based on the observations in the operating room, it is suspected that surgical lamps are the main source of increase in body temperature. Therefore, if it is proven that wound temperature has effect on the occurrence of surgical site infection, more studies have to be conducted to study the importance of the position and strength of surgical lamps.

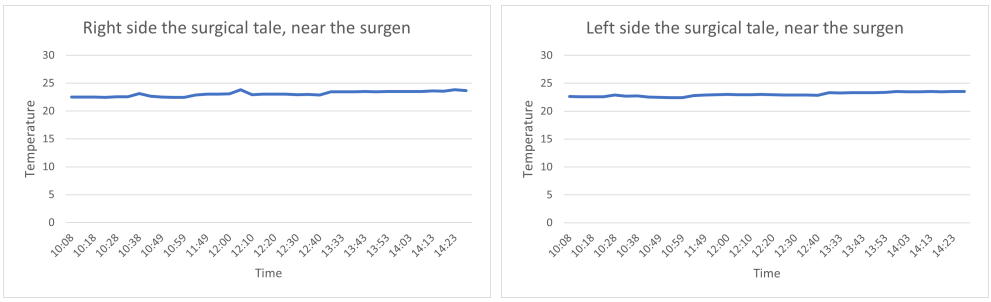


(a) Deep temperature of the wound (b) Surface temperature of the wound

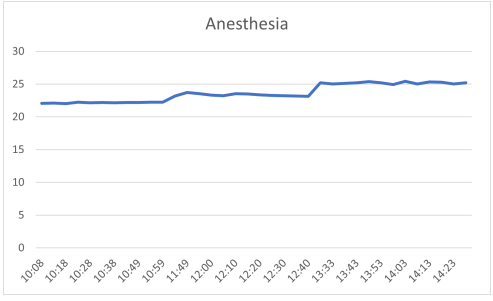
Figure 5.5: Temperature of the wound

5.3.4 Temperature around the personnel around the surgical table

Based on Figure 5.6, temperature around the operating table did not experience almost any changes. However, by comparing to changes in surface temperatures, walls and other surfaces showed more change. But, in general it seems that the temperature almost stays the same during the operation. Since, the CFU level is fluctuating over time, it means that the variables that do not change over time cannot influence on CFU level. Therefore, all the variables in this section will be eliminated from further studies.



(a) Temperature of the right side the surgical tale, near the surgeon (b) Temperature of the left side the surgical tale, near the surgeon

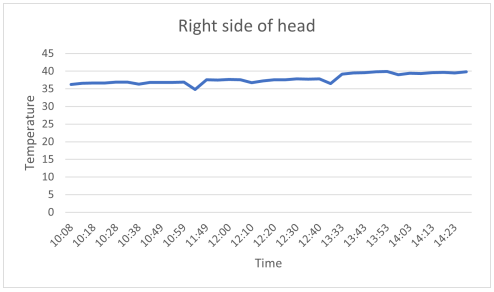


(c) Temperature of around anesthesia

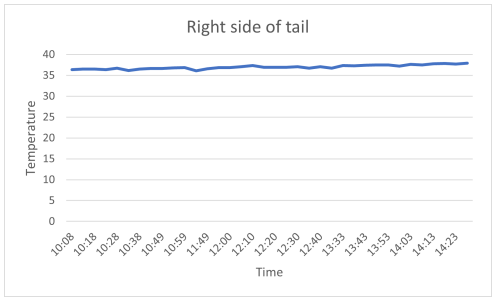
Figure 5.6: Temperature around the operating table

5.3.5 Temperature around the animal

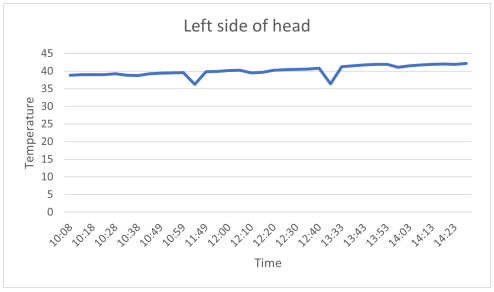
Based on Figure 5.7, no significant increase or decrease can be detected. This pattern can be seen also in Figure 5.6, where temperature around the surgical table is measured. Temperature around the animal also showed no significant change. Since, the CFU level is fluctuating over time, it means that the variables that change over time cannot influence on CFU level. Therefore, all the variables in this section will be eliminated from further studies.



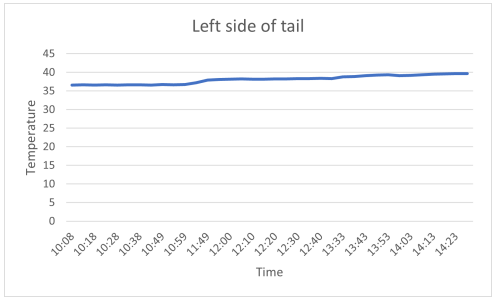
(a) Temperature around right side of the head



(b) Temperature around right side of the tail



(c) Temperature around left side of the head



(d) Temperature around left side of the tail

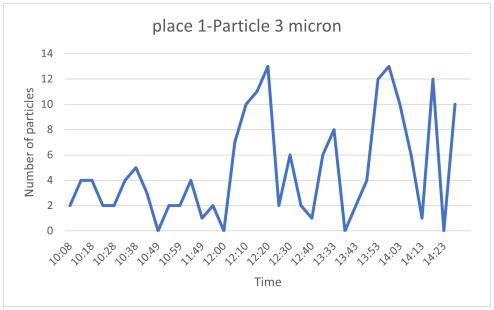
Figure 5.7: Temperature around animal

5.4 Airborne particles

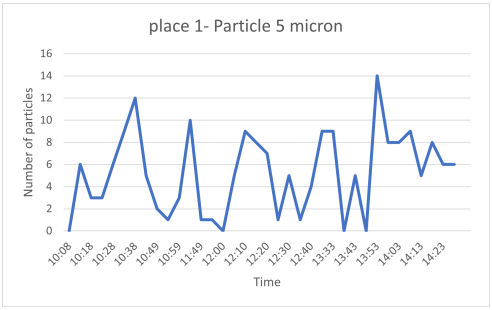
5.4.1 Level of airborne particles

A study has shown that airborne microbial contaminants are an important source of SSIs in clean operation.[1] Number of particles,s measured by Aerotrak 9306, which were placed in 4 different places. Studies have suggested that activity level is an important factor in changing the level of CFU.[1] They show how due to human activity, particles are released into the air. Particles having 0.3, 0.5, 1, 3, 5, and 10 microns in diameter were measured. However, since only big scale particles (3,5, and 10 micron) can carry bacteria, the plot of these particles is shown in this part,

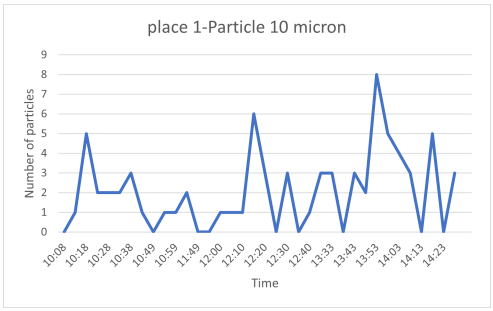
and others are disregarded from the study. By looking at the trend of particles, it can be concluded that there is no pattern that can be relied on. So, only based on these plots, future trends of particles cannot be imagined. The reason for this is because of the uncertainty in the nature of the surgery. This makes people's movement unpredictable, resulting in chaotic particle release. In all places, at the end of the experiment, around 14 o'clock, a huge increase in all particle levels can be seen. Based on the observations during the experiment, the blanket which was covering the pig had to be removed due to high and dangerous body temperature of the animal. The reason to do so was that this body temperature was so high that remaining at that temperature for the animal was lethal. Aside from that, airborne particles do not follow a specific pattern. It is suspected that when the level of human activity is high, more particles will be released into the air, so at the end of the surgery, while taking off the gloves and surgical gowns, this number has to be high.[9]



(a) Level of 3 micron size particles

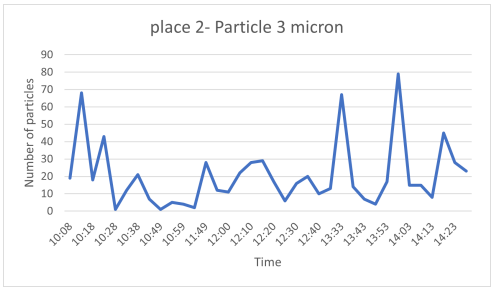


(b) Level of 5 micron size particles

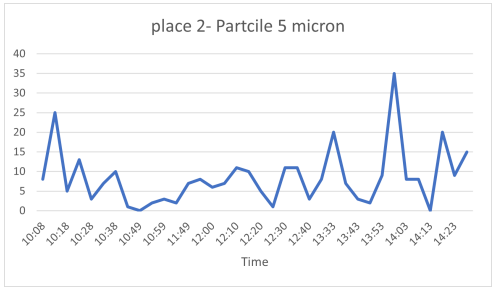


(c) Level of 10 micron size particles

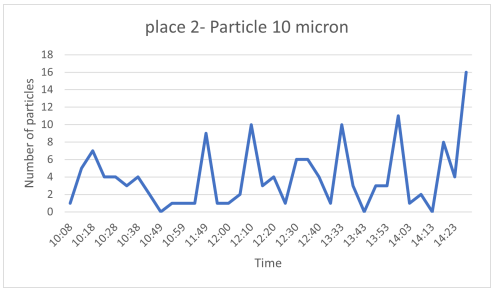
Figure 5.8: Level of airborne particles in place 1



