Informatics

Predicting and Communicating Outcome of COVID-19 Hospitalizations with Medical Images and Clinical Data

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Data Science

eingereicht von

Oliver Stritzel, BSc
Matrikelnummer 11920598

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Assistant Prof. Dr. Renata Raidou

Wien, 15. August 2022

Oliver Stritzel

Renata Raidou

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU WIEN Informatics

Predicting and Communicating Outcome of COVID-19 Hospitalizations with Medical Images and Clinical Data

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Oliver Stritzel, BSc
Registration Number 11920598

to the Faculty of Informatics

at the TU Wien

Advisor: Assistant Prof. Dr. Renata Raidou

Vienna, 15th August, 2022

Oliver Stritzel   Renata Raidou

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Erklärung zur Verfassung der Arbeit

Oliver Stritzel, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 15. August 2022

Oliver Stritzel

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU WIEN Bibliothek Your knowledge hub

Danksagung

Das Wichtigste zuerst. Dieses Cover wurde unter Verwendung von Ressourcen von `Flaticon.com` gestaltet: Lungensymbol von Vitaly Gorbachev; Grabstein von Dreamstale; Beatmeter Patient von wanicon; Intensivpatient von eucalyp; Nieren-, Herz-, Diabetes-, Raucher-, Krebs-, Männer- und Frauensymbole von Freepik aus `Flaticon.com`.

Nun sind wir hier. Es gibt eine Menge Leute die nie erwartet hätten, dass wir es aus diesem vorstädtischen Loch, das wir unser Zuhause nennen, herausschaffen würden. Ich möchte einer riesigen Anzahl von Menschen danken. Danke an Kendrick Duckworth, Karim Joel Martin und Tyler Okonma dafür, dass sie meine Ohren immer warm gehalten haben. Danke, Lofi Girl, für das ruhig halten. Danke an Gauloises Zigarettenmanufaktur und die Pöschl Tabak. Ein großes Lob an die Brauerei Loscher und die Privatbrauerei Krug, die meinen Motor immer am Laufen hielten. Danke an die Eichhörnchen der Stadtbibliothek Nürnberg für ihre emotionale Unterstützung - ich weiß es zu schätzen. Danke an all die Leute, die mit ihrer Open-Source-Software diese Art von Arbeit erst möglich und viel einfacher machen. Danke ihr verrückten Nerds, Gabriel, Michi, Raphael und Valentin, die diese Domäne hervorragend gemeistert haben und immer ein offenes Ohr für mein fränkisches Gejammer hatten, ich hoffe sehr, dass wir in Verbindung bleiben. Danke an meine Freunde und Familie, die zu dieser Arbeit beigetragen haben, Katharina, Simone, Lukas, Simon und Tobias. Besonderen Dank an Max - ich fühle dich. An meinen Bruder Uwe, der immer an mich glaubt. An Renata, die eine der nettesten Personen ist, die ich je getroffen habe, die immer frische Ideen und warme Worte hatte, die mir einen kühlen Kopf bewahrt hat und mich auf dem ganzen Weg unterstützt hat - you rock! Danke Vanadis, dass du monatelang eine Einzimmerwohnung mit mir geteilt hast, dass du mich unterstützt, auch wenn es nicht immer einfach ist, dass du so bist, wie du bist ♡.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acknowledgements

First things first. This cover has been designed using resources from `Flaticon.com`: Lungs icon made by Vitaly Gorbachev; Tombstone icon made by Dreamstale; ventilated patient by wanicon; ICU patient by eucalyp; Kidney, Heart, Diabetes, Smoking, Cancer, Male, Female icons by Freepik from `Flaticon.com`.

Now we here. There is a lot of people that would never had expected us to make it out of this suburbian hole we call home. I want to thank an enormous amount of people. Thank you Kendrick Duckworth, Karim Joel Martin and Tyler Okonma for always keeping my ears warm. Thank you lofi girl for keeping me calm. Thanks go out to Gauloises and Pöschl tobacco industries. Big shoutouts to the one and only Brauerei Loscher and Privatbrauerei Krug who always kept my engine running. Thanks to the squirrels from the Public Library Nuremberg for their emotional support – i appreciate it. Thanks for all the guys that open-source amazing software which makes this work possible and a lot easier. Thanks to all you crazy nerds, Gabriel, Michi, Raphael and Valentin who were amazing fellows at mastering this domain and always had an open ear for my franconian moaning, i sincerely hope we are staying connected. Thanks to my friends and family who contributed to this work, Katharina, Simone, Lukas, Simon and Tobias. Special thanks to Max – i feel you now. To my brother Uwe who always believes in me. To Renata who is one of the kindest persons i have met, who always had fresh ideas, warm words, kept me cool and supported me all the way – you rock! Thank you Vanadis for sharing a single-room apartment with me for months, for supporting me even if its not always easy, for being who you are ♡.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Kurzfassung

In dieser Arbeit entwickeln wir ein visuelles Analysesystem, das die Vorhersage, Analyse und Kommunikation der Ergebnisse von COVID-19-Krankenhausaufenthalten unterstützt. Obwohl mehrere reale Datensätze über COVID-19 öffentlich verfügbar sind, konzentriert sich der Großteil der aktuellen Forschung auf die Erkennung der Krankheit in Röntgenbildern der Brust. Bislang gibt es keine Arbeit, die Erkenntnisse aus medizinischen Bilddaten mit denen aus klinischen Daten kombiniert und die Wahrscheinlichkeit eines Aufenthalts auf der Intensivstation, einer Beatmung oder eines Todesfalls vorhersagt. Darüber hinaus hat sich die bisherige Forschung noch nicht auf die Vermittlung der Ergebnisse an die breite Öffentlichkeit konzentriert. Um die Vorhersage, Analyse und Kommunikation der Ergebnisse von COVID-19-Krankenhausaufenthalten auf der Grundlage eines öffentlich zugänglichen Datensatzes zu unterstützen, der sowohl elektronische Gesundheitsdaten als auch medizinische Bilddaten umfasst, führen wir die folgenden drei Schritte durch: (1) automatisierte Segmentierung der verfügbaren Röntgenbilder und Verarbeitung der klinischen Daten, (2) Entwicklung eines Modells für die Vorhersage von Krankheitsverläufen und Vergleich mit modernsten Vorhersagewerten für beide Datenquellen, medizinische Bilder und klinische Daten, und (3) die Kommunikation der Ergebnisse für drei verschiedene Nutzergruppen (medizinische und klinische Experten, Experten für Datenanalyse und die allgemeine Bevölkerung) über ein interaktives Dashboard. Das Dashboard ist so konzipiert, dass die Benutzer benutzerspezifische Aufgaben lösen können, die ebenfalls in dieser Arbeit definiert wurden. Vorläufige Ergebnisse deuten darauf hin, dass die Vorhersageergebnisse durch die Kombination von medizinischen Bilddaten mit klinischen Daten verbessert werden, während sich die Analyse und Kommunikation von Krankenhausaufenthaltsergebnissen als ein breites und bedeutendes Thema im Rahmen der COVID-19-Prävention erweist.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Abstract

We propose a visual analytics framework to support the prediction, analysis, and communication of COVID-19 hospitalization outcomes. Although several real-world data sets about COVID-19 are openly available, most of the current research focuses on the detection of the disease through chest X-ray images. Until now, no previous work exists on combining insights from medical image data with knowledge extracted from clinical data, predicting the likelihood of an intensive care unit (ICU) visit, ventilation, or decease. Moreover, available literature has not yet focused on communicating such results to the broader society. To support the prediction, analysis, and communication of the outcomes of COVID-19 hospitalizations on the basis of a publicly available data set comprising both electronic health data and medical image data Saltz et al. [2021], we conduct the following three steps: (1) automated segmentation of the available X-ray images and processing of clinical data, (2) development of a model for the prediction of disease outcomes and a comparison to state-of-the-art prediction scores for both data sources, i.e., medical images and clinical data, and (3) the communication of outcomes to three different user groups (namely, medical and clinical experts, experts in data analytics, and the general population) through an interactive dashboard. The dashboard is designed to enable users to solve user-specific tasks, also defined in this work. Preliminary results indicate that the prediction results are improved by combining medical image data with clinical data, while analysis and communication of hospitalization outcomes prove to be a wide and significant topic in the context of COVID-19 prevention.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Contents

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 1

Introduction

1.1 Motivation

COVID-19 is a respiratory disease that became a pandemic in 2020. The virus spread quickly throughout the world, instantly affecting the daily lives of mankind. While the main danger of the disease is the actual illness, it also fuelled social disruption through fake news and alternative facts. This development, termed *infodemic* by the World Health Organization [Hua and Shaw, 2020], is playing down the disease and hindering strategies combating the pandemics spread [French et al., 2020]. At the time of writing, the majority of Europe is hit by a fourth wave with increased hospitalization numbers and deaths. This wave arguably could have been mitigated through higher vaccination rates in response to a more offensive information strategy in highly trusted media by governments [Gehrau et al., 2021; French et al., 2020]. It seems increasingly important to communicate novel insights gained by scientific institutions to the majority of the population in an understandable manner.

Several real-world datasets about COVID-19 are openly available. With the disease affecting the lungs, datasets often consist of chest X-rays solely [Akhloufi and Chetoui, 2021] or inherit other data like computed tomography (CT) scans as well as other clinical data, such as electronic health records [Saltz et al., 2021]. These data can be used to train models for disease detection [Ahsan et al., 2021; Khan and Aslam, 2020; Zhao et al., 2020; Li et al., 2020b] or prediction of high risk patients [Dai et al., 2020]. Until now, no previous work exists on combining insights from medical image data with knowledge extracted from clinical data for COVID-19, predicting the likelihood of an intensive care unit (ICU) visit, ventilation or decease. Neither did available literature focus on presenting results to a broader society including non-domain experts.

TU WIEN Bibliothek Your knowledge hub

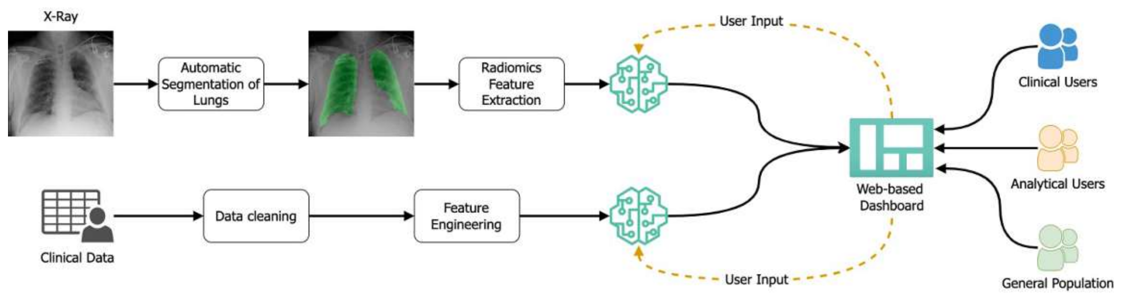


Figure 1.1: The workflow of the thesis. Radiomics features and clinical user data will be used to train respective models which can be combined for a final prediction. Models can also be re-used to predict user input from the application.

1.2 Aim of the Work

The aim of the work is to analyse a publicly available COVID-19 dataset, to predict hospitalization outcomes using electronic health data as well as medical image data and to finally build a visual analytics application that utilizes this data to communicate outcome prediction for different user groups. This is illustrated in Figure 1.1. Outcome communication shall provide novel insights into the complex prediction of COVID-19 in an effort to increase risk perception and provide decision making support to users, while also combining extracted image data features with clinical data records. Therefore, we define three research questions:

1. **Segmentation:** Does the automatic segmentation yield a reasonable performance? Reasonable is defined by an automatic segmentation strategy which reduces artifacts during segmentation on a quantitatively and qualitative evaluation. To what extent can radiomics features be used as biomarkers?

2. **Prediction:** When comparing prediction from chest X-ray images and clinical data prediction to state of the art approaches, does merging the predictions from X-ray and clinical data have a positive impact on the performance? To what extent can patients be clustered into groups of patients suitable for easy abstraction and simpler user input creation?

3. **Visual Analytics:** How does visual analytics help gaining an understanding into the complex prediction of COVID-19 outcome for the three types of users we consider? How do our generated dashboards fulfil the analytical process requirements of each group?

1.3 Methodology

1. **Automated segmentation**
   Segmentation of X-ray images, feature extraction and correlation with *radiomics*.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

The goal here is not to create a state of the art segmentation strategy but to reuse existing methods, in particular making use of transfer learning (Section 3.2.3, 4.4.1). A reasonable segmentation is achieved when we can assure that the automatic segmentation is to some degree free of artifacts. Since segmented ground truths are not provided within the dataset, popular similarity metrics may not be feasible, however other methods have been proposed and were considered [Kohlberger et al., 2012; Valindria et al., 2017] (Section 3.4.2). Besides radiomics it would also be possible to train a convolutional neural network (CNN) for this classification setup. However, as the dataset includes many artifacts, it would be possible that models would learn to predict COVID-19 based on tubes and patients orientation rather than representative lung features [Heaven, 2021].

2. **Prediction**
   Prediction of disease outcome and comparison to state of the art prediction scores for both data sources, i.e., medical images and clinical data. Reasonable performance is determined by the available results found in the literature which we will compare against [Ahsan et al., 2021; Khan and Aslam, 2020; Zhao et al., 2020; Li et al., 2020b], where we find Receiver Operating Characteristic area under curve (ROC-AUC) and other metrics. Appropriate metrics for unsupervised classification are considered when choosing clustering algorithms and their hyper-parameters (Section 3.4.2, 4.3.1).

3. **Visual Analytics**
   A list of requirements is defined that should be solved by the application using the multi-level typology of abstract visualization tasks proposed by Brehmer and Munzner [2013] . Three user groups are taken into account: medical experts and clinicians caring about clinical insights; users with an analytical background such as machine learning (ML) developers interested about prediction models and performances; general population in need of simplified insights. Each group focuses on different tasks. A detailed task analysis is conducted in Section 4.2 and tasks are evaluated in Section 5.2.

4. **Validation & Evaluation**
   The outcome is assessed respectively, using quantitative methods for the segmentation and prediction part and qualitative methods for the visual analytics part, conducting a case study with users from the general population as well as.

1.4 Outline

In Chapter 2 we describe the problem at hand and define our user groups, as well as describing the dataset. Chapter 3 presents related work and the theoretical background of the practical implementation. Here we define terminology, talk about traditional and deep learning approaches for automatic image segmentation and feature extraction, finishing with state of the art. We discuss solutions on how to deal with missing data

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

and large datasets, present data science foundations and finish with an overview about visual analytics. In Chapter 4 we present our implementations using data analytics as well as designing the visual analytics part. Chapter 5 summarizes our results, before we conclude our work and talk about limitations in Chapter 5.2.6.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 2

Background

2.1 COVID-19

COVID-19 is a respiratory disease which first was detected in China in late 2019. The virus quickly spread across the world, and with it unprecedented global restrictions went into place to in an effort to stop this spreading. This was mainly due to presence of high hospitalization rates in older and vulnerable patients for early virus variants, that have been reported between 5% and 18% based on age [Fontanet et al., 2020; Mahase, 2020; Cai et al., 2021] which even increases for patients with preconditions like autoimmune diseases [Akiyama et al., 2021]. Mortality rates were found to lie between 0.66% on average to 13% [Mahase, 2020; Akiyama et al., 2021; Cai et al., 2021]. This not only brought with it the risk of overwhelmed healthcare systems but did so in some regions [Conti et al., 2020; Legido-Quigley et al., 2020] and will be a risk going forward [Mahase, 2020]. Reduced hospitalization is reported for newer variants [Veneti et al., 2022] as well as for the help of quickly develop vaccinations, reducing severe cases with an initial efficacy of 95% for the original virus type also indicating less severe spreading [Polack et al., 2020; Anderson et al., 2020]. Nonetheless, the COVID-19 pandemic is accompanied by an *infodemic* with immensely spreading sources of fake news [Hua and Shaw, 2020], increasing vaccination hesitancy ultimately hindering successful vaccination [Puri et al., 2020; Reno et al., 2021] strategies even before COVID-19 [Carrieri et al., 2019]. Carrieri et al. [2019] show that after an Italian court officially recognized a causal link between autism and a measles-mumps-rubella vaccine, misinformation spread on nontraditional (internet-based) media, which lead to a significant reduction in child immunization rates. Risk perception plays a crucial role in the context of immunization strategies and was researched before [Deiana et al., 2022; Schmid et al., 2017]. Schmid et al. [2017] state that for influenza, risk perception has been a significant barrier to vaccination. This included low worry about the disease as well as not acknowledging vaccination would protect others, which would lead to a social benefit. Reno et al. [2021] report that low risk perception

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

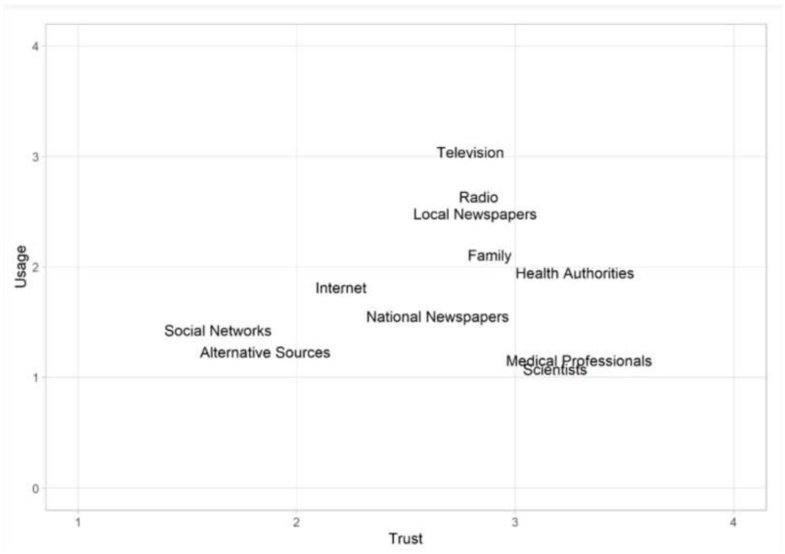


Figure 2.1: Interaction between information media type consumed and trust [Gehrau et al., 2021]. Data was gathered in Germany in late 2020 with 629 respondents.

was one of the major determinants to vaccine hesitancy and influences vaccine acceptance overall. They conclude that it is of fundamental importance to develop designated risk communication strategies for vaccinations, either by dialogue-based interventions or informative tools.