(a) Level of 3 micron size particles

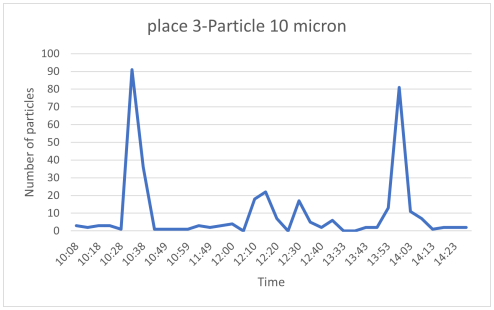


(b) Level of 5 micron size particles

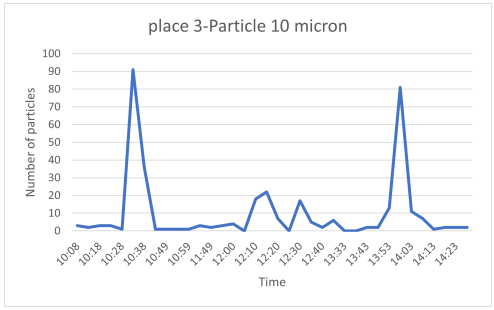


(c) Level of 10 micron size particles

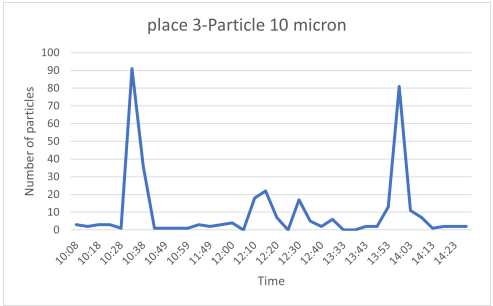
Figure 5.9: Level of airborne particles in place 2



(a) Level of 3 micron size particles

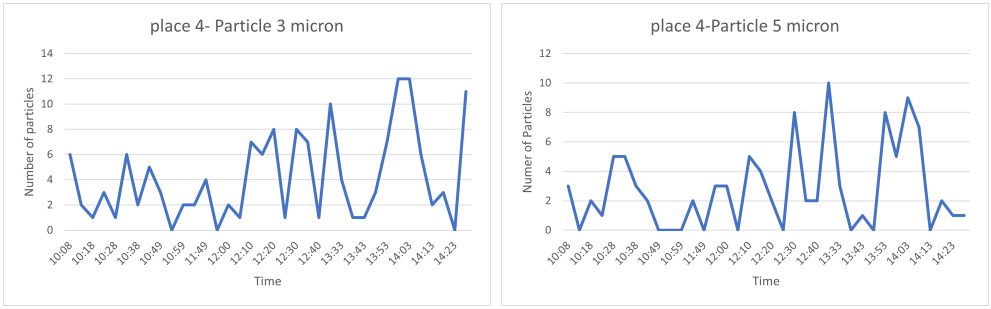


(b) Level of 5 micron size particles

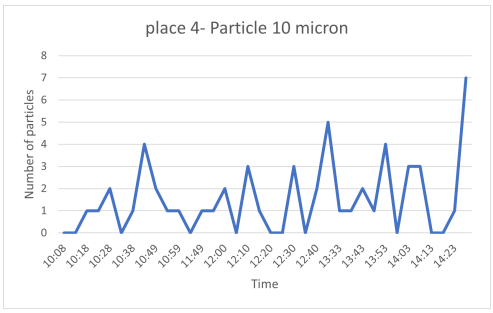


(c) Level of 10 micron size particles

Figure 5.10: Level of airborne particles in place 3



(a) Level of 3 micron size particles (b) Level of 5 micron size particles



(c) Level of 10 micron size particles

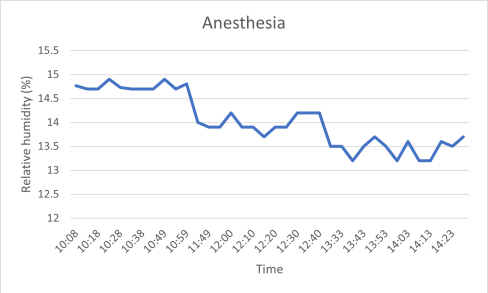
Figure 5.11: Level of airborne particles in place 4

5.5 Humidity

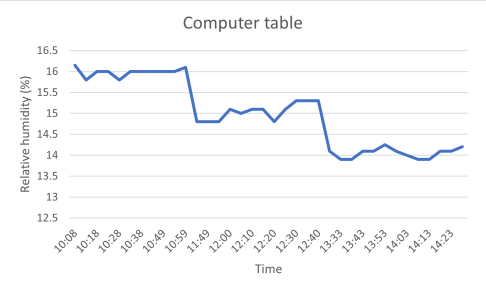
5.5.1 Relative humidity

Air humidity must be kept at acceptable levels since it is closely related to the space hygiene and thermal comfort conditions. High humidity levels favor the growth and transfer of bacteria that can easily become airborne on water molecules. In this study, high humidity level is not in favor as it increases the chance of moving of bacteria. However, low humidity level is dangerous as well, but it is more related to thermal discomfort and the safety of electrical equipment. Therefore, not discussed here. Based on ASHRAE, the recommended relative humidity level is between 30% to 60% for the temperatures between 20 to 24 Celsius. [21] Humidity in different places in the operating room, measured by tiny tags, shows that no specific relation can be founded. In all four places, humidity starts around 15.5 on average. However, in all three places, except for the left surgical table, the humidity experience a lot of fluctuations. The overall average humidity at the end of the experiment is 15, which is different from the beginning. The humidity in all places except for the anesthesia will be eliminated since humidity around anesthesia is the closest one to the patient. Also, two of the other humidity levels are so similar to anesthesia. The humidity around the left surgical table seems so far from the patient that it loses its effect,

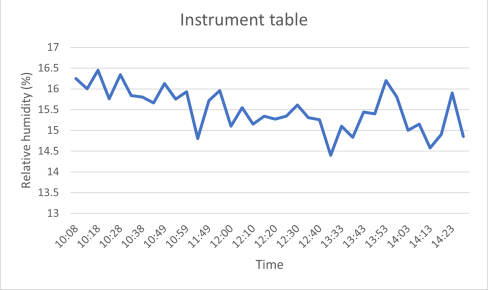
despite having a strange pattern.



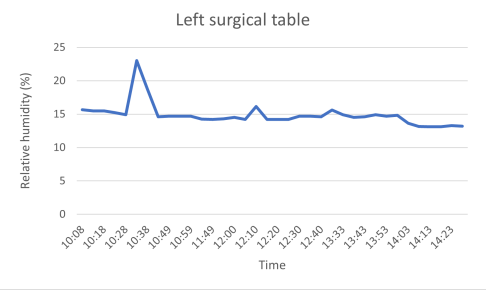
(a) Relative humidity around anesthesia



(b) Relative humidity around anesthesia



(c) Relative humidity around computer table



(d) Relative humidity around left surgical table

Figure 5.12: Relative humidity

5.6 Summery

This chapter visualises initial data in Excel to reach two goals. First, to get familiar with the evolution of data, and second is to investigate whether by only looking at different information presented by data, any knowledge about them can be retrieved. Based on the plots, most of them showed a chaotic trend during the whole time of surgery, and no repetitive trends were found. Also, ,no explanations could be given to justify the chaos in these plots. However, some plots follow more organized change. Finally, some plots were almost the same in the procedure and showed a little change.

The main objective of this chapter is to determine which variables have a potential effect on CFU. Since CFU level is very fluctuating, variables that stay the same during the surgery are irrelevant to CFU level, so they can be eliminated from the next phase. In this regard, variables related to temperature around the animal, around the personnel and surface are omitted, and their effects on CFU level will not be discussed. Moreover, humidity is collected from 4 different places, but all will be eliminated except one. By considering the large number of variables and difficulty in explaining the trend, the next step is to use machine learning techniques

to understand whether any useful knowledge can be extracted or not. The useful tools for machine learning will be discussed in the next chapter.

Chapter 6

A tentative test on initial data

This chapter serves the purpose to describes the framework used to process the data and build the machine learning models. It section uses the initial data to perform a supervised machine learning. This section only serves as a tentative exercises to discusses the problems within the initial dataset. Therefore, it highlights the problems and guides the thesis into future steps.

6.1 Implementing classification learner app

The classification learner app requires two types of data. One as predictors (currently the number of predictors are 18, but the maximum number of predictors has to be 15, since regression fitting cannot be done with more than 15 variables.) and response (CFU level). However, the response data needs to be a vector made of either 0 or 1. For our problem it is considered that 0 means CFU level is not dangerous and 1 relates to dangerous CFU level. The decision on which level of CFU is dangerous is decided by the standards determined by ISPESL in 2009 for air microbial contamination in operating theatres with turbulent which states maximum acceptable levels were taken as the air flow: less than 35 CFU/m3 at rest and less than 180 CFU/m3 in operational.[22] By considering the intensity of activity during surgery, above 35 CFU/m3 is considered as dangerous, therefore dedicated number 1 to it. Here, the number of observations is very limited, 22, which is very likely to cause problems. Before doing classification, data has to be imported into MATLAB software. For this purpose an Excel sheet compromising predictors and observer has to be provided. Each predictor and response is transferred into workspace in MATLAB.The next step is to import all the data into classifier app. After importing the data, predictors and response should be selected. To evaluate the classification, test data has to be provided as well. Therefore, the initial dataset is divided into two

subsets, train and test. The initial dataset was small at the beginning, therefore by splitting it into train and test set, the whole classification process may face a lot challenges. These challenges and the reasons will be discussed in analyzing results part.