Figure 2.1 shows usage and trust for media used in Germany. Health authorities and medical professionals as well as scientists are believed to have a high trust but are not used as often [Gehrau et al., 2021]. But in times of crisis when medical institutions are on the verge of collapse, medical personnel is scarce and busy. We can help by providing applications that communicate information, helping to save time for clinicians and medical experts as well as providing risk perception tools for the general population by creating online applications. For medical experts, this has been done already numerous times, either by providing decision making support and exploration [Furmanová et al., 2020; Bernold et al., 2019; Furmanová et al., 2021; Floricel et al., 2021] or in efforts to automate disease detection, for instance in radiology [Han et al., 2021; Chen et al., 2017; Shiraishi et al., 2000; Souza et al., 2019], and is a very active field of research. With the pandemic, a lot of research has been shifted towards developing applications with that in mind, focused on COVID-19 detection using medical images [Tamal et al., 2021; Vidal et al., 2021; Zebin and Rezvy, 2021; Li et al., 2020a; Zhu et al., 2020] or electronic health data records [Dai et al., 2020; Li et al., 2020b; Zhao et al., 2020]. Work has not been done on communicating insights from COVID-19 related data using interactive dashboards, where this work focuses on.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

2.1.1 User Groups

The dataset used in this work, the Stony Brooks COVID-19 [Saltz et al., 2021] is consisting of medical image data like chest X-rays as well as electronic health data records. It is presented in detail in the upcoming Section 2.2. That data is of particular interest for three user groups defined:

1. **Medical experts and clinicians**: Medical experts are interested in that data as it could provide decision making support for treatment of upcoming hospitalised patients. This includes comparing and filtering for similar patients based on electronic records and preconditions, which could also be used for risk perception dialogues in individual patient treatments. Medical images are of particular interest, viewing the disease progression of individual patients based on that data. Predicting the outcome for fresh images of new incoming patients could be of interest as well.
2. **Analytical experts**: Analytical users are defined as users having a background in data analytics, for instance machine learning engineers or data scientists. Feature engineering and prediction of this COVID-19 datasets has been a hot topic recently. Thus, users with this background are interested in how data has to be preprocessed and how valid predictions or segmentation strategies are in a real life scenario which can be provided by interactive machine learning solutions [Sacha et al., 2017].
3. **General population**: Everybody is affected by the measurements to detain COVID-19 spread. Layman users without specific backgrounds in medicine or data analytics are viable group to which insights of data available can be communicated which could increase risk perception and better support of certain measurements.

2.2 Dataset Description

The dataset we use in this thesis is introduced here. The Stony Brook University (SBU) COVID-19 Positive Cases Dataset [Saltz et al., 2021] includes a variety of medical information in form of tabular data on a per patient level. Patients included in the dataset were all hospitalized and tested positive using a polymerase chain reaction (PCR) test for COVID-19. The dataset is abbreviated in the following as SBU dataset. Medical data is available for one observation per patient, which was chosen using a specific algorithm [Saltz et al., 2021]. Variables are available in numeric, boolean and categorical form and will need different preprocessing strategies which are presented in Section 4.1. The dataset contains many missing values which are dealt with using imputation methods.

Additionally, among others, it contains medical image data in form of chest X-rays (CXR) images without segmentation masks. Here, multiple images can be available for a patient. The format of the images is DICOM[1] which is a standard format in medical imaging and output of many medical imaging machines.

[1] `www.dicomstandard.org`

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Age ([18-59; 59-74; 74-90])	Gender	Kidney Replacement Therapy
Hyptertension	Diabetes	Coronary Artery Disease
Chronic Kidney Disease	Cancer/Tumor	Chronic pulmonary disease
ACEi - HBP treatment	Antibiotics	ARB - Usage of blood widener
Smoking status	Cough	Shortness of breath (Dyspnea)
Vomiting	Diarrhea	Abdominal pain

Table 2.1: Medical data features including medical history, acute symptoms and general information. ACEi: Angiotensin-converting enzyme inhibitors, HBP: high blood pressure, ARB: Angiotensin Receptor Blockers.

2.2.1 Overview

Tabular data is available for download in form of a Comma-Separated-Value (CSV) file with $N = 1384$ patients. For each patient there is one record in the file which is consisting of information for the patients worst day during the hospitalization stay. In its raw form there are 131 features available, including identifiers, different potential target variables as well as partly redundant information.

In general the medical data, also called electronic health data or clinical data in the latter, consists of two parts in general. One part is historic health data including pre-existing conditions such as diabetes or heart problems as well as drug usage and symptoms, as shown in Table 2.1. The other part is mainly laboratory results following medical tests during the hospital stay, such as blood oxygen levels. Some of this information is also already specific to a certain state of the patient during the hospital stay. An example for that would be days ventilated, which indicates how many days a patient has been ventilated already. Those features have a high impact on some of the target chosen in the following work and might be removed in order to keep any trained models fair and prevent information leakage.

Overall there are $562,376$ medical images available in the dataset. Using the online tool from the cancer archive website[2] images were filtered by anatomical site and all *Port Chest* images were downloaded ($N_{CXR} = 10,526$, 170GB). This included 1330 patients. Port chest refers to portable CXR which have been recorded with mobile X-ray systems. The manufacterer is labelled as *Carestream Health* but no detailed model was provided. All images are recorded in anterior to posterior (AP) direction, meaning images are taken from front to back. Generally available CXR datasets are recorded in the opposite direction, posterior to anterior, [Jaeger et al., 2014] which naturally decreases the amount of hidden lung structure due to ribs and other corpora.

Images are often available per day. In most cases at least two images in different contrast settings are provided. Figure 2.2 shows high contrast and normal CXR images from the

[2]https://nbia.cancerimagingarchive.net/nbia-search/?MinNumberOfStudiesCriteria=1&CollectionCriteria=COVID-19-NY-SBU

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

SBU dataset as well as samples from the Montgomery County (MC) and Shenzen datasets (SD) which are used later for transfer learning. Both datasets have been proposed by Jaeger et al. [2014]. Images of those datasets are recorded posterior to anterior (PA). In order to have a solid basis for the transfer learning process, images should be as similar as possible between the datasets. Thus, high contrast images will not be applicable for transfer learning and are not used further. Nonetheless the remaining CXR will introduce increased artifacts due to tubes and cables being present in the pictures. In Section 4.4.2, image pre processing is evaluated which could increase similarity between transfer learning and available CXR images.

There were several patients in the dataset that did not have normal contrast images. Those patients have been removed, so that only patients in the dataset that have *at least* one rateCXR in a normal contrast setting are kept for further tasks. This resulted in a final number of $N = 1279$ patients and 4728 CXR (89GB). 174 (13.6%) patients deceased, 257 (20.1%) haven been admitted to ICU and 213 (16.7%) were ventilated during their hospitalization stay.

From those 1279 patient records left in the tabular data, only two were without any missing records. Only 13 features did not have any missing records. This included target variables, unique identifiers and some variables closely related to target variables but also patients age. The strategies with which this dealt are described in Section 4.1 in detail.

2.2.2 Prediction Targets

The dataset inherits multiple variables that could be chosen as prediction targets. In this work we chose:

(i) the outcome of the decease, as deceased or dismissed from the hospital;

(ii) ventilated during the hospital stay and

(iii) admitted to the ICU.

Focusing on these three outputs makes sense from a clinical standpoint as it helps in resource planning for limited areas such as ICU or ventilation.

The prediction can now be defined as three separate binary classification cases, which would lead to several representation issues for the dashboard and potential users, as we would not be inherently able to produce a implicitly logical outcome for a user including all targets. Three separate classifiers would lead to three independent predictions. In order to increase representation and communication of the results, we opted to combine the classes to a multi-class classification setting. This reduces complexity and enables communication of results in form of natural probabilities for an outcome to potential users for the dashboard.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

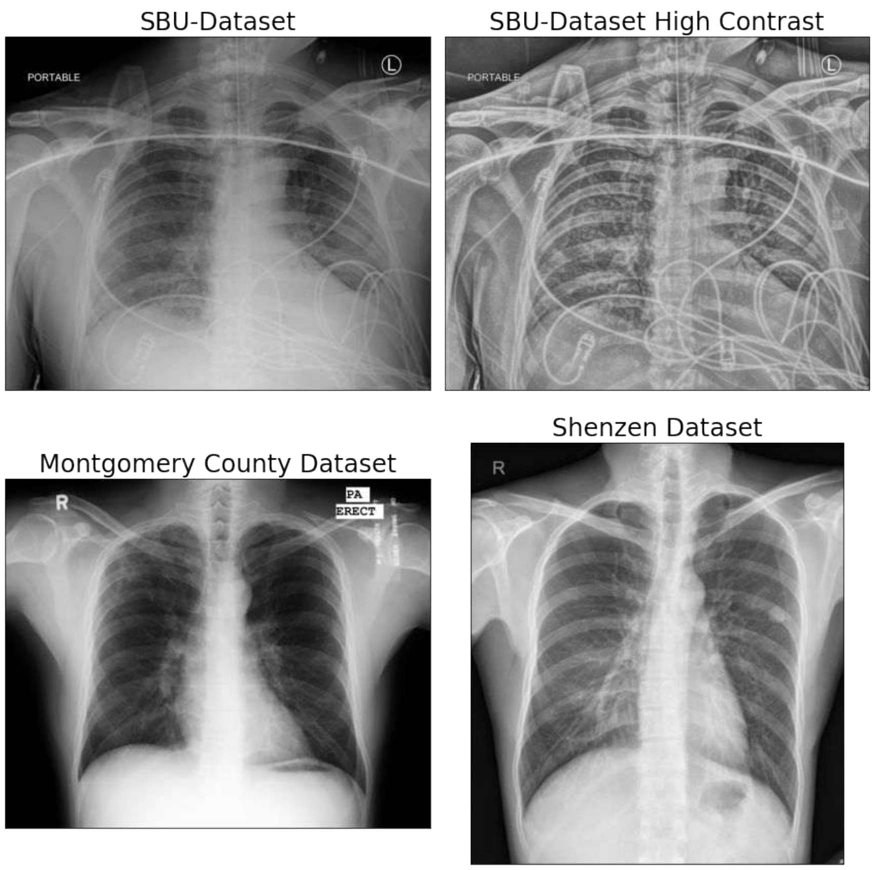


Figure 2.2: Images from the Stony Brooks COVID-19 Dataset, first row; and the Montgomery County and Shenzen dataset [Jaeger et al., 2014] used for transfer learning later on, second row. High contrast images were removed from the dataset.

But this also introduces several other problems like class imbalances, how many classes to present to a user or multi-class classification as a more complex prediction setting per se which need to be handled by the application and during the thesis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

CHAPTER 3

Theoretical Background & Related Work

This chapter gives an overview about the current state of the art in the fields relevant to the thesis. Currently, new COVID-19 related research is published at a high rate. The evolving field is influenced by different disease variants, daily updated vaccination rates and growing data sets over time. First, an overview is given about recent work in COVID-19 related classification and disease prediction.
This is followed by a recap of traditional and state-of-the-art automated computer vision approaches including deep learning, where we focus on image segmentation in medical problem definitions, while repeating basic understandings and problems from the domain. Additionally, a recap about working with clinical data of tabular form is presented, given the dataset used in the thesis is composed of both medical images and clinical data. Common issues including missing or dirty data are presented, as well as strategies on how to deal with them. Finally we present established views on presenting data and designing interactive visualizations using *visual analytics*, defining user groups and respective tasks for a visualization.

3.1 Recent Work about COVID-19

In general, related literature can be divided between approaches dealing with classification of COVID-19, mainly using medical image data like CT or X-rays, and prediction of the disease outcome. Prediction approaches focus on clinical data in form of patient records.

3.1.1 Classification

There have been several publications in the field of computer science about COVID-19 since the outbreak. Many works focus on the detection or classification of the disease while

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

some of them use CT scans instead of X-ray images. Ahsan et al. [2021] a Convolutional Neural Network (CNN) and Histogram of oriented Gradients (HOG) to detect COVID-19 in chest X-rays. For the same problem, Khan and Aslam [2020] applied a transfer learning approach to create a CNN classifier by combining publicly available datasets. Furthermore, online challenges exist tackling the problem of COVID-19 detection in X-rays [Akhloufi and Chetoui, 2021].

3.1.2 Prediction

Publications about predicting the outcome of the disease are not so commonly encountered.

Dai et al. [2020] categorize patients into different risk groups using a combining data from different hospitals in China (N=419). They create risk groups definitions by determining four core features that influence the risk of a severe COVID-19 infection. Determination of important features was done using statistical methods and classification is done using three risk classes by utilizing a scoring model. Severity was diagnosed using heart and pulmonary function recovery and lung CT findings. There is no further information on how lung CT was used defining the severity and it was not included in the final risk score model.

Zhao et al. [2020] use a different version of the Stony Brooks Dataset with N=641 patients to predict ICU admission and mortality on clinical data which includes laboratory findings as well. They train Logistic Regression models for ICU admission and mortality and identify key predictors from the resulting models without giving information about how well the Logistic Regression models performed in the first place. From that, similarly to Dai et al. [2020], they choose most important predictors by selecting highly significant features distinguishing non-admitted and admitted patients to ICU as well as non-deceased and deceased patients. Patients that received ventilation were counted into ICU group. The final model classifies between five risk groups for ICU admission and seven risk groups for mortality using a scoring model that is limited to the most significant features.

Li et al. [2020b] compare several machine learning models based on underlying health conditions, travel histories and chronic diseases predicting the likelihood of a patients death. They investigate two datasets available online with N_{github}=28,958 and $N_{wolfram}$=1,448 patients with positive rates 1.83% and 8.5% indicating a mortal disease outcome. Due to the low mortality rates, they treat the prediction as an anomaly detection. They have no access to medical image data for the patients.

3.1.3 Biomarkers

The most used definition for biomarkers comes from Strimbu and Tavel [2010]. They define biomarkers as any “objective, quantifiable characteristics of biological processes“, which may correlate with a reported patient’s current feeling. They differentiate between biomarkers, *Clinical Endpoints* and *Surrogate Endpoints*. Clinical Endpoints are solely

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

biomarkers that are variables representing a patient's current feelings from their own perspective, including symptoms and well-being. These Endpoints have been thought of as the only biomarkers for some time. But with the ultimate goal of improving survival rates in clinical research and medical supplement trials, other well-defined variables can also be used to model a patient's disease progression. Those variables exceed the definition of Clinical Endpoints by being not necessarily actively felt by a subject anymore. Those variables are extremely helpful in early diagnostics when a treatment is taking a long time and the clinical outcome may not be available as a near-immediate response to the medical therapy used, or when a new disease is tackled with novel therapies where no or infrequent survivals are present. When a biomarker is used over a clinical outcome as a target, they are defined as a Surrogate Endpoint. There has to be solid scientific evidence though that a biomarker is able to predict a clinical outcome, for instance by showing high correlation (and causation).

Medical biomarkers have been defined in multiple publications regarding COVID-19 disease progression [Ponti et al., 2020; Kermali et al., 2020; Malik et al., 2021]. They agree on lymphocyte, D-dimer, aspartate aminotransferase, creatine kinase, troponin, C-reactive protein, procalcitonin and erythrocyte sedimentation rate as clinical biomarkers found to be correlated with severe COVID-19 cases including ICU admission, ventilation and mortality.

However, biomarkers can also be defined in medical image data, most prominently in cancer research [Zwanenburg et al., 2020; van Griethuysen et al., 2017; Chen et al., 2017; Udeshani et al., 2011]. Image biomarkers are special regions of an image, such as volumes or intensities, that characterize image contents, also referred to as image features [Zwanenburg et al., 2020][Chapter 1]. The extraction process from medical image data was thought to be not clearly defined, with a lack of consensus guidelines, which led to the *Image Biomarker Standardisation Initiate* [Zwanenburg et al., 2016], proposing a standardized way of working with medical image data, including segmentation, feature extraction using reproducible methods based on and other preprocessing techniques [Zwanenburg et al., 2020][Chapter 2]. Software such as *PyRadiomics* is developed under close consideration of these standards.

In the context of COVID-19, no literature focused on the definition of biomarkers has been found for CXR data. Pu et al. [2020] investigate how CNNs can be used to identify biomarkers in CT images, especially those that distinguish COVID-19 from pneumonia. They conclude that there may not exist image features that would distinguish COVID-19 from pneumonia. Apostolopoulos et al. [2020] also use CNNs to discover new reliable biomarkers in CXR images. Their work focuses on classifying COVID-19 from six other most common pulmonary diseases given CXR images. The authors conclude that due to the high classification accuracy, biomarkers for COVID-19 could exist and propose further work focused on extraction via radiomics might be successful.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

3.2 Segmentation & Feature Extraction in Medical Imaging Data

Automatic image processing is a very active field of the current scientific landscape in general, with medical image processing being one special domain. Different goals have been formulated in this context, stemming from classification and object detection, clustering or segmentation of images. Outcomes of modern systems vary greatly in terms of goals and expectations, ranging from automatically detecting diseases, semi-automatic systems for clinicians — including humans in the loop — to interactive solutions providing decision support for experts solely. Solutions are motivated by disease prevention, detection or classification and are derived from medical imaging data. Medical image data comes in many different shapes and formats, from CT scans to X-ray images and magnetic resonance imaging (MRI) data. Complexity is only increased by different diseases that need to be detected by computer vision systems, which results in a high number of specialized, tailored solutions for diseases based on available data. The lack of sharing data and solutions with the public is another aspect fuelling challenges about understanding and reproducing successful solutions [van Griethuysen et al., 2017].

3.2.1 Traditional Approaches

When processing image data with computers, it is first important to understand how humans actually process images. Humans have developed a great sense of object detection through evolution for the sake of survival in ancient times. Much of human perception happens unconsciously, making use of high-level features, as proposed by Ullman [2000]. Different theories exist on how this is achieved in such a quick way, taking into account prior knowledge, expectations and temporal continuity. A simple but not very likely way of thinking about this is that we simply store a lot of views of objects in our brain, and, whenever we perceive something, we compare it to stored images in our brain, similarly to a database lookup. With so many memories that are kept and accessed unconsciously, this theory does not align with the fact of the sheer speed with which humans are able to perceive images. Therefore we are arguably able to abstract things in a very efficient way.

When trying to teach machines how to perceive images, we ultimately struggle in describing how we do it so quickly. While explaining why we perceive certain objects as we do, we rather fall back to reason with low-level features that make up the high-level features. Therefore we define typical shapes and characteristics of objects, made up of edges and colors, backgrounds or other context information. For instance, when we are defining the characteristics of a cat in an image, we would most likely argue that it typically has whiskers, a pair of ears, a tail and four legs. Teaching that to a machine in a deterministic way, for example when building a simple rule-based approach, is not straightforward, especially when objects occur in various angles, sizes and variations as depicted in Figure 3.1. Hard coding rules would not make much sense and the results would be highly unreliable while the design of such a system would be very expensive and inflexible.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

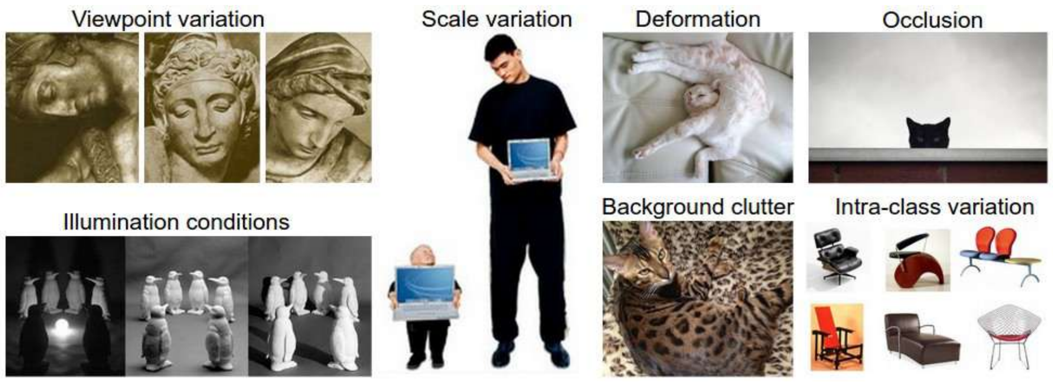


Figure 3.1: Different challenges when dealing with image processing in general [Li et al., 2022].

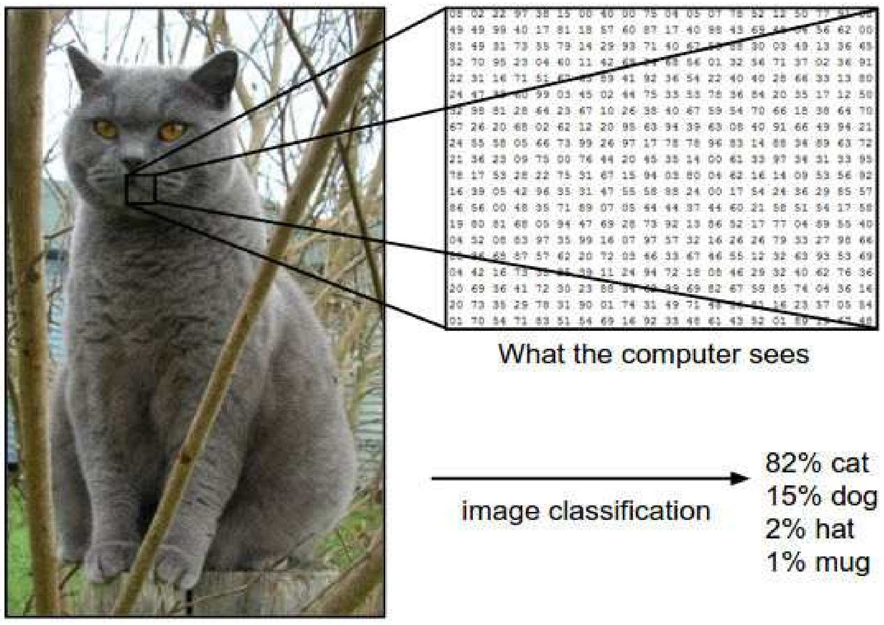


Figure 3.2: How computers see images. Information is defined as pixel values that are the basis for computation [Li et al., 2022].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Figure 3.2 shows a cat and how the computer stores this image in form of pixels as a 2-D array. In fact this can be represented as a stacked 2-D array in three dimensions for each color channel, if available. This allows for reliable low-level feature extraction techniques based on color, edges and shapes. Prince [2012] offers an extensive overview of models and techniques used in this context, including SIFT, Bag of Words and Histogram of Oriented Gradients (HoG) to name a few. While those algorithms exploit the fact that color information in images is available, others have been proposed that focus on gray scale images where only one color channel is available (see 3.2.4). Dealing with gray scale images is in fact most often the case in the medical domain, given that radiological and CT devices generate gray scale or intensity varying images.

Feature extraction methods may result in a high number of dimensions once applied, thus are often followed by dimensionality reduction step (subsection 3.3.3). Those reduced features have been used as the input for previous state-of-the-art machine learning algorithms like logistic regression or support vector machines in automated image processing applications [Girshick et al., 2014]. This is also shown in Figure 3.4.

While working towards building robust techniques for image processing, one important topic is also how different objects occur in images as depicted in Figure 3.1. Objects can be observed in different viewpoints, scales and illumination which influence the feature extraction process, especially for low-level pixel based representations. Deformation, illumination and intra-class variation are not solely a challenging aspect in medical imaging data, but they represent a very specific problem statement that should be explicitly highlighted.

Segmentation

Image segmentation has been addressed through a number of approaches, including different problem definitions given the characteristics of a dataset and defined outcomes. The goal in segmentation is to label pixels to a set of class labels [Prince, 2012, Chapter 7.9.3], aiming at describing regions in an image that are homogeneous, having similar texture or intensity levels [Pham et al., 2000]. This describes a per-pixel classification setting. Prince gives another definition of segmentation, as the attempt to infer the boundaries of objects which can be solved by contour detection [Prince, 2012, Chapter 11.8.4]. In each way, segmentation extends the problem definition of simple object detection by providing a solution for each pixel in the input. Generally, segmentation in medical images refers to isolating pixels or voxels that satisfy a predetermined criterion (homogenity, contour, etc.).

Multiple segmentation strategies have been proposed [Prince, 2012; Van der Walt et al., 2014]. Contour based approaches estimate boundaries in the image based on force computation that can handle data without GT available. They are generally known for outputting smooth segmentations. Known models are active contour segmentation based on *snakes*, *chan vese* and well as morphological enhancements for increased computational performance [Chan and Vese, 1999; Kass et al., 1988; Márquez-Neila et al., 2014]. When

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

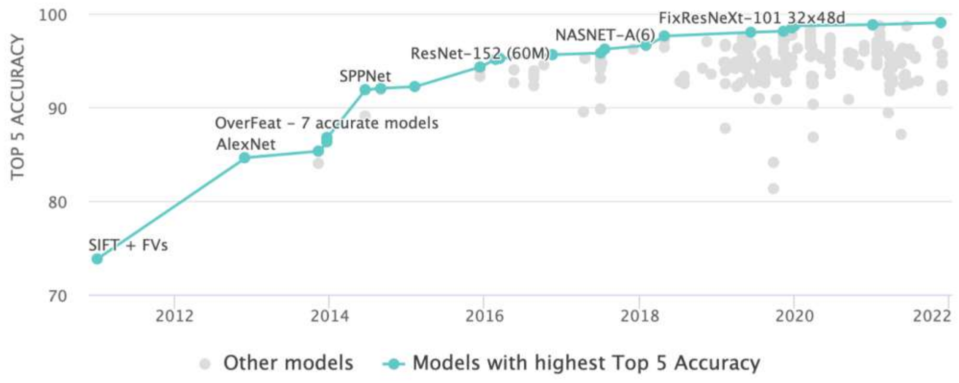


Figure 3.3: Top 5 Accuracy on the ImageNet Dataset. A high increase is observed after the introduction of AlexNet in 2013, going beyond traditional low-level feature detection (SIFT) [ima, 2022]

segmentation is defined as an unsupervised machine learning setting without GT available, clustering methods like *K-means* or hierarchical clustering (see Section 3.4.4) can be leveraged that provide per-pixel classification [Pham et al., 2000].
Making use of standard classification models in a supervised learning setting with GT available for training, defines another per-pixel classification method that is often preceded by feature extraction techniques. This also included earlier versions of neural networks that have been used in state-of-the-art applications for segmentation now, as described in the upcoming section.

One particular problem in image segmentation in the medical context is the availability and generation of labelled data [Ronneberger et al., 2015]. While ordinary people are able to create labelled ground truth segmentation masks of, again for instance a cat in a picture, labelling organs in medical images is not trivial and asks for the input of medical domain experts or radiologists in that particular case. This makes labelling non-trivial, error-prone and time expensive.

3.2.2 Deep Learning

Until 2012, the best performing applications for automatic image processing were based on techniques that leveraged on low-level feature extraction as described in Section 3.2.1. Only marginal improvements were being achieved over time [Girshick et al., 2014]. The proposal of neural networks for computer vision tasks significantly improved results in the are, as depicted in Figure 3.3 for the ImageNet classification dataset. Since then, the top performing algorithms on open access datasets have been built on the basis of deep learning, specifically convolutional neural networks (CNN, Section 3.2.2).

Deep learning networks aim to not only abstract from low-level features previously extracted manually, or learning from pre-defined features, but learn those representations by itself. This is called *representation learning*. Figure 3.4 illustrates the schematic

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

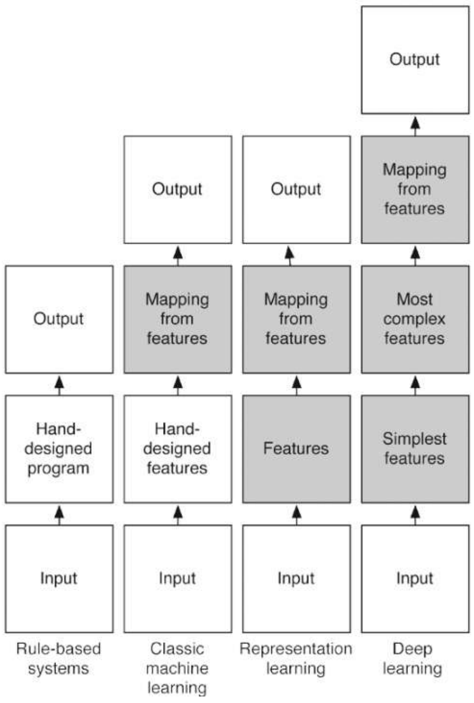


Figure 3.4: Schematic differences between image processing techniques. Gray boxes depict automated processes by a machine. Rule-based systems were created with domain knowledge. Machine learning learns these rules based on extracted features. With representation learning, the goal is to also learn feature extraction as well, while deep learning enables learning of different feature layers. Illustration adapted from Bengio [2015].

differences between the different approaches. They also stand out by being learnt end-to-end, meaning no manual work is required once training is completed.

This section focuses on presenting the theoretical background of deep learning which includes presenting theories about how those networks are trained, what they are built on and how optimization works. Finally, state-of-the-art segmentation strategies are presented.

Linear Model

The starting point for deep learning based networks is the simple linear model. In the context of object classification in visual computing we start by defining a function that maps inputs, e.g. the pixel values of an image, to some output, e.g. confidence scores for each class. We thus define

$$f(x_i, W, b) = Wx_i + b \tag{3.1}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

with x_i as the input pixels of the image flattened to a one-dimensional vector, the weights or parameter matrix W and the bias vector b. The shape of x_i is defined by the image dimensions, as $Width \times Height \times Channels$, so in an 24×24 with 3 color channels one would deal with $24 \times 24 \times 3 = 1728$ inputs. The weight matrix W is of shape $x^T \times c$, with c being the number of classes to be predicted, resulting in a row for each class in the weight matrix. The bias vector b is added to the equation. The output of the function is again a vector of length c, providing scores for each class, defined as vector s for scores. By using the *kernel* or *bias trick*, b can be appended to W. As a result 1 has to be appended to the input vector x_i, optimizing the equation [Goodfellow et al., 2016, Section 5.7.2]. Based on this definition the weights and bias vector can be trained *end-to-end* using a set of training images with the number of parameters to train defined as $N_p = x_i \times c + c$.

One can now stack many layers of this form next to each other to perform a chain of functions, so that $\bar{y}_i = f_1(f_2(f_3(x_i, \theta_3), \theta_2), \theta_1)$, with $\theta_1 = W_1 b_1$ of layer 1, $\theta_n = W_n b_n$ of layer n, which are then called *deep networks*. The layout of those networks is commonly thought of as a oriented computational graph (compare Figure 3.6).

Defining Cost Functions

The outcome vector of the linear model s_i, defined as the solution of Equation 3.1 for an image x_i, defines the confidence scores of the model for each class c. In order to train the model a loss function is defined that changes the weights in W, so that the equation is optimal for a training set. Different loss functions can be used in order to train deep learning networks, depending on the task definition. Most commonly used is the *cross entropy loss* in combination with the *softmax* function classifier [Ronneberger et al., 2015; Goodfellow et al., 2016].

The class scores of the model s_i may be arbitrarily high positive or negative numbers. The *softmax* function is used to transform those scores to a probability distribution, so that

i) gives a class confidence score between $[0, 1]$ for each class,

ii) sums up all class confidence scores to 1

The *softmax* function has been formally defined as

$$softmax(s)_i = \frac{exp(s_i)}{\sum_j exp(s_j)} \tag{3.2}$$

for class probabilities s_i and other classes s_j in C, with C being the set of classes c. The weights of the network are now trained by minimizing the *cross-entropy* score. This describes the attempt to model the empirical distribution of the true class labels in a training dataset by the models score distributions. This is done by maximizing the log-likelihood of the *softmax* linear models scores. The idea is to achieve a similar

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

distribution of output scores from the input sample by minimizing the cross-entropy score. The cross-entropy loss for weights and biases θ is defined as

$$L(\theta) = H((\hat{y}_i, softmax(x_i, \theta)) \tag{3.3}$$

Cost functions can be varied over implementations and enable networks to optimize over arbitrarily many problem definitions (compare section 3.2.3).

Back-propagation

When a prediction $\hat{y}_i$ over input x_i is produced, we traverse the network in form of a forward propagation. Training is done backwards by deriving information from the loss function. This is called *back-propagation*. In order to train the weights of the network, we compute the gradient of the cost function. For each sample i in the training dataset, a forward propagation is performed to receive the prediction scores from the classifier in form of vector s_i. The weights in each layer are adjusted backwards in order to decrease the loss by recursively going back to earlier layers and computing the derivatives for each neuron to connected neurons in previous layers. If this is done for each training sample in the training set, the action that is performed is termed *stochastic gradient descent* (SGD).

In practice, training samples are randomly shuffled and grouped into batches and the weights are optimized over each batch. This increases computational performance and the optimization is done in larger steps. The optimization is continued for all batches to include all training samples, for one *epoch*. Multiple epochs are used, in which training data is shuffled and randomly augmented by rotating, cropping and randomly masking images to increase generalization and number of training data over a training set.

Multiple enhancements for the back-propagation algorithm have been proposed, which are summarized by Goodfellow et al. Goodfellow et al. [2016, Chapter 8]. A *learning rate* is introduced to adjust how fast the weights are adjusted over epochs, helping battle overfitting and the vanishing or exploding gradients problem. Using *momentum* adaptively increases the step size for steps oriented in the same direction over continuous training batches. Most prominently *ADAM* is used here. Skip connections were introduced to reduce the distance from the loss to earlier layers, fastening training speeds and making them more robust (compare section 3.2.3).

Non-Linear Models

The previously described linear model can already do a reasonable job for linear classification, but most problems are non-linear. How a linear model is unable to solve simple non-linear problems can be observed when trying to model a XOR function [Goodfellow et al., 2016, sec. 6.1]. In order to model non-linear phenomena we need further tools. To solve non-linearity so called *activation functions* are used which are added to models in

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

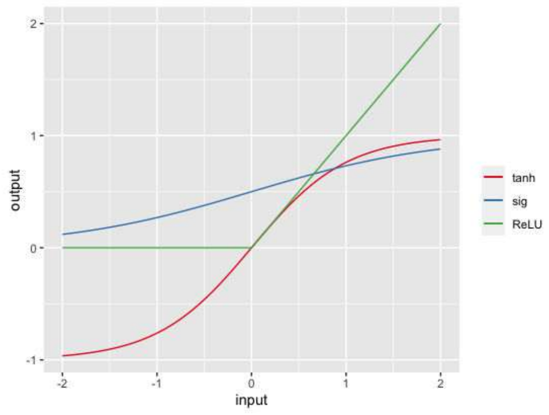


Figure 3.5: Most popular activation functions.

the form of additional layers that feed of previous layers weights. The most popular activation functions used today are shown in Figure 3.5. Models using non-linear activation functions have been termed feed-forward Multi Layer Perceptrons (MLP) or feed-forward neural networks. By definition, those networks flow is only one-directional, while multi directional networks called Recurrent Neural Networks, prominently used in language detection, can also have connections into previous layers [Goodfellow et al., 2016, sec. 10.6].

Convolutional Neural Networks

Training previously described models that try to learn representations using only one layer might be possible in theory, but in practice it turns out that these so called *shallow networks* are difficult to optimize [Bengio and Delalleau, 2011]. They tend to generalize not very well, leading to overfitting models, unless presented with a huge amount of training data. Additionally they are quite cost-intensive in terms of computational resources, which delayed usage of neural networks until the late 2000's, when computing became more affordable and training became less expensive. Apart from that, learning is still based on pixel values being used as input. When stacked to deeper networks, MLP's are designed as fully connected networks, connecting all layer outputs to all inputs in the next layer. This radically increases the amount of parameters to be learnt in deeper network and with that, the computational effort for training such deep networks. Increasing input sizes of millions of pixels in high resolution images have the same effect as the layers become very large.