6.2 Feature selection

There are many difficulties when more than five independent variables are in a regression equation. One of the most frequent is the problem that two or more of the independent variables are highly correlated to one another.[23] Currently, the dataset consists of 18 variables, which is already reduced from 33 using the result from visualization. The number of input variables has to be reduced for two main reasons: first, in many apps in MATLAB, it is not possible to add more than 15 inputs, and secondly, these 18 variables will have interactions if put in an equation, making the problem complicated. So, due to MATLAB limitations and for work simplicity, 18 variables have to be reduced. To reduce the number of data to 15, feature selection tool from statistics and machine learning toolbox in MATLAB was used. Feature selection reduces the dimension of data by selecting only a subset of measured features to create a model. Feature selection algorithms search for a group of variables that optimally model measured responses. The main advantage of feature selection is to improve prediction performance and provide faster and a better understanding of the data generation process.[24]
There are several feature selection functions in MATLAB that perform selecting variables based on different factors. These functions are separated also based on the supported problem (classification and different types of regression) and supported data type (categorical and continuous features).[24]
Now, the discussion is directed to which feature selection function to use, how to implement it, and which variables will be selected. The selected function is "relieff", which works both for regression and classification and all types of data. This algorithm chooses the data based on k nearest neighbours. The code line below is how relieff function should be written:
[idx,weights] = relieff(X,Y,k)
where X is predictors, Y is response, and k is the number of neighbors. If Y is numeric, relieff performs RReliefF analysis for regression by default. Otherwise, relieff performs ReliefF analysis for classification. Since the response values are numbers, relieff performs for regression analysis.
By running the codes below in command window, the 15 predictors are selected:

```
[idx,weights] = relieff(data,CFU,10);
```

idx(1:15);
The result indicates all variables related to number of particles in place 4 should be eliminated.

6.3 Analyzing the dataset

After selecting 15 variables the dataset was ready to be imported to MATLAB classifier tool. In this part, deciding which algorithm gives the best possible result is challenging. So, the it was decided to use the the option of training all algorithms.Let us consider one of the result of one of the algorithms:

Table 6.1: Summary of confusion matrix or algorithms run on the initial dataset

	TP	FP	FN	TN
Tree	4	0	1	17
KNN	0	0	5	17

Table 6.2: Accuracy, precision and recall, and F1 for initial dataset

	Accuracy	Recall	Precision	F1
Tree	95%	80%	100%	90%
KNN	75%	0%	-	-

Looking at table 6.2, results are not satisfying at all. Although accuracy looks logical, but as said before, accuracy is not a trustworthy criteria in controversial datasets. So, other metrics are observed. For Tree, precision is 100% and recall for KNN is 0%. When precision is estimated as 100% on a training set, implies overfitting. In KNN algorithm, TP is 0, therefore, recall and precision are 0, resulting in F1 being 0. It makes the classifier useless because the classifier cannot predict any correct positive result.
When working with small dataset, like this experiment, with 22 observation, machine learning will not give a good result. In the current dataset, the lack of enough data and presence of noise is visible. Considering these facts and the results from running the algorithms on data set also approves the existence of a improper data. [25] In the following paragraphs, the proposed methods to stop overfitting are introduced:
1) Regularization technique is proposed to guarantee model performance while dealing with real-world issues by feature-selection and by choosing useful and less useful features. Two methods of Regularization are Bayesian ANNs (BANNs) and early

stopping that favor models with smooth decision boundaries. The third technique which will be used in this report is noise injection. Noise injection penalizes complex models indirectly by adding noise to the training dataset.[26] Another method which is not consider completely as regularization, but sometimes can be assumed part of it, is feature selection that only the useful features and remove the useless features from our model.[27]
2) network-reduction technique is used to remove the noises in training set
3) data-expansion technique is used for complicated models to fine-tune the hyper-parameters sets with a great amount of data [27]

6.4 Summery

By finding out that the current dataset is not suitable for creating a classification algorithm for future predictions, it is decided to do expand the dataset. one way for data expansion is by by performing more experiments, but due to limitations in time and budget, cannot this be done. Also, data accessibility through different sources in Internet is not applicable. So, data has be generated using different algorithms.Therefore, the question of what algorithm can predict CFU level will be on hold and the in the next chapter, the answer to the question of how more data can be generated will be answered.

Chapter 7

Expanding the dataset

This chapter introduces different methods, for predictor and observer, to enlarge the dataset, then discusses which method is practical for this work, and implements the strategy.

7.1 Methods in data expansion

Indicating how many observations are required is a difficult and in some cases, an impossible task to do.[28]. In general, there are general rules for defining the dataset size. One of these rules of thumbs are, a factor of certain characteristics of the prediction problem. One rule-of-thumb is that the sample size needs to be at least a factor 50 to 1000 times the number of prediction classes. Another rule-of-thumb is that the sample size needs to be at least a factor 10 to 100 times the number of the features.[29] In the Table 7.1 there is a general recommendation on how to set up the initial size of dataset. This table is not a final decision; since the numbers are gained from years of practice and experience, they are but a helpful guide to start with.[30]

Expansion of data can be done using the following suggestions:
1- Acquire more training data by repeating the experiments. This option is the ideal solution but needs enough time and budget, which makes it less practical in many cases, such as this case.
2- Add some random noise to the existing dataset. Although, due to many studies, random noise is considered to negatively affect the quality and accuracy of prediction results, this only happens when the dataset is small or has less representative data. This situation gives the noises have great chances to be learned, and later act as a basis of predictions. [31]

Table 7.1: Suggested numbers of data samples for each type of project

Project example	Result quality example	Amount of data samples
Classical ML projects	Least possible quality	less than 1 000
Detection and identification tasks	Least possible quality	1 000 - 10 000
Non-complicated natural language processing	Average quality	10 000 - 50 000
High-quality classification	Average quality	50 000 - 100 000
Image processing for non-complicated tasks	Good quality	100 000 - 500 000
Very high-level semantic analysis	Good quality	500 000 - 2 000 000
Complicated tasks with unlimited possible data units	Excellent quality	2 000 000 and more

3- Generate new data that follows the pattern and behavior of the existing dataset.[27] This method of expanding the dataset, in a more scientific name, refers to synthetic data. Synthetic data means to use different methods to artificially generate data rather than using data generated by real events with the aim of testing systems or creating training data for machine learning algorithms.[32]
Synthetic data is a type of data augmentation that is generated with the help of algorithms and is used for various activities and testings, including as test data for new products and tools. Also it can is used for model validation. Synthetic data is in demand when: privacy requirements limits data availability or how it can be used. [30]
Three general strategies for generating synthetic data include:

- Generating according to distribution:
  When there is a comprehensive understanding of how dataset distribution would look like, the sampling can be done by using any distribution such as Normal, Exponential, Chi-square, t, log-normal and Uniform.[33],[30][27]

- "Agent-based modeling: To achieve synthetic data in this method, a model is created that explains an observed behavior, and then reproduces random data

using the same model. It emphasizes understanding the effects of interactions between agents on a system as a whole."[30]

- "Deep learning models: Variational autoencoder and generative adversarial network (GAN) models are synthetic data generation techniques that improve data utility by feeding models with more data. "[30]

- One easy way to generate dataset is to use different equations. These equations can be anything, linear, quadratic, polynomial, etc. However, this project aims not to find the best possible way to generate data, but a logical way to generate data. In this regard, the process of generating a simulated dataset can be explained as follows. First, specify the model to simulate. Next, determine each independent variable's coefficient, then simulate the independent variable and error that follows a probability distribution. And finally, calculate the dependent variable based on the simulated independent variable.[34]

As said earlier, the aim is to propose a logical way to generate data, not the optimal solution, as this project is not a data science work. Therefore, among all the recommended methods, deep learning cannot be implemented due to being out of scope of this project. Therefore, for generating input variables the first method, generating according to distribution, is chosen. When the distribution and parameters of each variable are estimated, each column of data will be re-sampled again to reach a decent number of observations. However, one hypothesis should be added for this part so that data generation works flawlessly. The hypothesis is that the variables are considered independent, meaning that increase or decrease of any of the variable does not affect the other variables.
To generate the output variable, the last method, using linear regression, is chosen. In the following sections, the regression equation and the coefficient should e calculated. Afterwards, the work is dedicated to presenting fitted distributions and estimation of their parameters.