Deep, convolutional neural networks (CNNs) were introduced to address issues from traditional feed forward neural networks and provide ways to finally represent the possible abstractions in order to learn high level features. They are commonly used on grid-like data inputs like discrete time data and images. The name stems from the usage of mathematical operations called *convolutions*. A convolution in deep learning describes

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

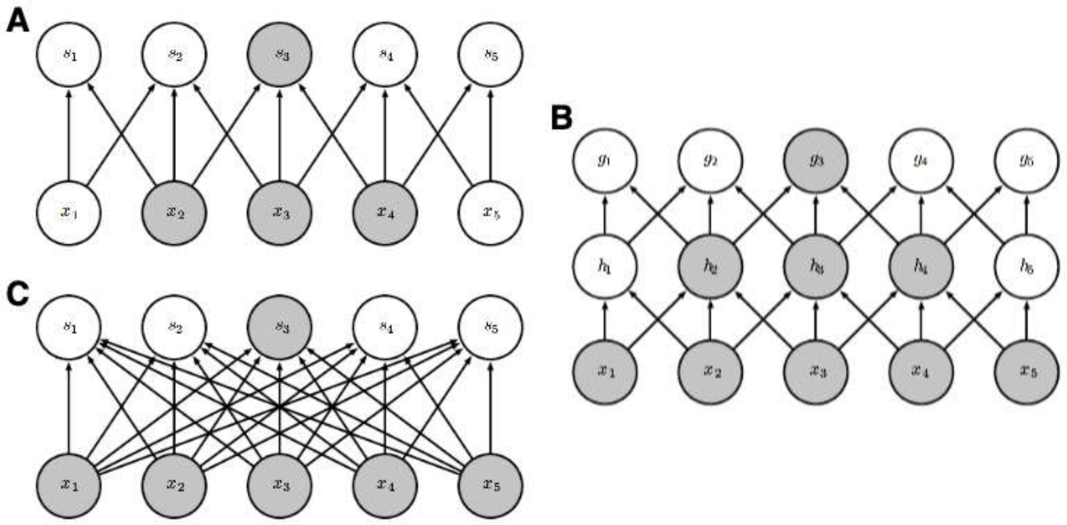

A
s_1 s_2 s_3 s_4 s_5
x_1 x_2 x_3 x_4 x_5
B
g_1 g_2 g_3 g_4 g_5
h_1 h_2 h_3 h_4 h_5
x_1 x_2 x_3 x_4 x_5
C
s_1 s_2 s_3 s_4 s_5
x_1 x_2 x_3 x_4 x_5

Figure 3.6: Sparse and dense network connectivity [Goodfellow et al., 2016, Section 9.2]. Highlighted neurons in earlier layers influence highlighted neurons in later layers. Connections are directed edges. **A** represents a Dense CNN, where the receptive field is smaller than in the equivalent network displayed as a fully connected network (**C**). **B** shows how the receptive field gets bigger over many convolutions in a dense CNN.

the method of sliding over the input with a constant window and performing a matrix multiplication with a specific matrix, called a *kernel*.

Kernels allow the network to reduce the size of the original image or preceding layers onto subsequent layers. By that, offering a way of extracting lower level spatial relations in an image in earlier stages and become high level representations of image regions in deeper layers of the network. Typically this is done in parallel with multiple kernels at once, resulting in a number of so called *feature maps* per layer, with each ideally specialising in different detections as depicted in Figure 3.7. Simply put, when comparing this to traditional approaches in Section 3.2.1, we are now able to train feature extractors such as edge detectors automatically. For each feature map a kernel is learned that is re-used in that particular feature map, sharing its values throughout the map. This parameter sharing is particularly useful as it reduces the number weights to be trained per layer by a large margin, depending on the size of the feature map. Stacking this process into multiple layers allows higher features to be detected by increasing the so called *receptive field* of a neuron, as illustrated in Figure 3.6.

A side effect of the convolution is that without enhancements, they shrink the input by $k-1$, where k is the size of the $k \times k$ convolution. This is illustrated in Figure 3.6, as in graph A and B, a 3×3 convolution results in one neuron in the subsequent layer. We can control this by padding images, usually with zeros, which is then called *zero-padding*. Padding a 2-D input by zeroes and using a 3×3 convolution results in the same output

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

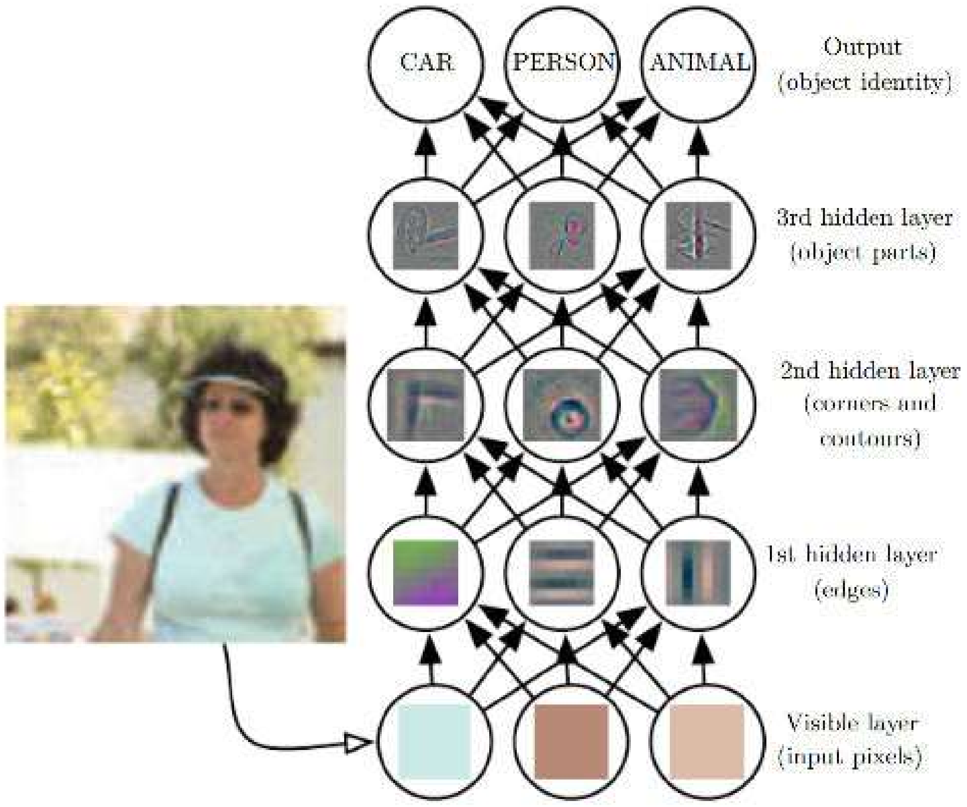

CAR
PERSON
ANIMAL
Output (object identity)
3rd hidden layer (object parts)
2nd hidden layer (corners and contours)
1st hidden layer (edges)
Visible layer (input pixels)

Figure 3.7: Illustration of the inner processes in a CNN. Multiple parallel feature maps are trained in hidden layers on the input pixel values using convolutions. Earlier layers specify lower level features while subsequent layers will gradually learn more high level representations. Image from Goodfellow et al. [2016].

dimensions as the input. Without padding, network learning would be harder because we either have to shrink the input quickly over consecutive layers in a CNN, or be limited to using only small kernels in order to keep the shrinking size small.

However, the convolution is only the first of actually three steps of one layer in a convolutional neural network [Goodfellow et al., 2016, chapter 9.3]. The outputs of the convolution step are first run through a non-linear activation function as presented in section 3.2.2, most commonly the rectified linear activation function (ReLU). This is done via a sparse 1-to-1 mapping. This is followed by a *pooling* layer, which is implemented to increase robustness of the learned features. Pooling is usually done with a sliding window over the input. Common networks implement 2×2 max-pooling layers, which would only choose the maximum value in a 2×2 pixel window. This increases the invariance to specific data transformations, e.g. moving the original pixel values to the right by one pixel would have no immediate effect since the maximum value should stay the same for most of the pooling windows. As a side effect, it reduces the size of the signal. This

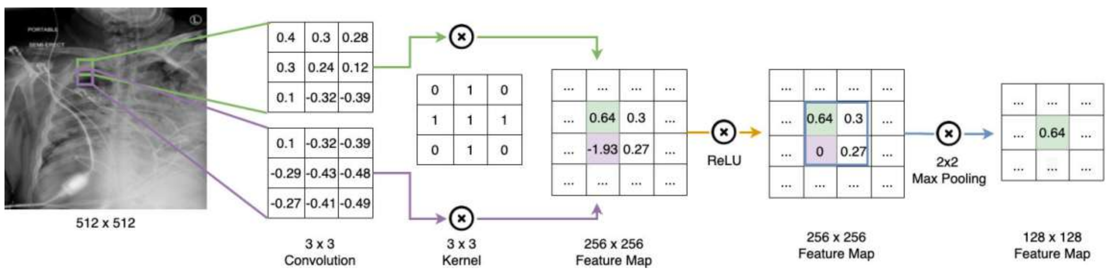


Figure 3.8: Convolution of one feature map with 3x3 kernel, 1-padding of the input and stride 2. The stride > 1 forces the feature map to be of smaller size, downsampling the image implicitly. An activation function is used (ReLU) which sets negative values to zero (compare Figure 3.5), which is followed by a 2x2 max pooling operation which again, is reducing the feature maps dimensions. This is repeated in parallel many times to create many feature maps with different kernels.

reduces runtime for subsequent layers while increasing the receptive field of the neurons in the next convolutional layer. By increasing the pooling sizes, optimization is increased as well but also a loss in detail is introduced to where exactly a certain feature has been located in a picture when using bigger pooling layers. Pooling enables effectively working with different sizes of images by consecutively reducing the images sizes, so that in the end, a fixed number of representations for an image is reached which can be used by a classifier (softmax) or as an input for other tasks shown in section 3.2.3. The process of shrinking the input throughout the flow of the CNN is also called *downsampling*.

Downsampling can be fastened by using convolutional kernels with a stride $s > 1$, in which we would jump over inputs. This implicitly combines the convolution and pooling layers into one operation in a computationally more efficient way. A schematic illustration of a convolution is given in Figure 3.8.

3.2.3 Solutions in Deep Learning

After presenting the foundations of visual computing using traditional low-level feature extraction approaches as well as deep learning foundations, this section continues with introducing the state-of-the-art solutions in regards to medical image segmentation.

U-Net

In 2015, Ronneberger et al. proposed the U-Net architecture for image segmentation, winning multiple challenges [Ronneberger et al., 2015]. The approach was motivated by reducing the high number of training images needed for training CNNs, which is often not applicable in biomedical image processing. Previous approaches were deemed too slow.

The U-Net embodies the idea of an encoder-decoder network. It first downsamples the input using convolutions and non-linear functions to learn a small representation vector,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

from which it creates an output again. From the input space X, a smaller representation C is learnt, which is then the basis for producing the output D of the network. This approach is most commonly used in several natural language processing approaches [Goodfellow et al., 2016, Section 10.2]. But the U-Net introduced an important difference to regular encoder-decoder networks: it makes use of skip connections which enable direct connections between the encoder and decoder parts, also illustrated in Figure 3.9. The name ultimately stems from the designs layout: since the expanding right decoder part of the network is also using multiple convolutions similar to the convolution side, it looks like a "U" or sometimes also called "V". In general, since deeper networks generalize better, deeper and deeper networks have been proposed, which take more effort to train in respect to time and resources. This led to a common problem in optimizing the gradient of deep neural networks, called the vanishing or exploding gradient. When networks became too deep and the gradient was not transported back to earlier layers in some cases [Glorot and Bengio, 2010]. Introducing skip connections allows the gradients flow to move faster between parts of the network, speeding up optimization [Goodfellow et al., 2016, Section 6.4.2]. This allowed the U-Net to train on a quicker and better on small train sizes. Deeper networks like *ResNet* proposed later also heavily rely on skip connections to improve learning speed and to deal with the vanishing gradient problem[He et al., 2016].

In order to generalize better and to increase the train size for the U-Net it was trained using data augmentation. Since the U-Net was proposed for image segmentation, loss optimization was adopted accordingly. The proposed softmax output and corresponding cost function for the U-Net is represented in Equation 3.4:

$$p_k(x) = \frac{exp(a_c(x))}{\sum_{c'=1}^{C} exp(a_{c'}(x))}; E = \sum_{x \epsilon \Omega} w(x) log(p_{l(x)}(x)) \tag{3.4}$$

The cost function is now computed over a 2-D matrix rather than a 1-D vector as seen in Equation 3.2. The output of the U-Net is of form $K \times W \times H$ for $K \in 1,\ \ldots,\ k$ classes in the dataset, referencing k different segmentation maps in one image. This is a good example about how easily neural networks can be tweaked to perform reasonable well in many different problem fields by adapting the cost function.

U-nets architecture has been enhanced and adapted to fit different problem statements [Islam and Zhang, 2018; Isensee et al., 2021]. Isensee et al. developed the nnU-net, short for "not-new U-Net", which is aimed to provide out-of-the-box implementations for various problem definitions in medical image processing[Isensee et al., 2021]. Image processing is still a field which yields best performances with (highly-)specialized applications and non-experts may be overwhelmed by the terminology and complexity that accompanies training neural networks. Islam and Zhang [2018] used the U-Net architecture for lung segmentation in a transfer learning approach, which was also re-used in this work (Section 4.4.1).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

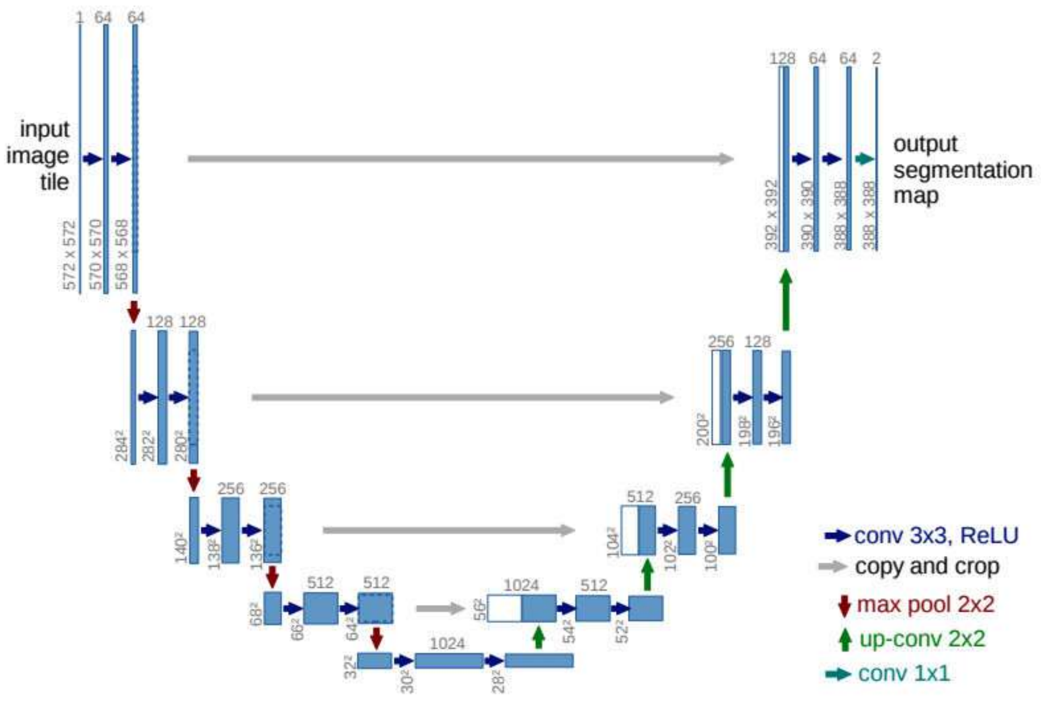


Figure 3.9: Originally proposed U-Net architecture by Ronneberger et al [Ronneberger et al., 2015]. Horizontal grey connections are skip connections enabling faster learning and combination of high and low level features in the expanding part of the network due to copying feature maps from the convolutions in the decoder.

Transfer Learning

When a feed forward neural network is trained, the last layers preceding layers are ultimately trained to create a *latent representation* of the input which is used by the last layers classification step, i.e. a softmax classifier stated earlier. This representation can be used in similar domains with related problem definitions. The process of transferring this representation is called *transfer learning* [Goodfellow et al., 2016, Chapter 15].

In practice transfer learning has become very popular in computer vision. Networks trained on large, open source, object detection datasets like CIFAR-10 or MNIST are shared online and can be re-used and further enhanced (fine-tuned) for other tasks, or just as a starting points with pre-trained model weights rather than random initialization in fresh networks, downloadable online[1] [Paszke et al., 2019] . In this work, transfer learning is used for automated lung segmentation of chest X-ray images. In the particular case of COVID-19, similar approaches have been implemented [Vidal et al., 2021; Zebin and Rezvy, 2021]. Unfortunately, parts of the data have either not been shared [Vidal et al., 2021] or model outcomes differ from this work [Zebin and Rezvy, 2021]. General lung

[1]`https://pytorch.org/vision/master/models.html`

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

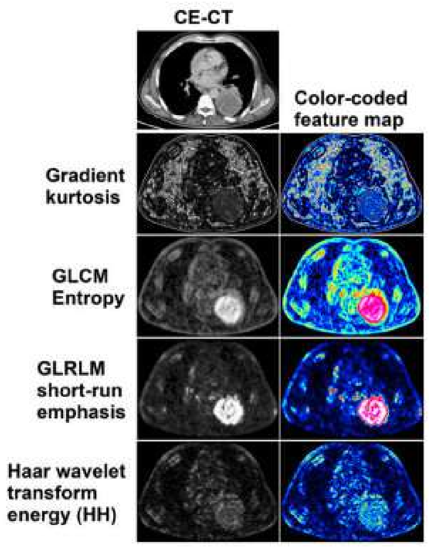


Figure 3.10: Examples of extracted radiomics features for necrotic lung-cancer in CT scan images. Feature-maps have been color coded and used as an overlay for the gray-scale CT scans. Slightly modified version adapted from Mayerhoefer et al. [2020].

segmentation works leaning on transfer learning are available though which are based on open data [Islam and Zhang, 2018]. Islam and Zhang combine two open datasets, the Montgomery County (MC) lung dataset as well as the Shenzhen dataset (SD) [Jaeger et al., 2014]. Additionally, implementations for the approach of Islam and Zhang are available online[2] following the design of Iglovikov and Shvets [2018], making use of a pre-trained VGG network as the starting point.

3.2.4 Radiomics

While the term radiomics is not finally defined, it generally describes the process of trying to extract quantitative features from medical image data in a reproducible way [Mayerhoefer et al., 2020]. *PyRadiomics*[3] is an open-source software library for reproducible feature extraction in medical image data used most prominently in cancer research [van Griethuysen et al., 2017]. It offers image loading, preprocessing and filtering techniques which enables handling a variety of medical image inputs such as CT, PET (Positron Emission Tomography), MRI. Besides that, radiomics implements different feature extraction classes using statistics, traditional feature extraction (Section 3.2.1) to extract common features of medical image data, illustrated in Figure 3.11. Most features are computed based on the Image biomarker Standardisation Initiative (IBSI) feature definitions, which strive to define standardization through implementing feature extraction with baselines and baseline datasets in order to achieve reproducible results [Zwanenburg et al., 2016].