7.2 CFU Prediction using the generated data

In the previous part, different independent values of temperature, humidity, and number of particles were generated using their distribution, the next step is to use these data to generate fake CFU levels. Therefore, it is essential to figure out how to use these generated values in the previous section to predict CFU. Although there is no evidence of presence of a mathematical equation that takes temperature,

humidity, and number of particles as inputs and gives back CFU level as an output; an other hypothesis has to be added which it is assumed that there is a function that takes temperature, humidity, and number of particles as inputs and returns CFU level as output. Although, there is no evidence on how inputs affect the output result, it can be said the problem has no specific solution.[35]
From Mathematical perspective, given an unique set of inputs and corresponding outcomes, there are infinite number of equations that produce exactly those results, even without allowing for the possibility of noise. Despite all the limitations in finding a a proper equation and a need of more data, it is decided to use linear regression for obtaining the equation. This decision is made by consulting with experts from computer science. However, having 15 variables in a linear regression by assuming that they do not interact with each other is problematic. It is stated that many challenges are possible when there are more than five independent variables in a regression equation. One of the most frequent is the problem that two or more of the independent variables are highly correlated to one another.[23] Therefore an other feature selection has to be done to reduce the number of variables. Going thorough these sections two decisions have been made so far:
1- Use an other feature selection to decrease the error and violating the hypothesis of variables being independent. By repeating the the steps in 6.2 ,this time the aim is to choose the first five important predictors. Therefore the following features are selected: Temperature of wound (deep and surface), humidity around the patient and around the left surgical table and number of 3 micron diameter particles in place 2. By looking at the initial values of humidity around the patient and surgical table, since these values are relatively closed to each other, it is decided that to eliminate one of these values as arbitrary, which is humidity around surgical table.
The last feature selection results in 4 final inputs. 2- To use the initial dataset to develop a linear regression which be later use to generate CFU level based on the newly generated variables in next section. The line code and the results of fitting a linear regression on the initial data based on the 4 important features are as followed:

model = fitlm(X,CFU);

Where X is a table that includes the initial values of the 4 selected variables, and CFU is the table containing of the corresponding CFU level.

	Estimate	SE	tStat	pValue
(Intercept)	282.24	102.41	2.7561	0.013495
x1	-4.0438	1.0083	-4.0107	0.00090604
x2	0.59251	0.8274	0.71611	0.48365
x3	-10.109	5.7221	-1.7667	0.095233
x4	0.12883	0.30919	0.41668	0.68212

Figure 7.1: Result of regression fitting

Where X1 indicates temperature in deep wound, X2 temperature in surface of the wound, X3 humidity, X4 number of 3 micron particles in place 2. By using the the coefficient above, and the random values of predictors from previous section, new fake CFU level can be obtained. Going through the made dataset, some negative values are observed. These negative values are all eliminated. As all the new CFU level are generated by a linear regression, if the dataset is given to any classifier, an over-fitting will be occurred, therefore,one of the methods for regularization introduced has to be implemented. For this project it has been decided to add noise to dataset to overcome overfitting problem.Noise injection is discussed in section 7.4.

7.3 Distribution fitting

For fitting distribution, distribution fitter tool from MATLAB was used.[36] Distribution fitter app allows to import the data, and it plots the probability density function (PDF), cumulative density funtion (CDF), Quantile, Probability plot, survivor function, and cumulative hazard. For this work, PDF plot is needed. In first step, each variable was imported. Using distribution fitter tool from MATLAB, different distributions will be tried to be fitted on density plot to decide which estimation is the closest to the optimal answer. However, for some variables, no distribution can desirably be fitted on density plot. For these variables kernel distribution should be used. [37] The parameters of kernel distribution are not shown in the distribution fitter toolbox, but are shown using the below codes in the command window of MATLAB. The variable "Temperature Deep Wound" was among those variables with no good fit from MATLAB distribution fitter.

pd = fitdist(Temperature Deep Wound,'Kernel');
mean = mean(pd);
standard deviation = std(pd);

After finding the related distribution, more data was generated for each variable. However, after observing the initial data, it was recognized that some variables had out-ranged data which influenced parameters in the distributions. This is due to the fact that , for example for normal distribution, the numbers can be generated in the range (mu-5*sigma, mu+5*sigma); though happens nor very rarely, sometimes values close to the beginning and end of the interval can be generated. Therefore, making the new variables unrelated to the work. For example, temperature around the animal started with values around 17 degrees, and increased to around 30 degrees after 30 minutes. Since, the source of hitting around the animal is surgical lamps, and during the surgery the lamps will stay on, it is obvious that the temperature around the animal will not be back lower that 30, so the values so much smaller than 30 were removed, as they influence the variance of data strongly. After removal of the outranged data, fitting process was performed once again. In Table 7.2 all the variables and their related distributions are shown and discussed. In Figure 7.2, there are plots of the PDF of body temperature and temperature in deep wound, and the suggested distribution that best describes these variables.

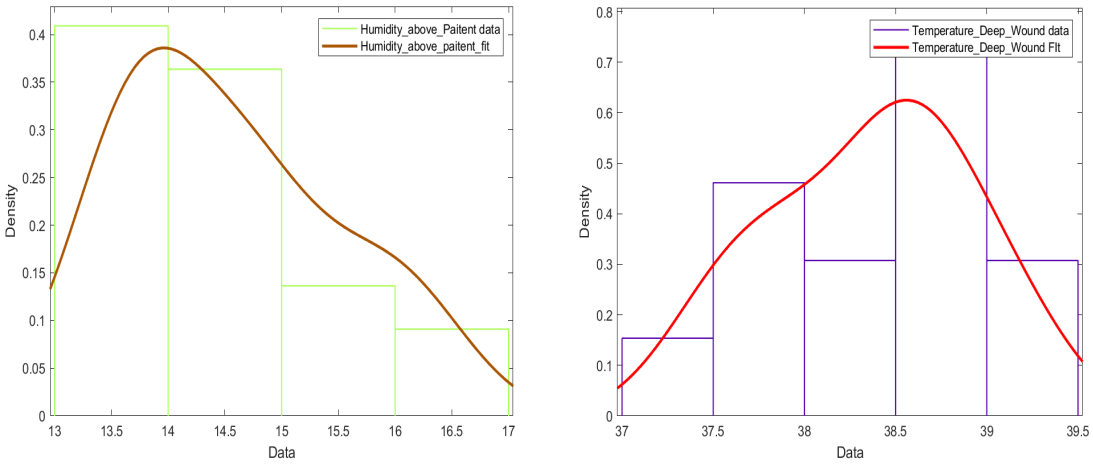
Table 7.2: Fitted distribution of variables

Name of variable	Name of fitted distribution	Estimated parameter
Temperature of deep wound	Kernel	mu= 38.3604 sigma= 0.6181
Temperature of surface of the wound	Normal	mu= 37.45 sigma= 1.46
Humidity above the patient	Kernel	mu=14.5453 sigma=1.0541
Number of 3 micron particles in place 2	Exponential	mu=14.63

7.3.1 Input generation

The next step is to generate input variables based on the fitted distribution for them. The variables follow three distributions: normal, exponential, and kernel. Variables that follow normal and exponential distribution can be easily generated in Excel using the following commands:

- Normal distribution:
  In excel: X= NORM.INV(Probability, mean, standard_dev).



(a) Distribution fitted for humidity above patient

(b) Fitted distribution for temperature in deep wound using kernel function

Figure 7.2: Fitted distribution for two variables

For variable "Body temperature" this code line can be extended to:
NORM.INV(ran(), 30.13, 1.26)
In MATLAB: r = normrnd(mu,sigma,sz)
generates an array of normal random numbers, where vector sz specifies size(r).[38]

- Exponential distribution: [39]

  X= ln(1-rand())/-lambda
  For variable "Number of 3 micron particles in place 2" this code line can be extended to: ln(1-rand())/-14.63
  Or in MATLAB
  X = [number of required observations 1];
  r3 = exprnd(14.63,X)

However, for variables following kernel distribution, it is easier to use MATLAB for data generation. The command line to do so for generating 100 samples, is as followed:[40]
pd = fitdist(X,'kernel');
Y = random(pd, [100,1]); or Y = pd.random(100,1);
When generating data, sometimes negative values are appeared. This happens always for variables that shows considerable deviations in plots in chapter 5. Considering estimated parameters at table, it is also excepted to see negative or out ranged values for some variables tat have big standard deviation compared to their mean values, by having the shape of normal distribution in mind and by considering mean at the middle, if two/three times of standard deviation is close or bigger than mean, there is a chance that generated data will be negative or out-ranged. For example, for variables related to temperature and humidity no negative or out ranged values

will be generated as the standard deviation is rather small compared to the mean value. The generation of undesirable values happens for number of particles in different places. Therefore, after generation the new dataset, it has to be investigated completely in order to remove the negative and out ranged values.