[2]https://github.com/IlliaOvcharenko/lung-segmentation
[3]https://www.radiomics.io/pyradiomics.html

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Feature Class	**Description**	**Nr. of Features**
First Order Features	Describe voxel intensity distributions within masked regions of interest through commonly used metrics. Extracts numerical information about the gray-values in the region of interest (ROI), including Mean, Median or inter-quartile range of pixel or voxel values in the ROI.	19
Shape Descriptors	Available for 2D and 3D shapes. Computing shapes in the original non-filtered images through marching cubes algorithm.	10
Gray-Level Co-Occurrence Matrix (GLCM)	Texture based analysis computed over co-occurring pixel values in the picture over multiple rotations/axis.	24
Gray Level Run Length Matrix (GLRLM)	Texture based analysis computed over run-length of same gray-level picture values in the picture.	16
Gray Level Size Zone Matrix	Rotationally independent version of the GLCM.	16
Neighbouring Gray Tone Difference Matrix (NGTDM)	Quantifies the difference between a pixels or voxels gray value in contrast to their neighbourhood, given a distance σ.	5
Gray Level Dependence Matrix (GLDM)	Computes the dependency of a voxels gray value to other gray values in an area given a distance and threshold value.	14

Table 3.1: Overview of available features in *PyRadiomics*.

Currently implemented radiomics feature classes in *PyRadiomics* [Radiomics, 2022] are presented in Table 3.1.

Figure 3.10 shows a visual representation of some of those extracted features. For now, only gray-scale images are supported. Radiomics can be used interactively through command line, in python or in the application *3D-Slicer* where users are able to segment images on the fly.

In order to derive features from input images, PyRadiomics needs segmentation masks in addition to the original images. Segmentation is thought to have a considerable impact on the performance of radiomics feature extraction [Mayerhoefer et al., 2020]. After extraction, the features can be used to train classification or regression models to predict

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

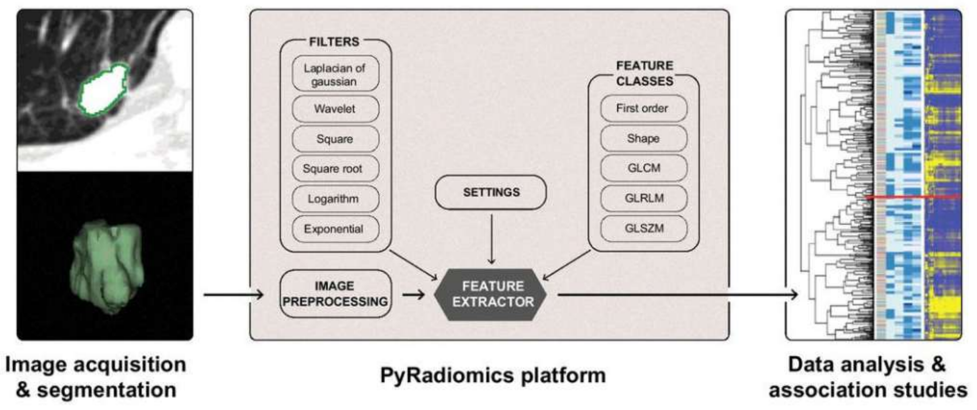


Figure 3.11: Feature extraction with PyRadiomics [van Griethuysen et al., 2017].

disease outcomes or to help in the decision making process for clinicians [Mayerhoefer et al., 2020].

Radiomics features have been used in many publications regarding chest X-ray images as of today [Tamal et al., 2021; Han et al., 2021; Chen et al., 2017]. An open question is still how many features are to be extracted. Tamal et al. [2021] use 71 out of 100 radiomics features in the context of COVID-19 classification in lung X-ray images. Segmentation of X-rays was done manually by radiologists prior to feature extraction. Han et al. [2021] use 102 radiomics features in combination with a CNN to train a contrastive learning algorithm for pneumonia detection. Chen et al. [2017] give an overview of multiple applications developed using radiomics feature extraction in the field of lung cancer detection. Works range from using multiple radiomics runs with different settings, resulting in the number of features typically lying in the hundreds. Redundant features are often removed afterwards to a far lower number, with Pearson-correlation $corr > 0.95$ as a threshold used in some works. This methodology is in line with argumentation by Mayerhoefer et al. [2020], as they point out the risk of overfitting due to the abundance of features, with the number of features that could be extracted being unlimited [Lambin et al., 2017]. This illustrates the idea of applications drifting apart in designs, so a standardization has been given in form of a score by Lambin et al. [2017]. They also stress out the importance of feature reduction (via Principal Component Analysis or clustering) due to the fact that radiomics feature extraction can be used to receive an unlimited amount of features which will most probably results in overfitting prediction models.

3.3 Dealing with Clinical Data

Clinical data for medical analysis and machine learning tasks refer to patient data in tabular form. Due to the high sensitivity of this data special treatment is advised. Thus

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

the data has often been *anonymized* prior to sharing. Often data is *missing*, for which imputation strategies are used. Popular strategies are presented here shortly. Clinical data is typically the quantitative result of some sort of patient examination which can quickly lead to high dimensional datasets (compare subsection 3.2.4). Therefore often *dimensionality reduction* is used to enable quicker computation and easier analysis of medical data. Here, Principal Component Analysis (PCA) and t-distributed Stochastic Neighbourhood Embedding (t-SNE) is presented.

3.3.1 Missing Data Types

Missing data can occur in different forms, with the most used classification given by Little and Rubin [2019] defining three types. Causes for missing data can be broad: from data transmitting errors, generally unavailable data, unwillingness to proceed in a survey, dropouts or death of a patient. The following types definitions are adapted from Van Buuren [2018, chapter 1.2]

Missing Completely at Random

Missing completely at random (MCAR) describes missing data points that do not depend on either observed data or missing data. There is no evidence of a systematic bias in data as it is understood that they are part of a random sub sample of available data. One can think of this case when sampling a random sub sample from an available population. Each member of the population would, in theory, have the same probability of being in the subset. In practice this is often an assumption difficult to fulfill and only used on very big datasets. In the medical context, e.g. when a doctor does not record one specific variable by accident on a random sub sample of patients, this would be MCAR.

Missing at Random

Missing data that can be traced or predicted based on the available data is termed missing at random (MAR). If only members of a population are included in a dataset given a certain known precondition, missing data in that dataset is then assumed to be MAR. This is often the starting assumption. E.g. for a sub sample of patients with a specific disease, like COVID-19, missing data for laboratory variables resulting from thorough blood test might be the result of less severe disease progression.

Missing not at Random

If both MCAR and MAR is not sufficient for the missing data, it is defined as missing not at random (MNAR) or not missing at random (NMAR). MNAR means having no knowledge about why the data is missing. In public opinion research, this occurs when people with weaker opinion respond less often. A prominent fix here is to gather more data in the attempt to be able to reason why it is missing and define it as MAR, making the missing data ignorable for an analysis standpoint again.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Missing Data Implications on Analysis

The types of missing data have implications on how to deal with them, as from a probability theory standpoint, it makes a difference on how different imputation strategies will work with the type of missing data. Simple fixes like imputing with mean values theoretically only make sense when used on MCAR, as we think of the missing data to be part of the distribution that we sampled the available data from. If MCAR is not the case, this would introduce bias ultimately shifting the distribution in the data affecting predictive models built upon it. In general, MCAR and MAR are described as ignorable types of missing data, where analysis can built upon, while MNAR is non-ignorable and needs further clarification on why data is missing.

3.3.2 Imputation

Dealing with missing data is an important prerequisite in machine learning and data analysis. Different methods exist, from removing missing data from the dataset completely, to imputing or interpolating data. The fit of such methods depend on the type of the missing data. In the following we give a glimpse overview of possible methods for data imputation and their effect on different types of missing data.

Deletion

Deletion or list-wise deletion refers to excluding records with missing data. Although it can be thoroughly wasteful in datasets with many missing variables, it is a common approach when data is sufficient especially in the context of big data with millions of records [Van Buuren, 2018, Section 1.3.1]. In the medical context though, data is often sparse and precious, thus deletion might not always be a good option. The option is also deemed not satisfactory for types of missing data which is not classified as MCAR, introducing bias in the data otherwise. A general rule of thumb on when deletion might be appropriate is not supported by the literature [Little and Rubin, 2019].

Mean or Median Imputation

Mean imputation uses the mean of available variables to fill in missing data for that particular variable. As Van Buuren [2018, Section 1.3.3] states, this will underestimate the variance of the variables distribution and as a result will introduce a divergence in the distribution which also affects related attributes. For categorical values one can use the mode instead. It should again be used cautiously as a quick fix in MCAR data with a low number of missing values.

Last Observation Carried Forward

Last Observation Carried Forward (LOCF) is a simple imputation models that repeat the last previously filled value for one missing or multiple missing records. This can be especially handy for longitudinal or temporal datasets. As Van Buuren [2018, Section

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

1.3.6] points out though, previous work has proved that it should be used with caution as it is able to introduce some bias even when data is MCAR and that it should be generally followed by another statistical method that distinguishes between the imputed and the originally present data, which is often not done. In general it should only be used when assumptions for using them are scientifically undermined.

Indicator Imputation

Indicator or Missing-Indicator imputation is describing the case when missing values are replaced by a constant value [Van der Heijden et al., 2006; Van Buuren, 2018]. This can be also be the mean, but it rather applies to imputation with zeroes [Van Buuren, 2018, Section 1.3.7]. It is a popular approach for dealing with health and epidemiological data since records are of utter importance and the method allows to keep all records in the dataset despite there being missing values. It is logical that, similar to mean or median imputation, the method introduces some distributional bias to the data, thus may severely bias predictors, even with MCAR and low numbers of missing data.

Regression Imputation

In regression imputation, a model is trained on the available data that is used to predict imputation values. It yields unbiased estimates for MCAR data as well as for MAR data when the model is trained on influencing variables for the missing data. Yet, Van Buuren [2018] describes it as the most dangerous of all methods, as one wrongly assumes the relations between the missing and explanatory variables may have been a very good fit since the prediction results seem realistic. In fact, regression imputation tends to artificially intensify observed relations in the data by increasing correlations while underestimating variance [Little and Rubin, 2019]

Hot Deck

Hot deck imputation is achieved using records of similar shape as an input for replacing missing values. Sometimes those values are chosen randomly, for instance in the traditional hot deck imputation that was used when data was still stored on card decks and only data that was available on that particular deck was used to impute missing values [Andridge and Little, 2010]. It is commonly used as a robust and fast way to impute data as it has the advantage that it does not need to be modelled in comparison to regression imputation. But it is making implicit assumptions over the data the imputation is computed through distance metrics. The method proved to produce consistent estimates when data is MCAR, while for MAR it only holds for unbiased under certain conditions. [Andridge and Little, 2010; Little and Rubin, 2019]

k-Nearest Neighbour

k-Nearest Neighbour (kNN) imputation is a special form of hot deck imputation. It takes the k most similar records into account, creating an average for available variables for

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

imputation. The method is also referred to as *predictive mean matching* [Van Buuren, 2018, Section 3.4]. Similarity is measured in form of a distance measures, by default the euclidean distance. The averaging can be done in a weighted approach, by determining the respective values importance by the distance to the record that will be imputed on. In the literature, nearest neighbour imputation is arguably one of the most robust method to use when imputing data, resulting in unbiased values proven in different experiments for MCAR and MAR data [Jerez et al., 2010; Van Buuren, 2018; Andridge and Little, 2010]. Usage becomes dangerous for small sample sizes though and limitations are limited to data values occurring in the dataset, with the method being unable to interpolate beyond them by design.

Imputing Categorical Variables

Imputing categorical values becomes a more difficult task in comparison to numerical values. Van Buuren [2018, Section 3.6] propose the use of multiple imputation, for example by using logistic regression to train a imputation model based on the available data. This is only feasible though when there is a high number data available. As a rule of thumb they define ten records per feature. If the feature is non-binary and data has to be one-hot encoded in order to follow a numerical representation, the available data multiplies to ten times the number of unique values. Given that data might not be distributed uniformly, this might introduce problems in datasets with a general low number of records. In those cases Van Buuren [2018] proposes changing to more robust methods like kNN or random-forest based estimators for imputation.

One additional effect when imputing categorical values, for example in binarized representations, is that one can end up with non-realistic values. When using kNN imputation you may end up with values in between the range $[0, 1]$ for a binary case. For example if a patient had a symptom or not, encoded as 0 for not having the symptom and 1 suffering from it, kNN could come up with a value of 0.4 based on the nearest neighbours and the weighting function (uniform, distance based). For cases like this, Ake [2005] weighs out the pros and cons for the option of rounding values back to a realistic state. He argues that in the earlier days, when data was shared via devices for analysis to a different user group, one might indeed round the values to make the data look quite realistic and reduce the probability for questions from a succeeding user group in the data analysis workflow. In the case of providing the data it might still be a good idea to round those values, but nowadays, when imputing the data, you follow up with an analysis task yourself, which then makes interpretation of the imputed data non-necessary.

Univariate & Multivariate Missing data

Previous methods, with the exception of kNN imputation, are methods that focus on univariate missing data imputation. Univariate missing data are defined as datasets that have missing data in only one feature [Van Buuren, 2018, Chapter 3]. In practice though, missing data often occurs in several variables, sometimes at once [Van Buuren, 2018, Chapter 4].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Multivariate missing data introduces several additional issues [Van Buuren, 2018, Section 4.2]. Missing data variables can have co-dependencies, where multiple missing variables are highly correlated with one another. Often, data types are different. Some might be categorical, others numerical or continuous, while others might be ordinal. There might be some temporal structure through pre-existing ordering in the data that could be important when dealing with the missing data. Some missing data might be introduced due to censoring or data anonymisation, given an underlying connection which needs to be protected between data variables. Multivariate imputation methods dealing with this type of data are able to create unrealistic values, for instance negative values for non-negative scales (weight).

Proposed literature shows that in case of multivariate data, best performances are achieved when using more sophisticated machine learning strategies Yenduri and Iyengar [2007] including kNN. Van Buuren [2018, Chapter 4] propose iterative imputation strategies. In general multivariate imputation is not straightforward and method applicability does depend on missing data patterns and data availability Van Buuren [2018, Chapter 4].

Summary

Imputation of missing data is a broad topic. Van Buuren [2018, Section 1.1] comes to the conclusion that the problem is often not thoroughly tackled in research and thus modern methods like multiple imputation are not fully utilized. Additionally missing data is often not communicated well because historically missing data was deemed as a sign of a weak study, leading to downplaying of the topic in multiple medical contexts. Apart from the mentioned methods, additional and more complex statistical models have been proposed in the literature, namely maximum likelihood or pattern mixture models which are also able to deal with MNAR data more robustly, but are not presented here in detail. Methods for statistical comparison of imputation models are available as well, and different approaches have been evaluated in experiments.

Yenduri and Iyengar [2007] experiment with six real-world datasets and 13 imputation strategies, including mean imputation, listwise deletion, ten different settings for hot-deck imputation and maximum likelihood estimation. They conclude that listwise deletion performs worst when the data is not MCAR and thus should be avoided. Best performing methods include hot-deck imputation with more robust results using Manhattan distance on datasets with MAR, amidst maximum likelihood approaches.

Jerez et al. [2010] compare different imputation models on breast cancer data, including incomplete multivariate data (several missing variables at once). In addition to the presented methods above they trained a two-layer MLP and Self-organisation map and compared the differences for prediction accuracy of a cancer survival prognosis model. They conclude that, with the exception of hot-deck imputation, the methods all increased prediction accuracy. The highest increase was observed in MLP and kNN imputation.

Since the aim of this work is not focused on creating the best possible representation of missing data but rather to acknowledge recent work in the field, those are not specified

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

further as well. The performance on the missing data will be chosen by the best performing one over the metrics of the models predicting the actual target defined later, and multiple imputation methods will be available for comparison for the respective interested user groups.

3.3.3 Dimensionality Reduction

Working with high dimensional, multivariate datasets as in this thesis introduces several challenges for training models and data analysis. Dealing with high dimensional data has even been termed as the *curse of dimensionality*, where analysts keep struggling with highly dimensional data due to data becoming exponentially more sparse the higher the dimensions get [Steinbach et al., 2004]. When clustering data, most methods rely on distance or similarity measures which can get hindered by higher dimensional datasets. Moreover, training models on high dimensional data can be complex. While models handle dimensionality differently, having multiple highly correlated features can be a computational waste while hindering performance and real-time visualization techniques [Munzner, 2009]. Additionally, high number of variables makes data visualization especially hard. In order to enable users to get an overview of data, it has to be transformed to a representation which can be understood in a common way by reducing dimensions to a maximum of three in order to use non-complex 2D or 3D representations and embeddings. By reducing dimensions, presentation performance can also be increased. Reducing dimension can be done using feature selection, for instance by removing features that have high correlation values with other features, or are a-priori known to have no impact on the further analysis. In addition, dimensions can be reduced by projecting features from higher dimension space to lower space by aggregating *similar* features with the goal to loose as less information possible. Here, two approaches are presented, namely principal component analysis (PCA) and t-SNE.

Principal Component Analysis

Principal Component Analysis (PCA) is a non-parametric method that enables projecting data onto a lower dimensional data space [Shlens, 2014]. PCA is using singular value decomposition (SVD) to project the original data to a new dimensional space, which is created by maximizing the variance of the data projection capability of the new dimensions. The generated dimensions are expressed as linear combinations of the original feature space and are normalized, sorted descending by their eigenvalues. Those linear combinations are called Principal Components. Earlier principal components are by definition able to inherit a higher variance of the feature space, making them more important than latter components. The principal components are found using the following procedure:

i) First the data is centered around the origin. The first principal component is then found by creating a line which is forced to go through the origin and gets rotated so that the variance in X is maximized. This is done by maximizing the sum of

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

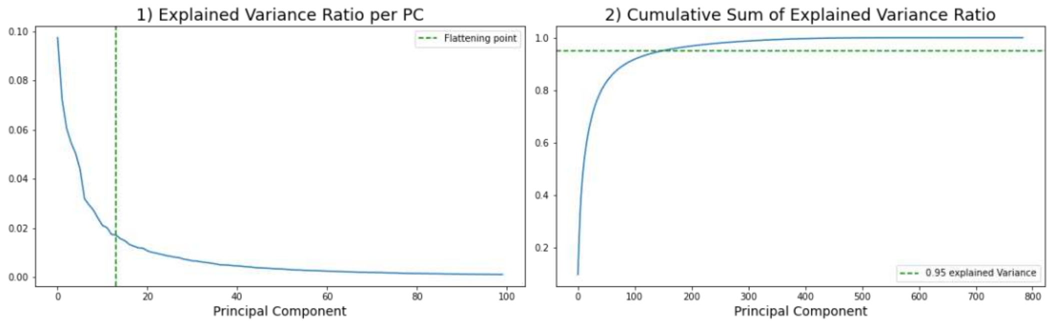


Figure 3.12: Illustration of Principal Component Analysis for the MNIST hand-written digits dataset in a scree-plot. The left plot shows the explained variance for the first 100 principal components, showing a massive decrease after the first 20 features in explained variance. When the line is flattening out one might decide to stop and use those number of components. This is a scree-plot. The right plot shows the cumulative sum of explained variance ratios. One common (and the default in *scikit-learn*) is to use 95% of explained ratio as a stopping criterion for the number of selected components. In this example, given the maximum is the number of dimensions is 784, representing one dimension for each pixel, 95% of the variance would be achieved after 150 principal components.

squared distances from all points to the line. The vector that represents the lines slope is normalized and by estimating the linear combinations weights (SVD) the features importances for that vectors direction can be expressed.

ii) next, another principal component is found which is again, going through the origin and is orthogonal to the previous principal component(s). Again, it is rotated so that the variance is maximized.

iii) this is continued until the previously defined number of PCs are found, a variance threshold is met or the number of principal components is equal to the variables in the dataset.

In practice, many stopping criteria have been proposed and are presented by Brown [Brown, 2009]. Most prominent is a scree plot, indicating the most important PCs by plotting the variances. Stopping at a certain number of cumulative explained variance is another popular option, for instance when components are able to express ≈ 90% of the variance. Both of those options are displayed in Figure 3.12 for a 10% subsample of the MNIST dataset[4].

t-SNE

t-SNE is another dimensionality reduction approach focusing on visualizing high dimensional data in a 2D or 3D space [Van der Maaten and Hinton, 2008]. It uses the

[4]http://yann.lecun.com/exdb/mnist/

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Stochastic Neighbourhood Embedding (SNE) algorithm [Hinton and Roweis, 2002]. The idea in SNE is that the dataset is represented as a similarity matrix, inheriting pairwise similarity scores of the data records. A low-dimensional matrix is learnt that represents the high-dimensional similarity matrix. The idea here is that points from the same cluster/region will have higher similarity scores. Using stochastic gradient descent with momentum, the low-dimensional matrix is learnt iteratively. The neighbourhood of the high-dimensional matrix is represented as conditional probabilities from a Gaussian distribution with variance σ that depends on a hyper-parameter *perplexity*. Points that are near to each other in the original space will have a higher probability of being a neighbour. In the iterative process, when optimizing the matrix, points are moved towards points that are more likely to be their neighbour, making the low-dimensional matrix more similar to the higher-dimensional one. New gradients computed during SGD can be thought of springs that pull similar points near to each other while pushing non-similar ones from each other.
Now what t-SNE effectively does different to SNE is using a t-distribution rather than a Gaussian which was proposed in the original SNE to overcome the crowding problem, and additionally tweaking the cost function of the original SNE. Probabilities also get scaled in order to make up for for differences in cluster densities.

A *good* value for perplexity is not learned per se, as authors propose to use different values for visualization to get a feeling on how the reduction performs [Hinton and Roweis, 2002; Wattenberg et al., 2016]. It is also worth noting that due to the SGD optimization and its random initialization of gradients, consecutive runs on the same dataset with exact same settings may yield different outputs, even if seeds are provided [Wattenberg et al., 2016]. In practice t-SNE is slow on data having a very high number of dimensions. This is often mitigated by using PCA prior to t-SNE.

Figure 3.13 shows a comparison between PCA and t-SNE in use on a 10% sample of the MNIST dataset, which is a collection of hand-written digits in 28×28 pixel representation, resulting in a 784-dimensional feature space. While in the space of the two most-explaining principal components, a high overlap between the numbers is still visible, t-SNE is doing a much better job at distributing the classes.

3.4 Data Science Foundations

This section briefly present data science foundations including data splitting for hyper parameter tuning as well as effective use of the data. Moreover, metrics and predictors used in the upcoming sections throughout the work are introduced shortly.

3.4.1 Data Splitting

When training a machine learning model to predict the outcome for a given set of input variables, the ultimate goal besides achieving high accuracy in prediction is to create a model that performs well on new, unseen data. This objective is termed generalization.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

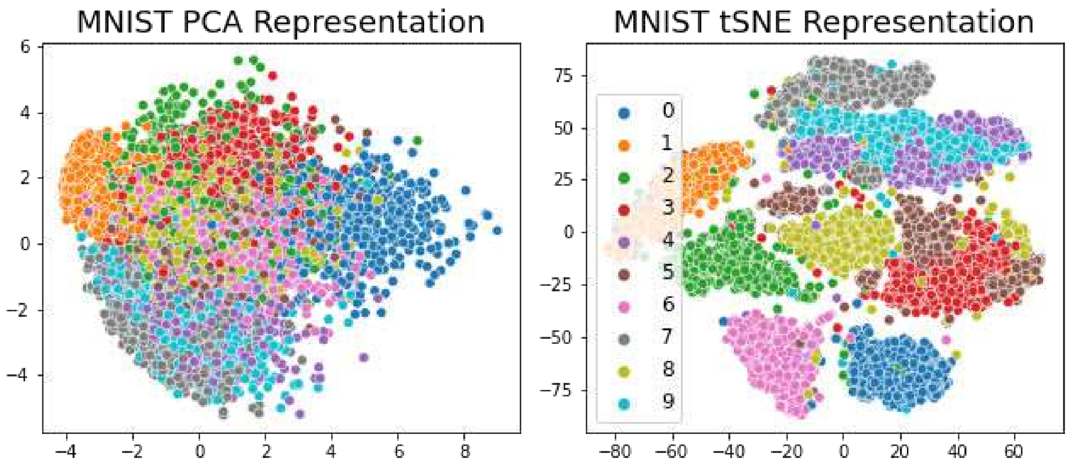


Figure 3.13: PCA and t-SNE representation of a 10% sample of the MNIST dataset. PCA is successfully distributing numbers that differ from one another, for example zeroes and ones which should be clearly distributable in pixel-space as well. It does struggle though with numbers that are more similar, hence eight and three. T-SNE is doing a much better job in the case of MNIST, being able to more clearly separate the classes, with similar ones being near each other or having overlapping boundaries, for example four and nine.

Models that do not generalize well are models that overfit. Overfitting occurs when models rather memorize training data instead of learning relationships between the input and respective outputs, possibly perfectly embodying their output representation but yielding low scores on unseen data. The trade-off between those goals is often referenced as the bias and variance dilemma [Reitermanova et al., 2010; Hastie et al., 2009; Kohavi et al., 1995; Grus, 2019, Chapter 11]

In the effort of detecting and avoiding overfitting a dataset D is randomly split into different mutually disjoint subsets. Most commonly, it is split into three parts, training X, test T and validation data V, where the size n of the splits is $n_X > n_V \geq n_T$. In this setting, the training subset is used to train different model families with different hyper-parameter settings which are tested against the test dataset. Common ways of dividing the The validation dataset is never used during tuning and training of the model. It is only used at the end of the process as a final estimator, testing generalization ability of the final precition model. This procedure is also called hold-out cross validation model. Often, the terms test and validation are exchanged, so that the test data is hold out as a final estimator, while the validation data is used consecutively to estimate best performing settings. In the context of this work, the first terminology is used. Hastie et al. [2009, Section 7.2] loosely propose split sizes of 0.5/0.25/0.25 for a train, test, validation split, which is flexible and should be chosen with the dataset in mind.

Due to the nature of splitting data only once, data variability is limited during the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

evaluation as the same training and test/validation set is used over all runs. K-Fold cross validation improves the process in that regard. In k-Fold cross validation, data is split into k equally sized portions. For each model evaluated, all splits are used once as a testing dataset now. This improves variability since all records are now used at least once as training and testing, so records with outliers are now guaranteed to be included in both parts of the process. In addition to that, a third exclusive validation dataset is not needed anymore, at the cost of increasing the computational effort, since each model is now trained k-times which can be problematic when dealing with large datasets. In the most extreme case, k is set to the number of records in the dataset n, which is then called Leave-One-Out Cross-Validation (LOOCV). The test set is consisting of only one record while remaining records are used for training. Metrics are finally averaged over k-folds in order to estimate the best performing model and respective hyper-parameters. Common values for k are 5 or 10 depending on the dataset [Hastie et al., 2009, Chapter 7.10].

While most often splitting is done randomly, in some cases this will introduce biases towards majority classes, or at worst inability to train for less represented classes. In this case, a stratified split is preferred, keeping class distributions equal among different splits.

3.4.2 Model Performance Evaluation

There is a variety of possible methods to evaluate the performance for models trained to solve different problems. In the following, metrics for image segmentation with and without ground truth are proposed. Additionally clustering metrics for unsupervised clustering are presented. Finally, we provide an overview of classification metrics.

Segmentation Metrics

Quantifying segmentation performance in a binary optimization setting is often done using similarity metrics. Most popular metrics used are *Dice* or *Jaccard* similarity in order to compare automatic segmentation to ground truth [Bertels et al., 2019]. Dice and Jaccard are defined as shown in Equation 3.5 and Equation 3.6.

$$Dice(y, \hat{y}) := \frac{2|y \cap \hat{y}|}{|y| + |\hat{y}|} \tag{3.5}$$

$$Jaccard(y, \hat{y}) := \frac{|y \cup \hat{y}|}{|y \cap \hat{y}|} \tag{3.6}$$

where y represents the GT and $\hat{y}$ the prediction. As long as the ground truth y is available similarity metrics can be used as performance metrics for classifiers. But in medical image segmentation, the ground truth is often not available. Therefore, other approaches are presented as well. Qian et al. compare segmentation evaluation without the ground truth to unsupervised clustering and thus propose to use evaluation techniques proposed

TU WIEN Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

in that domain [Huang and Dom, 1995]. They do not really go into detail on how they would approach this in a gray-scale image domain where pixel-domain differences are more scarce given the lower feature dimensions. Using color histograms for instance would not make sense to cluster pixels into different regions as they would be too weak to distinguish regions of interest from one another. In general, limitations presented in subsection 3.2.1 are applicable as well here.

Correia and Pereira [2002] propose metrics for evaluation of segmentation in video, focusing either on metrics per object or the whole scene. This is done by estimating how many objects are detected and how many are expected in a scene, averaging final metrics over multiple frames. In the context of this thesis, one would expect to receive two objects from a segmentation. But it would not make a big difference if we would have more, given that the artifacts are cleaned in post-processing by removing smaller areas in the picture. So a segmentation yielding many small artifacts around the edges of the image will be removed and should not be penalized if the key aspect of the segmentation is reasonably better (compare Figure 4.15 in Section 4.4.2).

Kohlberger et al. [2012] use 42 low-level image features extracted from the segmentation mask and source image to predict errors. The proposed method assumes that for a given dataset at least some GT segmentation is available. For those images, automatic segmentation is used and a classifier is trained to predict the error between the automatic segmentation and the actual ground truth segmentation. Subsequent observations without GT can be segmented and the error can be predicted on new, unseen data. This has also been implemented by Magg et al. [2021]. Nevertheless, the problem with this approach is still that the authors defined the problem when evaluating segmentation for a dataset in which one has at least some ground truth available that can be used for feature extraction and a classifier can be trained on in the first place. This is not the case for the dataset used in this thesis though. One possible way mitigate this issue would be to train the error predictor on available data used to learn the transfer model, where GT is available. The problem is though, that low-level feature extraction methods may be not very robust towards the differences in the datasets. As it is shown in Figure 2.2, the images may vary by a large margin, including artifacts, different angles and views and deviating in contrast settings.

Valindria et al. [2017] proposed a different solution for the problem which they termed *Reverse Accuracy*. After training a segmentation model, the model is used on unseen data. The outcome of this segmentation is then used as the input to train a new model, which is evaluated on the original training images for which GT is available. Using obtained metrics, segmentation on new, previously unseen images can now be flagged as potentially good or poor. This approach could be used as in a slightly modified way as a *reverse transfer learning* approach and would be feasible to use in this thesis given the data.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Performance Metrics for Unlabeled Data Clustering

When true labels are unknown and data characteristics that would suggest natural clusters are not known a-priori, natural choices for clustering metrics are insufficient. Here, two universal metrics are introduced to estimate clustering performance without knowledge of the ground truth labels by focusing on the fitted clustering model itself.

Silhouette Score: The Silhouette score [Rousseeuw, 1987] is defined for each observation in the dataset as the mean distance of the observation to all observations from the same cluster a and the mean distance of the observation to all observations from the nearest different cluster b. Formally it is defined as

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{3.7}$$

The overall score is then averaged over all records in the dataset. It ranges between $[-1, 1]$, with higher values indicating a perfect clustering, values near zero a undecided clustering, while negative scores would indicate wrong clustering based on distance metrics.

Calinski-Harabasz Index: The Calinksi-Harabasz Index (CH) [Caliński and JA, 1974] is a measure for clustering performance also known as Variance Ratio Criterion. It describes the ratio between the sum of between-clusters dispersions and the within-cluster dispersion for all clusters. Higher scores are awarded for dense and well separated clusters, but in general the index can be continuously rising or falling for increasing number of clusters. In that case it is likely that ideal clustering is not possible. Cluster scores should thus be plotted and read like an inverse scree plot, not looking for the biggest decrease but peak in CHI. The CHI s is mathematically defined as

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n-k}{k-1} \tag{3.8}$$

with $tr(B_k)$ defining the between group dispersion matrix, $tr(W_K)$ within-cluster dispersion, k number of clusters over a dataset with n observations. Dispersions are calculated with cluster centroids distance to cluster points.

Prediction Metrics

Next, prediction metrics used to evaluate model fitness are briefly introduced, with focus on classification metrics only. Most prominent metrics are calculated in respect to the confusion matrix, as depicted in Figure 3.14, which sets predictions and true labels into perspective. The default is a binary classification confusion matrix as seen here, but it can also be used for multi-label classification settings. Among many others, the following metrics derived from the confusion matrix are defined [Davis and Goadrich, 2006]:

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

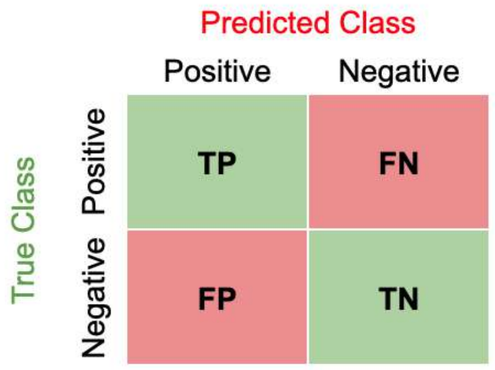


Figure 3.14: Typical illustration of a confusion matrix in a binary classification setting.

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN}$$
$$Accuracy = \frac{TP+FN}{TP+TN+FP+FN} \qquad F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

where TP, FN, FP, TN = *True Positive*, *False Negative*, *False Positive* and *True Negative* as defined by the confusion matrix in Figure 3.14.

The right choice of metrics for model evaluation really depends on the classification use-case. For medical disease prediction the metric should be including false-negatives and false positives, as individuals that are not infected with a disease for instance should not be given unnecessary treatment [Lever, 2016]. With that in mind, *Precision* and *Accuracy* are especially not that adept at doing so, with Precision not even bothering about false predictions per definition. In general there is a trade-off between *Precision* and *Recall* (or *Sensitivity*) in classification, as one ultimately is trying to optimize one of the two. When a predictor is optimized towards reducing FN (increasing the recall) it will most likely increase FP (reducing the precision). This is why aggregate metrics like *F1*, defined as the harmonic mean between the Precision and Recall (Specificity) are quite handy. Furthermore, another common metric in the medical domain is Specificity, or the true negative rate which is defined as

$$Specificity = \frac{TP}{P} = \frac{TP}{TP+FN} \tag{3.9}$$

More abstract derivations of the metrical score are Receiver Operator Characteristic (ROC) curve [Provost et al., 1998] or Precision-Recall (PR) curve [Schütze et al., 2008]. Their goal is to give a visual overview of the model performances. ROC has been proposed in order to increase transparency in binary classification by plotting the number

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

of incorrectly classified negative examples in contrast to correctly positive examples. This is plotted as a curve and the metric that is used in ROC is the area under that curve (AUC). The PR curve has a similar representation form which plots the precision and recall rates. As Davis and Goadrich [2006] point out, these curves are not redundant but complementary as they both give singular perspectives into the performance of a classifier.

In multi-class classification scenarios, most metrics can be used with several tweaks. For ROC, one-vs-one (OvO) or one-vs-rest (OvR) can be used to convert the multi-class problem into several binary problems [Bishop and Nasrabadi, 2006, Section 7.1.3]. In OvO, pairwise combinations of the classes are created and the metric is are averaged over all classes (uniformly weighted by default). In OvR, each class is predictions are run against all other classes.
Precision, Recall, Accuracy and F1 are altered similarly. Multi-class gets treated as several binary problems by averaging the original binary metrics. *Macro* averaging computes the mean of all binary metrics with equal weights. Often though in multi-label classification, there are large majority classes that make this approach infeasible. *Micro* averaging is averaging the score on a per sample-class pair basis, handing less importance to majority classes and more to low-presented ones.

3.4.3 Working with Imbalanced Data

In this work prediction tasks must be aligned with the general problem of *unbalanced data* [Barandela et al., 2003]. The term unbalanced is used for datasets that tend to have a highly skewed distribution of prediction targets. This is often the case in applications where a classifier is to detect a *rare* but *important* case, for instance in fraud detection, e-mail spam classification, or medical settings as in this thesis. In general, Barandela et al. [2003] point out that the majority of applications dealing with imbalanced data are solving this issue with one of three possible mitigations: assign distinct costs to classification errors; resample the original data by oversampling the minority class, or undersampling of the majority class; compensating for class imbalance by internally biasing the discrimination-based process. In this work, we adopted the first two approaches, namely by implementing balancing as well as different over- and undersampling strategies. Discrimination-based processes are often implemented by using auto-encoder neural networks to train a classifier focused on detecting outliers in a system with many observations [Eavis and Japkowicz, 2000; Li et al., 2020b], where Li et al. [2020b] propose such a system for COVID-19 detection with promising results. Due to the scarceness of our data in terms of pure observations, we ultimately lack the amount of data required to successfully train such a model and thus focus on the other approaches.