7.4 Noise injection

Several empirical studies state that noise plays a vital role in the effective and efficient training of neural networks. The theory behind it, however, is still largely unknown.[41] In machine learning tasks, it helps to reduce overfitting, while in data privacy protection, it adds uncertainty to personally identifiable information. [42] The addition of noise can be done in several ways. For example, noise in the input data of a neural network during training can, in some cases, lead to significant improvements in generalization performance.[43] Adding noise leads to smaller network weights and a more robust network with lower generalization error. This is because the network is less able to memorize training samples since they are changing all of the time. [44]

Noise generation can be categorized by three main features:[45]
1- The place where the noise is injected. Noise may affect input predictors or output class, impairing the learning process and the resulting model.
2-Distribution of noise. Noise in the dataset can be presented as uniform or Gaussian.
3-The focus and immensity of noise values. The number of noisy data depends on each data value of each attribute or relative to the minimum, maximum and standard deviation for each attribute.
The most typical type of noise used during training is the addition of Gaussian noise to input variables. The method of noise injection refers to adding noise artificially to input data during the training process. Jitter is one particular method of implementing noise injection. With this method, a noise vector is added to each training case in between training iterations. This causes the training data to "jitter" during training, making it difficult for the algorithm to find a solution that fits the original training dataset exactly, resulting in reducing overfitting. The noise vector is typically drawn from some probability density function known as a kernel. For this case, a one-mean and one-standard deviation Gaussian kernel was used.[26]
In the following, there is a description of how noisy data having Gaussian or normal distribution is made.
Uniform attribute noise. x% of the values of each attribute in the dataset are cor-

rupted. To corrupt each attribute Ai, x% of the examples in the data set are chosen, and their Ai value is assigned a random value from the domain Di of the attribute Ai. An uniform distribution is used either for numerical or nominal attributes.
Gaussian attribute noise is similar to the uniform attribute noise, but in this case, the Ai values are corrupted, adding a random value to them following a Gaussian distribution of mean = 0 and standard deviation = (max-min)/5, being max and min the limits of the attribute domain. Nominal attributes are treated as in the case of the uniform attribute noise.
In this project, the gaussian method is chosen to add noisy data for each input. However, in practice, because the effect of a particular variance value of the noise kernel with respect to the underlying class distribution of the training cases are not known a priori or not known at all, one must select an appropriate value for the variance of the noise kernel-from analyzing the training data. One of the methods to do so is described by Holmstrom and Koistinen to identify an appropriate value for the variance of the noise kernel.[46]
An other topic which requires discussion is the amount of added noise. Too little noise has no effect, on the contrary too much noise makes the mapping function too complicated to learn.[45]In real-world datasets, the initial amount and type of noise are unknown. Therefore, no assumptions about the noise type and number of noisy data can be made.[45] Therefore,adding noise will continue until around 5% of total observations will be consisted of dangerous level of CFU.
In the first attempt is to add noise, 50 noisy rows generated. This 50 could not reach the desired numbers in a dangerous level of CFU, so continuous attempts were made to reach the desired level. Data generation was stopped by reaching 508 observations. 39 observations were categorized as dangerous level. These 39 observations were around 5.5% of the whole dataset. The next step is to import these data into the classifier app in MATLAB and analyze them.

7.5 Summery

In this chapter, methods to expand the dataset were discussed. Based on the constraints and limitations of the study and the available equipment, the chosen method is to use the distribution of parameters to generate fake values for inputs. The next step is to again reduce the number of predictors, as the aim is to minimize the possibility of interaction between variables. Based on different studies, five input values are considered the maximum number of inputs that can be put in an equation with minimum multiplication effect. Therefore, feature selection was performed to determine those five variables. Among five variables, two of them were in different locations in the operating room. Since their values were rather close to each other, with very low variance, it was decided to keep the humidity level around the patient's bed. Implementing these steps results in reaching the number of inputs to four. The next step in data generation is to achieve the CFU level, as response values. The chosen method for generating CFU level is linear regression, as it is the simplest method that fits well for 22 observations. However, the linear regression fits the dataset so well that if the generated data is given to classifier recognition tools, overfitting will accrue. Adding noise to generated CFU levels will solve the overfitting problem. For noise to perform well, the number of CFU levels above 35 should consist of around 5% of the whole observation. Therefore, noise injection and generating CFU levels were done in parallel. After preparing the acceptable dataset, the final step is to import response and predictor values to classifier prediction. In the next chapter, different algorithms are trained for the dataset.

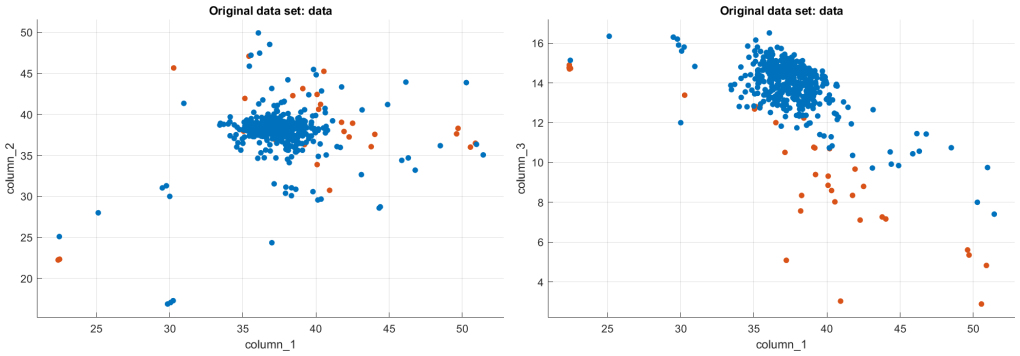
Chapter 8

Use of machine learning in risk analysis

This chapter discusses how to implement the generated dataset into classifier app in MATLAB and shows the results. Also, there will be a discussion about how practices within data analysis contributes to risk analysis. The initial objective of this work is to set up a data-driven method for forecasting the CFU level in operating rooms using data.

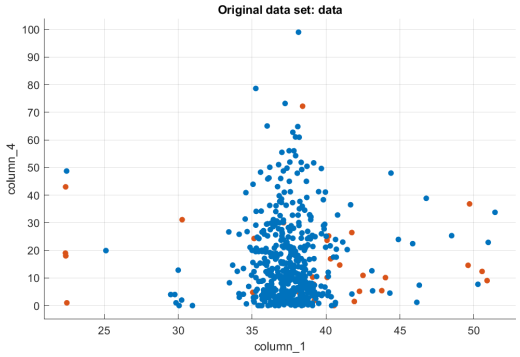
8.1 Implementing machine learning on dataset

Before inserting dataset into the classier app, a small modification should be done. Since the data is going to be classified, CFU level cannot be used as response. Instead, the CFU level being dangerous or not is a response. Therefore, it is decided that for CFU level above the value of 35, 1 will be assigned, indicating the CFU level being dangerous and for CFU level below 35 0 is assigned.[22] In Figure 8.1, scatter plots of different predictors against each other are shown. Blue and orange dots represent class 0 and class 1 respectively. Column 1 refers to body temperature in deep wound, Column 2 refers to body temperature in surface wound, Column 3 refers to humidity, and column 4 is the number of particles in place 3. In Figure 8.1.a the majority of class 0 can be found when temperature in deep wound and surface wound both varies between 35 and 40 degrees. In Figure 8.1.b dispersion is more obvious. However, orange dots, class 1, are more in the ranges where humidity is low.



(a) Temperature in deep wound vs temperature of surface wound

(b) Temperature in deep wound vs humidity



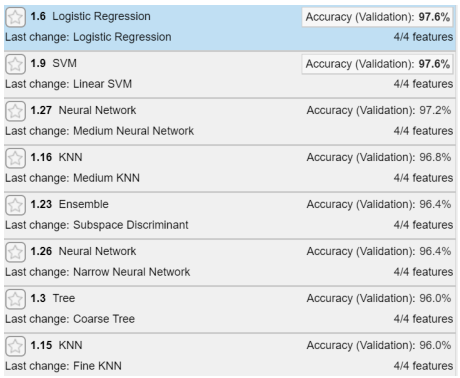
(c) Temperature in deep wound vs number of 3 micron diameter particles in place 2

Figure 8.1: Scatter plots, temperature in deep wound, surface wound, number of particles and humidity

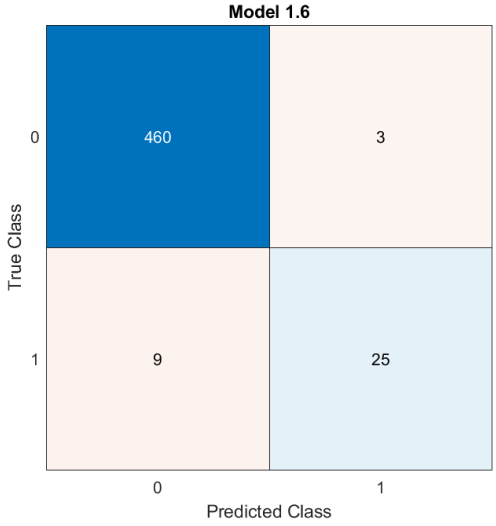
In Figure 8.1.c , there is also a huge dispersion in class 1, however the aggregation of class 0 is found when the temperature in deep wound varies between 35 and 40 degrees, and while the number of particles are between 0 and 25. There is a huge chance that these plots does not represent the real world events, but in the case of having real data, it should be interpreted in such way.

8.2 Classification results

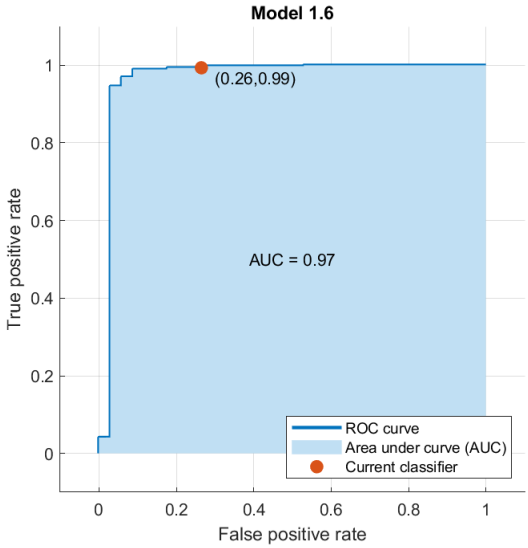
After importing the dataset and identifying the response and predictors, to protect from over fitting, cross validation is also done, choosing the class validation folds as 5. When importing the data into classifier app, all the available classification algorithms can be run like: decision tree, logistic regression, support vector machine, neural network. It was decided to run all the all the algorithms and then rank them based on the accuracy of the validation set. The result of one of the runs is as follows in Figure 8.2:



(a) Algorithms ranking based on accuracy



(b) Confusion matrix of logistic regression



(c) ROC and AUC of logistic regression

Figure 8.2: A sample of results from the classifier app

MATLAB offers to export the classifier, so it can be used later, when new data

are available, or used by an other person.[47] To grasp a better understating of the performance of a classifier, other tools or criteria can be used. In the following, these tools and criteria will be discussed.

8.3 Interpretation of results

When deciding which model offers the best prediction, the answer can be complicated. For example, if the accuracy is 99.9%, it is assumed that this predictor is the best tool for prediction. However, when it comes to deciding the suitability of predictors, the type of outcome for prediction should be highlighted. For example, if the positive case is someone who is sick and carrying a virus that can spread quickly, or a model separates a terrorist from a non-terrorist. In these cases, the cost of having a misclassified actual positive (or false negative) is very high and can have severe consequences. Furthermore, high accuracy means the number of correct predictions is high. One weakness of accuracy is that when a dataset contains a few numbers of one class, the result of accuracy is biased and cannot be trusted. Accuracy is a good measure when an equal number of observations is in the dataset. Even if accuracy is going to be used in imbalanced data, the costs of false positives should be low to minimise the loss. Therefore, the metrics introduced in chapter 4 are useful to give a better vision when working in the dataset where the number of one class is noticeably few.
When working with imbalanced data, other metrics are more useful. One of the most popular ones is F1, which consists of precision and recall. Recall and precision can also be used as judging tools individually. A model with high precision is the one that maybe cannot find all the positives, but the ones that are classified as positive are very likely to be correct. Recall actually calculates how many of the actual positives the model detects through labelling it as positive (true positive). It means that recall brings a more safety to prediction, as high recall means most likely all the positive cases are labelled as positive, but maybe some negative is among them. Recall is the metric to select the best model when there is high cost is associated with not finding positives. For instance, in fraud or sick patient detection. If a fraudulent transaction (actual positive) is predicted as non-fraudulent (predicted negative), the consequence can be adverse for the bank. Likewise, in sick patient detection. If a sick patient (actual positive) is tested and predicted as healthy (predicted negative), the cost associated with false negative will be high if the disease is contagious. High F1 indicates precision and recall are high (more than 0.7) , average F1 refers to both precision and recall are low (0.4-0.7) , and low F1 means that one of precision and recall is low and the other is high (0-0.4).
In Table 8.2, all these 4 metrics are calculated. For all the models, accuracy is very

high, indicating a very good classification of infection and non-infection. For deciding which algorithm perform the best accuracy cannot be used since the accuracy values are close. Also, recall and precision have scored acceptable values. Table 8.2 shows that recall and precision values are apart except for the Tree algorithm.
To know whether to use precision or recall, let's assume that the goal is to find infection in a very high-risk patient, then a model with high recall is needed. It leads to choosing logistic regression or neural network. However, an ideal model has high values both in recall and precision. This leads to choosing logistic regression, with 89% precision.
, However, if only F1 measure is the basis of judgment, there should be a choice between LR and linear SVM. Both have F1 81%, so which one is better? In this case, the recommendation is to remind again what we are looking for and what is the cost of a false detection. A false detection results in mental, physical and economic damage to a person; the model with high recall should be chosen since the damage is serious. Again, LR is chosen as the best model.

Table 8.1: Summary of confusion matrix for each method

	TP	FP	FN	TN
LR	25	3	9	460
Linear SVM	25	2	10	461
Neural network	25	5	9	458
KNN	19	1	15	462
Tree	21	12	13	451

Table 8.2: Accuracy, precision and recall, and F1

	Accuracy	Recall	Precision	F1
LR	98%	74%	89%	81%
Linear SVM	98%	71%	93%	81%
Neural network	97%	74%	83%	78%
KNN	97%	56%	95%	70%
Tree	95%	62%	64%	63%

Regarding the descriptions about the evaluation criteria, if the objective of the prediction is to highlight all surgeries susceptible to SSI for further screening by

healthcare personals, then a high recall would be desirable at the expense of a lower precision, because the healthcare personals can then manually remove the false positives. However, if the objective is to reject surgeries with high SSI, to avoid wasting hospitalization after the surgery, then a high precision (lower recall rate) might be more suitable. The other factor important to know is that what type of surgery the data comes from. For example, different surgeries can pose different threats of SSI due to their nature. In head and neck surgeries SSI is much more common that abdominal surgeries. Moreover, the health state of a patient has a vital effect on which factor should be used to choosing the model. In high risk patient, like with diabetes, high BMI, or other medical difficulties, the threat of SSI can be different.

8.4 Summery

This chapter investigated the generated data using the classifier app in MATLAB. The classifier app offers different features for modelling and assessing the model. However, the machine learning method acts as a black box, so different aspects should be considered when choosing the model. One important question to help start this investigation is to answer, "what are we looking for when looking at the result?". For example, is the main objective to find how many surgeries have been correctly labelled out of all the surgeries? Then accuracy is needed. If the question is how many surgeries labelled as having dangerous risk of SSI are actually dangerous, precision is needed. If the question is, among all the people surgeries labelled as SSI positive, how many of those were correctly predicted, recall is needed.

Chapter 9

Conclusion

This chapter provides the results and conclusion of the work. It summarizes the aim of the work, limitations and challenges, and the techniques used in the work. It also suggest methods to improve the work.

9.1 Results

The main objective of this study is to answer the question of can a predictive model be used as a dynamic model to make decisions regarding high infection in a surgery? The answer to this question depends on how far the model is expected to answer accurately. Choosing the optimal algorithm is challenging, and it is still done manually. Therefore, different measures for evaluating machine learning algorithms are proposed. Accuracy is a simple term which means the classification is done correctly. Accuracy of 90% means that 90% of data is in the correct groups. However, there is one problem with accuracy. Accuracy lacks efficiency when it comes to suggesting a good model when more data points of one class are observed than of another. Using accuracy to determine infection cannot be a good answer for all surgeries in this work. In head, neck, and orthopedic surgeries, accuracy can offer a good model since the rate of infection is higher. But for other surgeries where infection happens for fewer patients, accuracy is not efficient. To solve this issue, other metrics should be used.
F1, which is composed of precision and recall, is a popular metric. However, precision and recall can be used separately. For this study, high precision means maybe all the positives could not be found, but the ones that the model classes as positive are likely correct. On the other hand, high recall means all infections have been classified correctly, but some non-infection are also classified as infection. The ideal model is the one that classifies all the infections as infections correctly, which means

a model with high recall and precision. Instead of handling the challenge of finding a model both high in recall and precision, a model should be founded where the mean of precision and recall is high.
Finding a prediction model is a challenging task. However, when analyzing the results, the first step is to know about the nature of the problem, what is expected to be discovered, what are the expenses of a missed classification.

9.2 Recommendations for future work

The result of data related works is related both to the data analysis process and on the quality and quantity of data. In this work, the initial data lacks both quantity and quality. The limitations in time and budget to perform more surgeries were the main reasons leading to low quality and quantity of data. By working with data that is not acceptable in quality or quantity, it is complicated to draw any solid conclusions. Although data expansion was done, so the data can be reached to an acceptable quantity, but the quality cannot be trusted completely. Therefore, the result of the thesis revolves mainly around the concept that if proper and trustworthy data was available, what could have been done.

In order to improve the results of this thesis, the plan for data gathering should be revised in many ways. Firstly, gathering input variables is not challenging since it is done using equipment that automatically saves and logs data. The main problem is collecting output variable since it is done manually by changing the blood adgers regularly. Secondly, one of the problems in this thesis was the limited number of observations. Therefore, more observations will be provided by increasing the number of experiments, however the problem of budget is still a problem. Thirdly, it is suggested, if possible, to include some data from real surgeries. Finally, it is claimed surgical site infection can be due to factors that are completely different in nature. It is stated that four groups of factors are related to infection: room factor, patient factor, HVAC factor, and surgical factors.[48] For studying the risk of infection, all these factors have to be studied. However, whether studying one type of factor or all factors, this project is an interdisciplinary project. It is recommended that the next work should be written with regular consulting with a healthcare specialist and an energy specialist.