Balanced Weights

When training a classifier in general, a cost function is optimized. When doing so in a balanced dataset by default, each observation is assigned the same weight in the training phase. In imbalanced datasets or when dealing with outliers, one can manually adjust

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

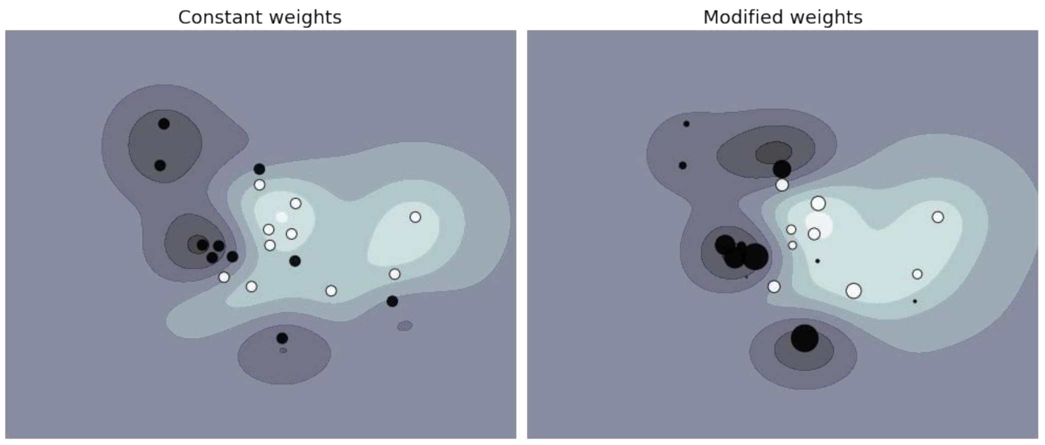


Figure 3.15: Balanced weights effect displayed with a SVM classifier and a binary toy dataset. Larger points depict higher weighted observations, changing the decision boundaries of the classifier.

those weights in order put more emphasis on the classifier on getting these higher weighted observations predicted correctly. In multi-class imbalanced datasets, most often the sheer number of observations is used to determine the weight applied to each observation. A visual representation of this effect is shown in Figure 3.15 using a binary toy dataset. One can see that higher weighted observations, encoded by point size, are affecting the decision boundaries of the classifier towards those observations.

Undersampling

Undersampling describes the effort to choose a smaller sample of observations from the majority classes in an effort to balance out prediction targets, ultimately creating a balanced subset of the available dataset. This is often done randomly but more sophisticated methods have been proposed to limit the amount of information lost due to dropping a potentially high number of samples [Barandela et al., 2003]. This method can often lead to harshly reduced datasets, especially when dealing with multi-label classification datasets as the class with the lowest occurrences dictates the overall amount of data used to train a classifier, potentially leading to a huge loss of information that would be available.

Oversampling

In contrast to undersampling, oversampling replicates observations from minority classes by replicating them [Barandela et al., 2003]. This approach does not add more information to the classification system as only already known combinations of features are replicated. Classifiers that are sensitive to high-volume datasets may be affected negatively in terms

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

of training and classification time by adding more observations, but this can be neglected here.

SMOTE

The Synthetic Minority Over-sampling Technique (SMOTE) is an extension to random oversampling Chawla et al. [2002]. Each observation of the minority class that is to be oversampled is used to create synthetic examples along the line segments joining a subset of nearest neighbors. Most often a form of k-neighbors with $k = 5$ is used to reduce the computational effort. For each synthetic observation's features, a value is chosen using a random weight between the lines of a chosen observation and its nearest neighbor (s).

While again, imbalanced learning and efforts to overcome problems when dealing with the classification of unbalanced datasets is a wide field, we merely scratch the surface here and follow best practices without focusing too much on the most sophisticated methods available.

3.4.4 Clustering Methods

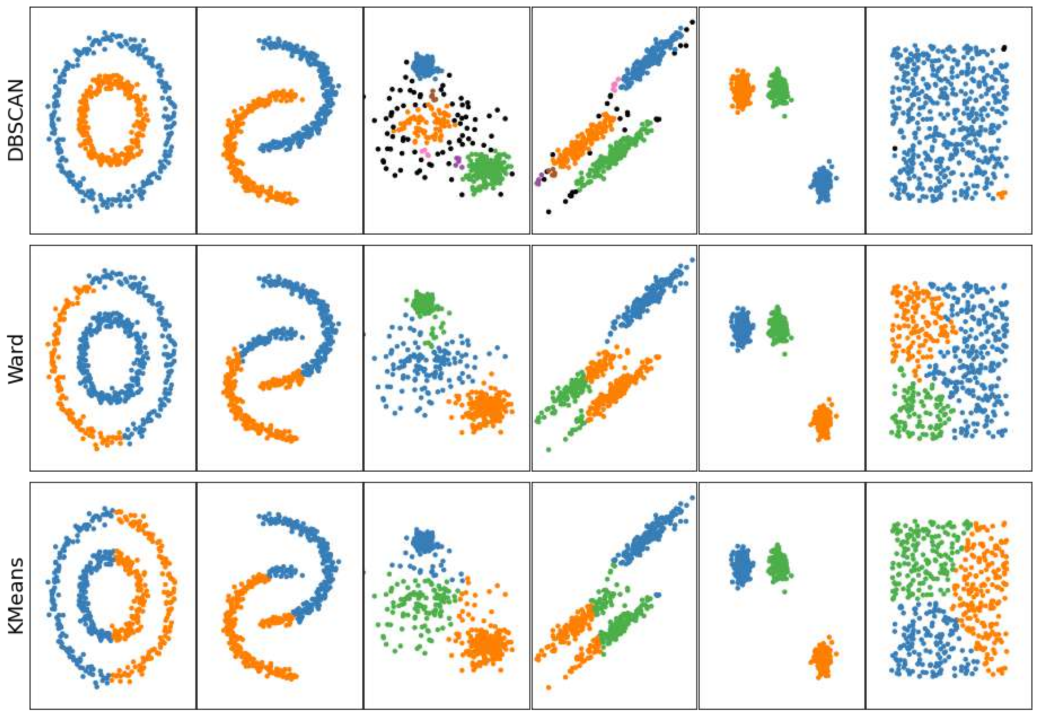


Figure 3.16: Clustering methods compared on toy datasets, illustrating different clustering results for each respective method. Figure adapted from *scikit-learn* docs[5].

[5]https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Clustering describes the attempt of clustering or grouping observations into groups of observations. This is often done in the context of *unsupervised learning* where no ground truth is available or when one wants to group observations into clusters of observations that are similar [Rokach and Maimon, 2005]. The similarity of instances is often described as a distance metric. Additionally, Goodfellow et al. [2016, Section 5.8] describes clustering as an attempt to create a representation that preserves as much information about a dataset X as possible while keeping the representation *simpler* or more *accessible* than X itself. Clustering is an important part of multiple scientific disciplines where clusters represent different typologies of entities.

Different clustering methods can be grouped into different categories based on their approaches [Rokach and Maimon, 2005]: *Hierarchical methods*, which connect instances in either top-down or bottom-up fashion; *partitioning methods*, which relocate cluster instances from on to another given some convergence function, starting with an initial partitioning; and *density based* approaches, assuming points belonging to one cluster are drawn from a specific probability distribution. For each of those categories, one method is used in this thesis and explained here in more detail. Further methods exist but are not used in this thesis to limit complexity.

Figure 3.16 shows the results of chosen methods on toy datasets, pointing out different strengths and weaknesses of the respective algorithm.

K-Means

K-means is a simple representation learning algorithm that divides a training set x into k different clusters [Goodfellow et al., 2016, Section 5.8.2]. It is a partitioning method. The k-means algorithm starts by, often randomly, generating k-cluster centroids and alternates between two steps until convergence:

1. label each observation as cluster i, where i is the cluster of the nearest centroid, with $i = k$.
2. each centroid is updated to the mean of all training examples x_i assigned to cluster i.

This process is repeated and measured by decreasing an error function. Most often the sum of squared distances is used [Rokach and Maimon, 2005].

K-means clustering is one of the most popular clustering methods due to its linear complexity even with a high number of observations [Rokach and Maimon, 2005]. Shortcomings of the algorithm are the initial cluster centroid creation, which may have an impact on the final results, as well as the algorithm's sensitivity towards noisy data and outliers. Hence the algorithm being most dependent on distance metrics, data must be of scaled numerical type. Optimization towards cluster initiation most prominently has been the proposed weighted approach by Arthur and Vassilvitskii [2006] called k-means++.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Hierarchical Clustering

Hierarchical clustering is either done using a top-down or bottom-up approach. This describes the starting point of the clustering. In agglomerative clustering, each object initially represents one cluster. Those clusters are then merged until the desired number of clusters may be obtained. In divisive hierarchical clustering, all observations are in one cluster, which is divided until the desired number is reached. Merging or division is performed using similarity or distance metrics which optimize some penalty criterion, for instance, the sum of squared distances. A number of hierarchical clustering methods are proposed that use different metrics [Rokach and Maimon, 2005].

In this work, we focus on Ward-hierarchical clustering which is considered as an average-linking clustering strategy — or *minimum variance* method. It considers the distance between two clusters to be equal to the average distance of points in the respective clusters. When compared to single-linkage or complete-linkage methods, which define cluster distances by the nearest or farthest neighbors of respective clusters, average-linking provides a more robust solution by taking all points into account [Yim and Ramdeen, 2015].

Density-Based Spatial Clustering of Applications with Noise

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based method that assumes the data is drawn from a mixture of several probability distributions [Rokach and Maimon, 2005]. Clusters are formed by growing them until the density of observations in a given radius, or neighborhood, exceeds a certain threshold. DBSCAN enables the clustering of arbitrary shapes in an efficient way for large spatial databases.

In contrast to hierarchical and the k-means clustering method, DBSCAN uses two variables to cluster available data which does not restrict results to a pre-defined number of clusters. The algorithm is illustrated by Schubert et al. [2017]. DBSCAN uses a minimum number of points as well as a radius ϵ to analyze each point in a dataset. Each point is considered a *core point* if it has a minimum of n neighbors in the radius ϵ. This can be defined as a minimum density area that forms a cloud of points, which can be considered a cluster. Each core point discovered is (recursively) joined by its neighbors to form a cluster. There are also tuned implementations with only affect core points themselves without recursively iterating through its neighbors, leading to improved runtime. Non-core points that are reachable by other core points through this transitive approach are defined as border points. Points that are not reachable are defined as outliers or noise and are not considered to be part of any cluster but get assigned to a cluster that is solely meant to inherit outliers.

3.4.5 Prediction Models

Supervised prediction models applied later on in this thesis are shortly described here. In general, the idea was to use models of different families. Therefore, Random Forest,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Logistic Regression, Support Vector Classifier, MLP and XGBoost were used in order to have a comparison of different model types and their ability to fit on the dataset. The multi-layer Perceptron (MLP) was already covered in Section 3.2.2 and is not mentioned here again.

Random Forest

Random forest was introduced by Breiman [2001]. Random forest utilizes the training of many decision tree classifiers in different settings. Many decision trees build a *random forest* of trees which are used in an ensemble classifier where multiple classifiers vote for a final classification label. This is done in an effort to combat overfitting which is a major problem in decision trees.

Random forest starts with *bootstrapping* the original dataset by generating a random subset of size $n = N$, with N being the number of observations of the original dataset, by sampling with replacement. The number of distinct observations in the bootstrap dataset is denoted as n_1. The remaining observations $n_2 = N - n_1$ form the out of bag dataset. Now when training the decision trees, only the bootstrap sample is used to fit the tree first. Additionally, in each leaf, only a sample of all available p variables is considered for each leaf split. After fitting the tree in this fashion, the out-of-bag data is used to compute the tree impurity and feature importances for each variable p. In classification settings, a majority decision is generated by using each trained tree to classify new observations.

Support Vector Classifier

The Support Vector Classifier (SVC) terms the usage of Support Vector Machines (SVM) for classification purposes. There is also a regression solution that is called Support Vector Regression (SVR). Support vector machines in combination with traditional feature extraction have been used heavily prior to deep learning [Goodfellow et al., 2016, Section 5.7.2].

SVC first builds upon the assumption that a given problem is linearly separable. By that, the algorithm aims to find an optimal, linear decision boundary by maximizing the margin, where the margin is defined as the perpendicular distance between the decision boundary and the closest data points. Those points are then called the support vectors and observations not defined as support vectors may be moved, added or removed freely in any dimension without changing the decision boundaries as long as they are not moved inside the margins [Bishop and Nasrabadi, 2006]. The main key innovation used by SVM is the use of the so-called kernel trick [Goodfellow et al., 2016, 5.7.2]. The linear decision function can be re-written as

$$w^T x + b = b + \sum_{i=1}^{m} \alpha x^T x^{(i)} \tag{3.10}$$

TU Bibliothek WIEN Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

with $x(i)$ being training examples and α a vector of coefficients. Now, x can be replaced by the output of a kernel function $\phi(x)$, with only α being optimized [Goodfellow et al., 2016, 5.7.2]. This kernel representation now allows for non-linear decision making, in other words the original feature space is transformed to a new transformed space in which a classifier is trained. Common kernels are Sigmoid, radial basis function (RBF) or Gaussian.

Now in practice data is not always linearly separable. As a fix for that the margin is *softened*, meaning it allows miss-classifications to happen by penalizing them to some degree using a *slack* variable ξ, leading to a cost function as followed:

$$C \sum_{n=1}^{N} \xi_n + \frac{1}{2}||w||^2 \tag{3.11}$$

The parameter $C > 0$ controls the slack variables penalty and the margin, with miss-classifications on the wrong side of the decision boundary resulting in $\xi > 1$, correctly classified observations in $\xi = 0$, with $C \rightarrow \inf$ penalizing much, leading to a more simple decision boundary.

Multi Class Problem: By default, the SVM is always a binary classifier. In more complex problems, like in this thesis where multiple prediction targets exist, two common approaches are available [Bishop and Nasrabadi, 2006, Section 7.1.3]. One way is to train $K = C$ classifiers, where C defines the number of classes available. For every classifier trained, observations not labeled as the current class are considered negative examples. This is often referred to as *one-vs-rest* (OvR). The downside of this approach is that each classifier is now trained on different tasks, given the combination of different labels as the negative ones. This also leads to the introduction of imbalance in many cases, even when the data is balanced. For example, when 100 training samples are available for ten classes, every classifier has to deal with a 9:1 imbalance now. Another idea would be to train binary classifiers for each pair of classes available and combine those via a majority vote. This is called *one-vs-one* (OvO). This leads to significantly more computational effort, as instead of K, now $K(K-1)/2$ classifiers are trained. Nonetheless, in the latter stages of this thesis, the OvO approach was used.

Logistic Regression

Logistic regression builds upon the linear regression model by generalizing it to a classification scenario [Goodfellow et al., 2016, Section 5.7.1]. Logistic regression inherently uses probabilities to determine class affiliation, thus input variables are transformed using the logistic sigmoid function, squashing outputs of each linear function defining the classification process into values spanning from [0,1] interpreting those values as probability. This leads to the following equation

$$p(y = 1|x; \theta) = \sigma(\theta^T x) \tag{3.12}$$

with σ as the sigmoid function and θ defining the parametric vector that is to be optimized. In a two-class classification setting, $p(y = 0|x; \theta)$ would just equal $1 - p(y = 1|x; \theta)$.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Parameters can be calculated maximizing the log-likelihood by minimizing the negative log-likelihood, for instance using gradient descent and the cross entropy error [Goodfellow et al., 2016; Bishop and Nasrabadi, 2006] (see Section 3.2.2 for more details). In multi-class settings either several binary OvR models are trained or Multinomial logistic regression is used, which details are not explained here in detail [Bishop and Nasrabadi, 2006, Section 4.3.4].

XGBoost

Extreme Gradient Boosting (XGBoost) is an ensemble method of the Gradient Boost family originally proposed by Friedman [2001], which uses a simple classifier like decision trees as the base classifier. These models build upon the iterative training of a base classifier numerous times on previous errors. XGBoost combines this approach with parallel computing, cache access patterns and more computational tweaks to improve learning time and performance [Chen and Guestrin, 2016].

Trees are built using a *gain* parameter that decides the best split points and is controlled via a regularization parameter λ. Tree sizes are limited by a complexity parameter, γ. Consecutive trees are building upon the prediction of previous trees, namely their residuals. This leads to stepwise optimization towards the ground truth. The consecutive tree's impact is again controlled using a learning rate parameter η to prevent overfitting. Trees are being built until the residuals are small or a maximum number has been reached.

3.5 Visual Analytics

Visual analytics (VA) has been formally defined as "the science of analytical reasoning facilitated by interactive visual interfaces" by Cook and Thomas [2005]. The authors argue that, in order to fully capitalize on analytical processes, solutions for analysts need to be created that enable them to make use of their cognitive and perceptual capabilities during the analytical process. With exploding types (variety), volumes and velocities of data and its accompanied desire for analytical processing and interpretation [Katal et al., 2013], VA is a research area that combines many existing fields. It does so while trying to build innovative solutions to help analysts and experts fulfill their analytical tasks by providing interactive visual representations. While the amount of data we create and the necessary hardware to store this data are evolving simultaneously, it also enables us to conduct a more complete analysis. But we as humans are not as quick to evolve and thus are limited in what we can process visually and conceptually (compare Section 3.2.1). Therefore, applications enabling analytical processes must deal with a number of scalability issues [Cook and Thomas, 2005]: *Information*, *Visual*, *Display*, *Human* and *Software* scalability.