Bibliography

1. Cao G, Pedersen C, Zhang Y, Drangsholt F and Radtke A. Can clothing systems and human activities in operating rooms with mixing ventilation systems help achieve 10 CFU/m3 level during orthopaedic surgeries? 2021; 120:110–6. DOI: 10.1016/j.jhin.2021.11.005
2. Disease Prevention EC for and Control. Surgical site infections - Annual Epidemiological Report 2016 [2014 data]. 2016
3. Owens C and Stoessel K. Surgical site infections: epidemiology, microbiology and prevention. Journal of Hospital Infection 2008; 70:3–10. DOI: 10.1016/S0195-6701(08)60017-1
4. Persson M. Airborne contamination and surgical site infection: Could a thirty-year-old idea help solve the problem? Med Hypotheses 2019; 132. DOI: 10.1016/j.mehy.2019.109351
5. Sajadi B.and Saidi M and Ahmadi G. Computer modeling of the operating room ventilation performance in connection with surgical site infection. Scientia Iranica 2020; 27:704–14. DOI: 10.24200/SCI.2018.5514.1359
6. Hopkins B, Mazmuda A and Driscoll C. Using artificial intelligence (AI) to predict postoperative surgical siteinfection: A retrospective cohort of 4046 posterior spinal fusions. Clinical Neurology and Neurosurgery 2020; 192. DOI: 10.1016/j.clineuro.2020.105718
7. Kuo PJ, Wu SC, Chien PC and Chang SS. Artificial neural network approach to predict surgical site infection after free-flap reconstruction in patients receiving surgery for head and neck cancer. Oncotarget 2018; 9:1115–22. DOI: 10.18632/oncotarget.24468
8. Alfosno-sanchez J, Martinez I and Martín-Moreno J. Analyzing the risk factors influencing surgical site infections: the site of environmental factors. Canadian journal of surgery. Journal canadien de chirurgie 2017; 60(3):155–61. DOI: 10.1503/cjs.017916

9. Noguchi C, Koseki H, Horiuchi H and Yonekura A. Factors contributing to airborne particle dispersal in the operating room. BMC Surgery volume 2017; 17(78):454–70. DOI: 10.1186/s12893-017-0275-1

10. Spin Air – Air Sampler. Available from: https://iul-instruments.com/product/spin-air-air-sampler/ [Accessed on: 2022 Jun 5]

11. 4 Benefits of Data Analytics in Healthcare. Available from: https://online.maryville.edu/blog/data-analytics-in-healthcare/ [Accessed on: 2022 Jun 1]

12. data analysis-process. Available from: https://www.tutorialspoint.com/excel_data_analysis/data_analysis_process.htm [Accessed on: 2022 Apr 13]

13. , Rachida D and Mohamed Adel S. Big Data Pre-Processing: A Quality Framework. IEEE International Congress on Big Data 2015 :191–8. DOI: 10.1109/BigDataCongress.2015.35

14. What Is Machine Learning? 3 things you need to know. Available from: https://se.mathworks.com/discovery/machine-learning.html [Accessed on: 2022 Apr 5]

15. Classification. Available from: https://se.mathworks.com/help/stats/classification.html?s_tid=CRUX_lftnav [Accessed on: 2022 Apr 27]

16. Classification Learner App. Available from: https://se.mathworks.com/help/stats/classification-learner-app.html [Accessed on: 2022 Apr 5]

17. Chow JC. Analysis of financial credit risk using machine learning. arXiv preprint arXiv:1802.05326 2018. DOI: 10.13140/RG.2.2.30242.534492

18. Paltrinieri N, Comfort L and Reniers G. Learning about risk: Machine learning for risk assessment. Safety Science 2019; 118:475–86. DOI: 10.1016/j.ssci.2019.06.001

19. Davis J and Goadrich M. The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd international conference on Machine learning 2006 :233–40

20. Li Q. Overview of Data Visualization. Embodying data 2020 :17–47. DOI: 10.1007/978-981-15-5069-0_2

21. Balaras C, Dascalaki E and Gaglia A. HVAC and indoor thermal conditions in hospital operating rooms. Energy and buildings 2007; 39(4):454–70. DOI: 10.1016/j.enbuild.2006.09.004

22. Napoli C, Marcotrigiano V and Montagna M. Air sampling procedures to evaluate microbial contamination: a comparison between active and passive methods in operating theatres. BMC Public Health 2012; 12. DOI: 10.1186/1471-2458-12-594

23. MULTIPLE REGRESSION. Available from: https://home.csulb.edu/~msaintg/ppa696/696regmx.htm [Accessed on: 2022 May 2]

24. Introduction to Feature Selection. Available from: https://se.mathworks.com/help/stats/feature-selection.html [Accessed on: 2022 May 2]

25. Overfitting, more than an issue. Available from: https://towardsdatascience.com/overfitting-more-than-an-issue-fac2d8b1fb5d [Accessed on: 2022 Jun 1]

26. Zur R, Jiang Y, Pesce L and Drukker K. Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. Medical physics 2009; 26(10):4810–8. DOI: 10.1118/1.3213517

27. Ying X. An Overview of Overfitting and its Solutions. Journal of Physics 2019; 1168(2). DOI: 10.1088/1742-6596/1168/2/022022

28. Alwosheel A, Cranenburgh S van and Chorus CG. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. Journal of Choice Modelling 2018; 28:167–82. DOI: 10.1016/j.jocm.2018.07.00

29. Figueroa RL, Zeng-Treitler Q, Kandula S and Ngo LH. Predicting sample size required for classification performance. BMC medical informatics and decision making 2012; 12(1):1–10

30. The Ultimate Guide to Synthetic Data: Uses, Benefits Tools. Available from: https://research.aimultiple.com/synthetic-data/ [Accessed on: 2022 Apr 27]

31. Gupta S and Gupta A. Dealing with Noise Problem in Machine Learning Datasets: A Systematic Review. Procedia Computer Science 2019; 161:466–74. DOI: 10.1016/j.procs.2019.11.146

32. Bolon-Canedo V, Sanchez-Marono N and Alonso-Betanzos A. A review of feature selection methods on synthetic data. Knowledge and information systems 2013; 34(3):483–519. DOI: 10.1007/s10115-012-0487-8

33. Synthetic Data Generation: Techniques, Best Practices Tools. Available from: https://research.aimultiple.com/synthetic-data-generation/ [Accessed on: 2022 Apr 19]

34. Generate Simulated Dataset for Linear Model in R. Available from: https://towardsdatascience.com/generate-simulated-dataset-for-linear-model-in-r-469a5e2f4c2e [Accessed on: 2022 Mar 21]

35. Hi, I have experimental data and I need to fit them. The fitting function is unknown and I would like to solve the problem avoiding the use of the fitting tool. Can someone help me? Available from: https://se.mathworks.com/matlabcentral/answers/331651-hi-i-have-experimental-data-and-i-need-to-fit-them-the-fitting-function-is-unknown-and-i-would-lik [Accessed on: 2022 Apr 27]

36. Fit a Distribution Using the Distribution Fitter App. Available from: https://se.mathworks.com/help/stats/fit-a-distribution-using-the-distribution-fitting-app.html [Accessed on: 2022 Apr 27]

37. Fit Kernel Distribution Object to Data. Available from: https://se.mathworks.com/help/stats/fit-a-kernel-distribution-object-to-data.html [Accessed on: 2022 Apr 27]

38. normrnd. Available from: https://se.mathworks.com/help/stats/normrnd.html#d123e677941 [Accessed on: 2022 May 5]

39. Simulation - 3 - Generate Exponentially Distributed Random Numbers. Available from: https://www.youtube.com/watch?v=1V7_OO9hBsY [Accessed on: 2022 Apr 28]

40. Generating random data from Kernel density estimator. Available from: https://se.mathworks.com/matlabcentral/answers/20402-generating-random-data-from-kernel-density-estimator [Accessed on: 2022 May 1]

41. Zhou M, Liu T, Li Y, Lin D, Zhou E and Zhao T. Towards Understanding the Importance of Noise in Training Neural Networks. International Conference on Machine Learning 2019 :7594–602

42. Rodriguez-Garcia M, Batet M and Sanchez D. A semantic framework for noise addition with nominal data. Knowledge-Based Systems 2017; 122:103–18. DOI: 10.1016/j.knosys.2017.01.032

43. Bishop RM. Training with Noise is Equivalent to Tikhonov Regularization. Neural Comput 1995; 7(1):108–16. DOI: 10.1162/neco.1995.7.1.108

44. Train Neural Networks With Noise to Reduce Overfitting. Available from: https://machinelearningmastery.com/train-neural-networks-with-noise-to-reduce-overfitting/. [Accessed on: 2022 May 2]

45. Noisy Data in Data Mining. Available from: https://sci2s.ugr.es/noisydata#Creating [Accessed on: 2022 May 5]

46. Holmstrom L and Koistinen P. Using additive noise in back-propagation training. IEEE transactions on neural networks 1992; 3(1):24–38

47. Machine Learning Applications in Risk Management: Classifying Credit Card Default Using the Classification Learner App. Available from: https://se.mathworks.com/videos/machine-learning-applications-in-risk-management-classifying-credit-card-default-using-the-classification-learner-app-1536227997787.html [Accessed on: 2022 May 16]

48. Memarzadeh F. Comparison of Operating Room Ventilation Systems in the Protection of the Surgical Site. ASHRAE Transactions 2004

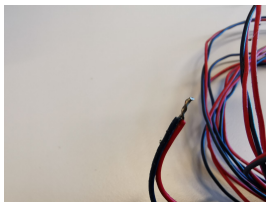
Appendix

A Pictures of equipment used in the experiment

(a) Aerotrak 9306

(b) Tiny Tag

(c) Thermal camera



(d) Thermal Couple

(e) Hioki LR 8400

(f) Flir A60

(g) Air sampler

Figure 1: Used equipment in the experiment

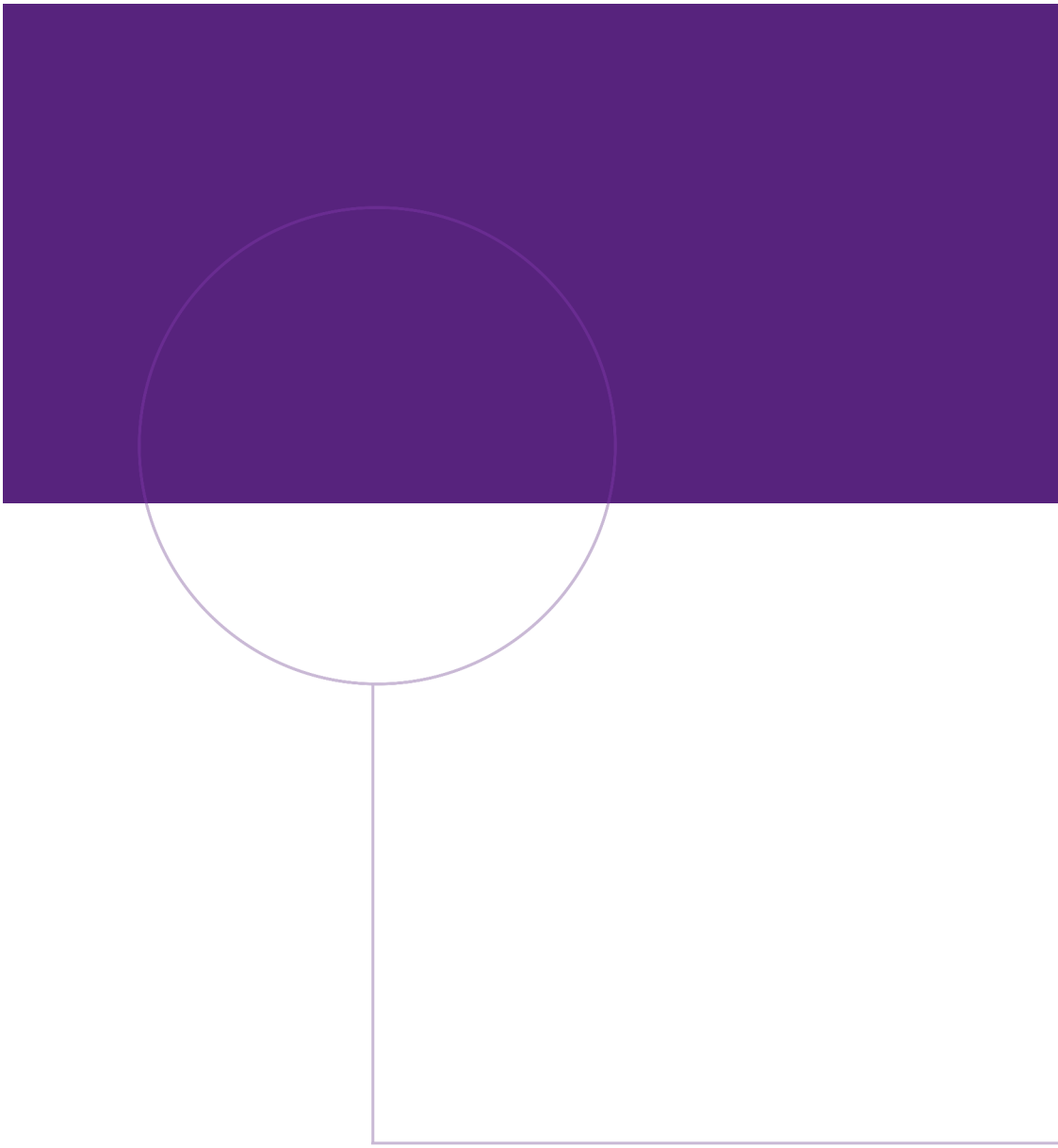
B MATLAB code for extracting temperature data from thermal pictures

```
% Get the name of the image the user wants to use.
baseFileName = 'IR_5003.jpg'; % Base file name with no folder
↪  prepended (yet).
\thispagestyle{empty}
% Get the full filename, with path prepended.
folder = pwd; % Change to whatever folder the image lives in.
fullFileName = fullfile(folder, baseFileName);
% Append base filename to folder to get the full file name.
fprintf('Transforming image "%s" to a thermal image.\n',
↪  fullFileName);

%================================================================================
% Read in a demo image.
originalRGBImage = imread(fullFileName);
% Display the image.
subplot(2, 3, 1);
imshow(originalRGBImage, []);
axis on;
caption = sprintf('Original Pseudocolor Image, %s', baseFileName);
title(caption, 'FontSize', fontSize, 'Interpreter', 'None');
xlabel('Column', 'FontSize', fontSize, 'Interpreter', 'None');
ylabel('Row', 'FontSize', fontSize, 'Interpreter', 'None');
drawnow;

grayImage = min(originalRGBImage, [], 3);
% Useful for finding image and color map regions of image.

%================================================================================
% Need to crop out the image and the color bar separately.
% First crop out the image.
%imageRow1 = 70.5;
%imageRow2 = 154 ;
%imageCol1 = 96.5;
%imageCol2 = 66;
% Crop off the surrounding clutter to get the RGB image.
%rgbImage = originalRGBImage(imageRow1 : imageRow2, imageCol1 :
↪  imageCol2, :);
```

```
rgbImage= imcrop(originalRGBImage, [95.5 100.5 117 63]);

% Next, crop out the colorbar.

% Crop off the surrounding clutter to get the colorbar.
%colorBarImage = originalRGBImage(colorBarRow1 : colorBarRow2,
↪  colorBarCol1 : colorBarCol2, :);
%b = colorBarImage(:,:,3);
colorBarImage= imcrop(originalRGBImage, [306.5 67.5 9 106]);

%================================================================================
% Display the pseudocolored RGB image.
subplot(2, 3, 2);
imshow(rgbImage, []);
axis on;
caption = sprintf('Cropped Pseudocolor Image');
title(caption, 'FontSize', fontSize, 'Interpreter', 'None');
xlabel('Column', 'FontSize', fontSize, 'Interpreter', 'None');
ylabel('Row', 'FontSize', fontSize, 'Interpreter', 'None');
drawnow;
hp = impixelinfo();

% Display the colorbar image.
subplot(2, 3, 3);
imshow(colorBarImage, []);
axis on;
caption = sprintf('Cropped Colorbar Image');
title(caption, 'FontSize', fontSize, 'Interpreter', 'None');
xlabel('Column', 'FontSize', fontSize, 'Interpreter', 'None');
ylabel('Row', 'FontSize', fontSize, 'Interpreter', 'None');
drawnow;

% Set up figure properties:
% Enlarge figure to full screen.
set(gcf, 'Units', 'Normalized', 'OuterPosition', [0 0 1 1]);
% Get rid of tool bar and pulldown menus that are along top of
↪  figure.
% set(gcf, 'Toolbar', 'none', 'Menu', 'none');
% Give a name to the title bar.
set(gcf, 'Name', 'Demo by ImageAnalyst', 'NumberTitle', 'Off')
```

```
%=========================================================================
% Get the color map from the color bar image.
storedColorMap = colorBarImage(:,1,:);
% Need to call squeeze to get it from a 3D matrix to a 2-D matrix.
% Also need to divide by 255 since colormap values must be between
↪  0 and 1.
storedColorMap = double(squeeze(storedColorMap)) / 300;
% Need to flip up/down because the low rows are the high
↪  temperatures, not the low temperatures.
storedColorMap = flipud(storedColorMap);

% Convert the subject/sample from a pseudocolored RGB image to a
↪  grayscale, indexed image.
indexedImage = rgb2ind(rgbImage, storedColorMap);
% Display the indexed image.
subplot(2, 3, 4);
imshow(indexedImage, []);
axis on;
caption = sprintf('Indexed Image (Gray Scale Thermal Image)');
title(caption, 'FontSize', fontSize, 'Interpreter', 'None');
xlabel('Column', 'FontSize', fontSize, 'Interpreter', 'None');
ylabel('Row', 'FontSize', fontSize, 'Interpreter', 'None');
drawnow;

%=========================================================================
% Now we need to define the temperatures at the end of the colored
↪  temperature scale.
% You can read these off of the image, since we can't figure them
↪  out without doing OCR on the image.
% Define the temperature at the top end of the scale.
% This will probably be the high temperature.
highTemp = 36;
% Define the temperature at the dark end of the scale
% This will probably be the low temperature.
lowTemp = 24;

% Scale the indexed gray scale image so that it's actual
↪  temperatures in degrees C instead of in gray scale indexes.
```

```
thermalImage = lowTemp + (highTemp - lowTemp) *
↪  mat2gray(indexedImage);

% Display the thermal image.
subplot(2, 3, 5);
imshow(thermalImage, []);
axis on;
colorbar;
title('Floating Point Thermal (Temperature) Image', 'FontSize',
↪  fontSize, 'Interpreter', 'None');
xlabel('Column', 'FontSize', fontSize, 'Interpreter', 'None');
ylabel('Row', 'FontSize', fontSize, 'Interpreter', 'None');

% Let user mouse around and see temperatures on the GUI under the
↪  temperature image.
hp = impixelinfo();
hp.Units = 'normalized';
hp.Position = [0.45, 0.03, 0.25, 0.05];

%================================================================================
% Get and display the histogram of the thermal image.
subplot(2, 3, 6);
histogram(thermalImage, 'Normalization', 'probability');
axis on;
grid on;
caption = sprintf('Histogram of Thermal Image');
title(caption, 'FontSize', fontSize, 'Interpreter', 'None');
xlabel('Temperature', 'FontSize', fontSize, 'Interpreter', 'None');
ylabel('Frequency', 'FontSize', fontSize, 'Interpreter', 'None');

% Get the maximum temperature.
meanTemperature = mean(thermalImage(:));
fprintf('The mean temperature in the image is %.2f\n',
↪  meanTemperature);

⋮
```



NTNU

Norwegian University of
Science and Technology