Furthermore, Cook and Thomas [2005] define four core areas that are key parts of VA, as they propose that research is done in those areas to increase the impact of VA:

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

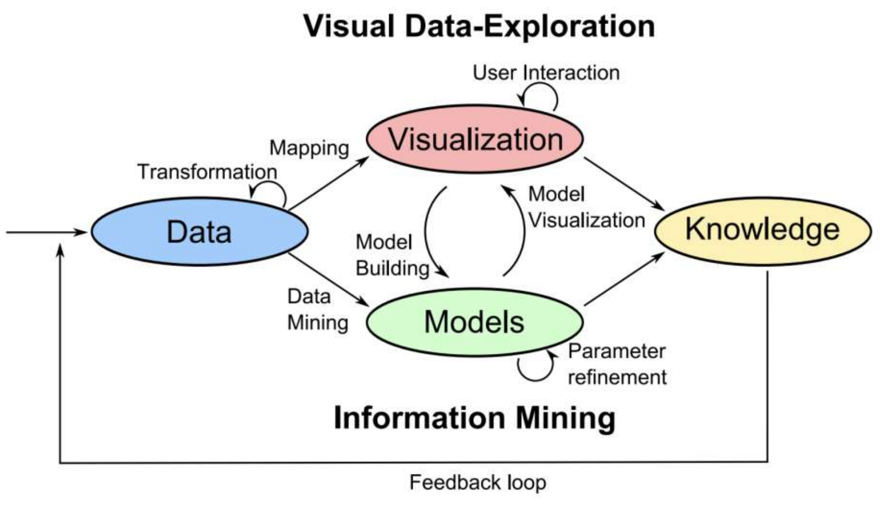


Figure 3.17: Visual analytics process by Keim et al. [2008], characterized by interaction between data, visualization and models while users discover knowledge.

- Analytical Reasoning: support decision making, user insights and situation assessment with respect to the scalability issues and that some analytical tasks may be solved under extreme time pressure.
- Visual Representations and Interaction Techniques: Developing applications that enable interactive use with visual representations fitting user groups.
- Data Representation and Transformation: Data preprocessing and unifying types to enable comprehensive information overview.
- Production, Presentation and Dissemination: Results must be communicated to different audiences, from policy-makers to the general population in unambiguous and meaningful ways. Therefore flexible tools may be used that enable analysts to produce reports and save created representations in order to communicate insights.

Keim et al. [2008] build upon the previously mentioned terminology by Cook and Thomas [2005], generalizing their theories. Keim et al. [2008] think of VA as the intertwining discipline between data analysis, visualization and interaction. By that, it not only inherits known best practices and difficulties from those disciplines but also influences back into those disciplines. Figure 3.17 displays the visual analytics process including data preprocessing (transformation), mapping or encoding, model building and evaluation, and knowledge extraction from that application by a user or user group. This process can be replayed which enables faster or more accurate insights through *feedback loops*.

With machine learning and deep networks being so popular, the question might arise why a human is even necessary for certain analytical tasks. Deep learning based solutions aim

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

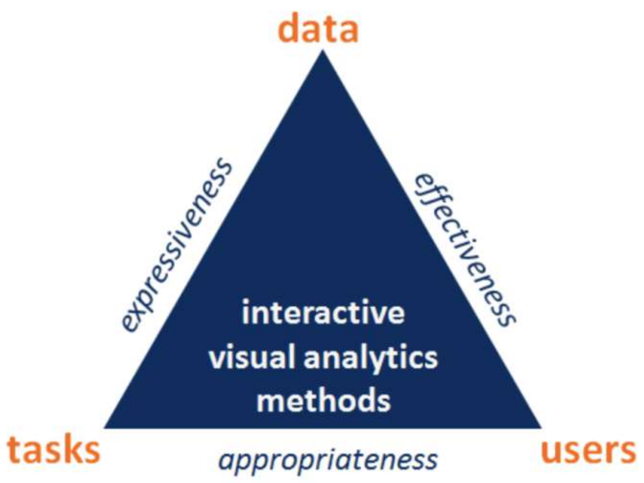


Figure 3.18: Design triangle as proposed by Miksch and Aigner [2014].

toward end-to-end pipelines, which deal with pre-processing and data transformation by themselves by learning from huge datasets [Holzinger, 2018] (compare Section 3.2.2). While in less sensitive contexts like high-frequency trading or facial recognition of the mass it might be appropriate to deploy certain systems without humans in the loop. In high sensitive contexts like the medical domain, where the decision process is not entirely machine-translatable (Section 3.2.1) and systematic errors would be of high influence, analytical tools to provide decision making including humans in the loop are necessary to increase transparency and trust in algorithms. When legal aspects are added this gets transparency and *why* a decision was made gets even more important. Thus, VA applications should help enable humans fast pattern detection properties and leverage on building efficient applications doing so. [Munzner, 2014; Ullman, 2000; Holzinger, 2018]. Furthermore, deep and machine learning is shifting from huge automatic black-box models towards explainable AI (XAI), where focus is set on debugging, refining as well as explaining decisions to humans using VA and interactive visualization [Holzinger, 2018; Choo and Liu, 2018].

3.5.1 Visualisation Principles

When designing VA dashboards Miksch and Aigner [2014] propose the Design Triangle shown in Figure 3.18. We already described data characteristics in Section 2.2 and will describe requirement analysis and task definition in the upcoming Section, while users are already described in Section 2.1.1. Still one thing to discuss here are the ground principle of visualization that fill the triangle in Figure 3.18. One of the more important views in this context have been introduced by Shneiderman [2003]. The *Visual Information Seeking Mantra* that is:

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

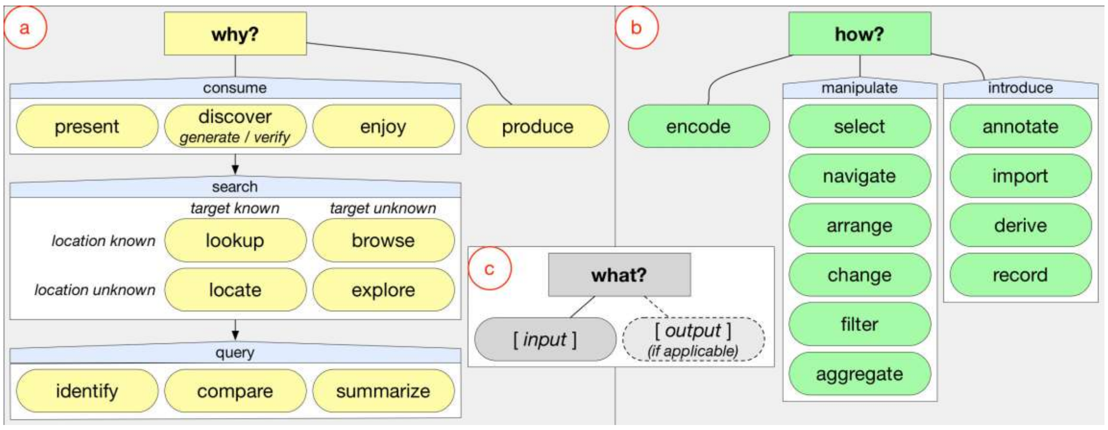


Figure 3.19: Multi-Level typology of abstract visualization tasks by Brehmer and Munzner, adapted from [Brehmer and Munzner, 2013]. Task motivation *why?* is defined in a top-down way, from high level definitions to detailed low level taxonomy (consume → search → query). Methodology on how the task is solved is defined in *how?*, while also defining the tasks input and output in *what?*. Defining input and output allows for the definition of subsequent paths.

Overview first, zoom and filter, then details on demand

The dashboards designed in this work later (Section 4.7) are focused on filling out that exact mantra and key insights find itself in the typology used for requirement analysis [Brehmer and Munzner, 2013]. Hence its still an important building block of todays design principles of visualisation.

3.5.2 Requirement Analysis

Remembering the main idea behind VA is knowledge derivation, which can also be referred to as solving an analytical task. This process can be widely defined [Brehmer and Munzner, 2013; Munzner, 2014]. When does a task end, and when does a potentially new one start? What is a user's context or motivation behind using a VA application, and what is the user's background knowledge? In order to verify or evaluate the outcome of visualizations and VA applications, tasks and user groups need to be defined in a standardized way.

The *mutli-level typology of abstract visualization design* shown in Figure 3.19 proposes a low-level typology which enables VA developers to define user-tasks [Brehmer and Munzner, 2013]. The typology is designed via answering three simple questions defining the "means and ends of a task". *Why* is the task performed, *how* is the task performed and *what* are its inputs and – if available – outputs. This also allows for evaluation by answering those questions after designing an application.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Why: A top-down approach is proposed which helps describe why a task is performed from high-level to low-level definitions. The first question to be answered is if the user is trying to *produce* or *consume* visual artifacts. *Producing* can be introducing new data transformations, annotating data, creating new visualizations or saving created ones for exports. *Consume* is fulfilled when a user is either consuming information or seeking new information-driven by an analytical task. When *presenting* data, the typology defines a way of data-storytelling and guiding an audience. *Discover* terms the desire of a user to generate or verify a hypothesis. If a user casually stumbles upon an application it is termed as *enjoy*. The user then is not driven by some hypothesis but may want to explore. This translates very well to a museum setting. To specify in more detail, the medium-level term *search* is about how elements of interest are found in the visualization. Therefore it distinguishes between previously known targets or locations. If a user is for instance searching for targets with known characteristics this refers to *browse* and *explore*. In contrast, if a user wants to find a specific item in a visualization, for instance, a country on a map, and he does not know where this country lies, he must first locate it. If he is familiar with the geographic aspects he can locate it. Lastly, a user can *query* for a condensed portion of information after a successful search. Here, *identify* refers to single targets, *compare* to multiple and summarize may refer to the whole dataset or a set of possible targets.

How: Here the tools and functions a user can interact with are defined. In the majority of visualizations, data has to be encoded somehow in order to grant quick information gain. Encodings can be chosen between arranging data – for instance by connecting or ordering data – and mapping data to color, shape or angles. A user can be enabled to *manipulate* data for example by reducing items (select, filter, aggregate) or *introducing* new artefacts.

What: The most important question might be should be the context of the visualization. This is broadly defined and ranges from pixel to values or nodes and even clusters [Brehmer and Munzner, 2013]. Here, an input needs to be defined for every task, and if one is to model consecutive tasks, an output of some form has to be defined.

This taxonomy is used in Section 4.2 to define our user specific tasks.

3.5.3 Validation

Finally, when creating VA applications, some kind of validation is important. Many applications are ineffective in what they try to achieve [Munzner, 2014, 2009; Cook and Thomas, 2005]. In this thesis we will use the Nested Model by Munzner [2009] illustrated in Figure 3.20. It consists of a four-level top-down design.

The first level is the domain problem characterization. This includes domain-specific users and interests as well as their data and sets the overall setting of the solution. Requirements are best met and validated by conducting user interviews with domain experts or observing them to gain knowledge of current shortcomings and problems. The threat is to ask the wrong questions, thus mischaracterizing the problem and making

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

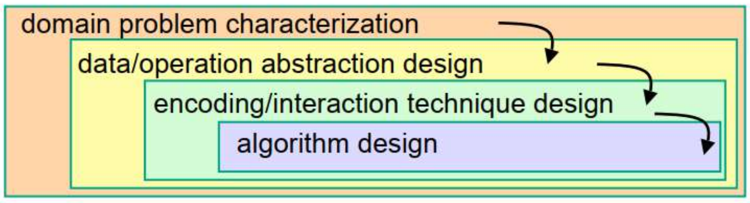

domain problem characterization
data/operation abstraction design
encoding/interaction technique design
algorithm design

Figure 3.20: Nested Model for Visualization Design and Validation. Outer layers operate as the input for the succeeding inner layer. Image from Munzner [2009].

wrong assumptions [Munzner, 2009, 2014] The second level is about the data. The goal is to abstract domain-specific knowledge into generic representations. In particular, domain-specific vocabulary is mapped to the vocabulary of information visualization. This also includes transforming and preprocessing data into visualization-ready types and forms. Threats like bad wrong operations or data types can be minimized by conducting target user interviews. At the third level, the abstracted data is visually encoded. This also includes how potential users may interact with the data given the chosen encodings. Pitfalls at this level include ineffective encodings or visualizations – the user does not understand it. VA design can be validated by conducting small lab studies or result image analysis. Lastly, algorithms used by an application can be validated. This might include consecutive rendering or encoding of data due to interaction techniques, which may be slow and hinder user performance. Measuring time and making sure potential constraints are met would suffice here. Also, the correctness of the algorithm should be validated. In this thesis no potential constraints are defined regarding runtime other than usability of the application should be interactive. Nielsen [1994] defines three response time limits for applications that effect user experience and how to deal with them in order of feedback provided by applications:

- 0.1 seconds: This is the limit for having the user feel that the system is reacting instantaneously. No feedback is needed for the user.
- 1 second: The range of 0.2 to 1.0 seconds depicts the feeling of freely navigating the data space. In this time-span, users have a feeling that the application is doing work. Delays more than 1 second should be indicated by the application, f.e. by changing mouse cursor.
- 10 seconds: Operations taking longer than 10 seconds should be indicated by the application, and feedback should be provided in some form of a *percent-done* feedback. Wait times like this risk losing the users attention.

In addition to the nested model, among others, Isenberg et al. [2013] observe *usage scenarios from visualization researchers* as a possible way of validation. Here, a designer or researcher mimics the target group and uses the visualization to prove its correctness.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 4

Data Analytics

This chapter gives an overview of the design and the environment the practical implementation was done in, and documents findings along the way. Section 4.2 introduces task and requirement analysis for the previously defined user groups following the abstract visualization task design [Brehmer and Munzner, 2013]. In Section 4.1, data preprocessing is compared and defined as it was used for most of the subsequent tasks. Clustering of patients and comparison of different strategies for outcome prediction is conducted in Section 4.3. Comparison of different transfer learning approaches, image pre, and post-processing, radiomics feature extraction as well as comparison of prediction strategies is presented in Section 4.4. In Section 4.5 we investigate how features can be combined to possibly increase prediction performance. We formally define the environment used in the thesis in Section 4.6. This includes a short presentation of possible front-end environments and argues why one was chosen over other options. Finally, Section 4.7 provides an overview of possibrle ways to communicate results given the data by fulfilling the previously defined tasks and requirements.

4.1 Data Preprocessing

For data preprocessing, target variables were transformed first. The three binary targets depicting the outcome of the hospitalization, *last.status*; ICU admission *is_icu* and ventilation status, *was_ventilated*, have been transformed from string to binary form in order to be working with a numeric representation which is necessary for algorithms to be fit towards target variables. From that, we created a numeric representation by combining the outcomes for each patient, leading to eight target variables that defined our multi-label classification setting. The targets and their corresponding encoding are shown in Table 4.1. Original target columns have been removed before any subsequent tasks. The very under-represented class 5 of *Deceased + Ventilated* patients was merged into class seven.

last.status	is_icu	was_ventilated	Multi-label	Description	Count
0	0	0	0	Hospitalized (only)	948
0	0	1	1	Ventilated	11
0	1	1	2	ICU	46
0	1	0	3	ICU + Ventilated	100
1	0	0	4	Deceased	61
1	0	1	5	Deceased + Ventilated	2
1	1	0	6	Deceased + ICU	11
1	1	1	7	Deceased + ICU + Ventilated	100

Table 4.1: Multilabel targets and occurrences in the dataset

Many features are redundantly present in the raw dataset. Some variables are provided in form of a numeric value. Additionally those variables are available in form of one-hot encoded information, in the form of pre-defined ranges suitable for the respective variable. For example, the Body Mass Index (BMI) was given in form of its numerical value, a positive float, and additionally, two variables were present indicating BMI over 30 and over 35. Those variables may be of use when creating manual decision processes but it is not helpful when training a classifier and thus, these redundant features have been removed. There have been 51 dropped columns of this form.

Duplicate columns are removed as well. This only affected one column in the dataset, namely sodium values in serum or plasma, which was available twice in the tabular data. Furthermore, prior to any imputation, features that were eligible for one-hot encoding and filling missing values with zeros were determined manually. Those features are *kidney replacement therapy* and *kidney transplant* which have been imputed with zeroes, because besides missing values, only positive values were present in the binary variables. Other variables, i.e. *age*, *gender*, *Urine.protein*,*hf_ef_v* (heart failure and ejection fraction) and *smoking status* have been one-hot encoded, as they inherit different categorical values.

Additionally, for non-explorative tasks, targets and columns that could implicitly indicate the outcome of any of the target variables are removed as well. As already mentioned in Section 2.2, multiple variables correlate or indicate the outcome of a patient's hospital stay. For example, the number of ventilated days of a patient is a sign for ventilated patients, and as there is a high co-occurrence between people being ventilated and being submitted to ICU, there is also a high correlation to that target. A longer hospitalization stay means there is a higher possibility of a severe case including ventilation or ICU stay. Suffering a kidney injury during the hospitalization was also an indicator for a severe case with a medium correlation factor. Those features were all removed. In addition to that, all columns that would indicate something about the patient's stay in the hospital also were removed. Those included information about the patient receiving therapeutical medicine, f.e. against thrombosis like heparin, which would — again — indicate a longer hospitalization stay and implicitly lead to information leakage. After that, no features

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

with strong correlation ($corr > 0.95$) to target variables remained in the dataset.

The state of the data is now at a point at which it would be possible to train a classifier. Generally, features can be split into two groups now, as mentioned in Section 2.2. On one hand, there are features that would be available prior to any patient arriving at the hospital or a medical facility at all. This is data that consists of symptoms, prior medical records with possible medical pre-conditions, and some basic features like oral temperature. It would be possible for a layman to input those features into a front-end application that could predict from those features alone. This feature set is comparable to data used in previous works mentioned before [Dai et al., 2020; Li et al., 2020b]. These features would be also fall under the definition of Clinical Endpoints defined by Strimbu and Tavel [2010] and presented in Section 3.1.3.

Additionally, there is data available that stem from medical or laboratory tests. This data would not be available to laymen users using such a prediction system by default. We thus define that only data that would be available to interested layman users are going to be worked into the dashboard design as possible inputs. Due to usability and improved results shown in previous works Li et al. [2020b]; Zhao et al. [2020], it should be possible to infer additional data from medical tests for a user as well. This could be done by using clustering (see Section 1.2) or neighbor-based methods. In general, temporal aspects of the data have not been used in prediction but only in an exploratory context through visual analytics.

Sklearn's *StandardScaler* was used to scale numerical variables by removing the mean and scaling to unit variance prior to any prediction, clustering, or exploration scenarios, when not documented otherwise, to eliminate any numerical problems due to differences in the data dimensions.

4.1.1 Missing Data and Imputation Strategies

We define the data as being Missing at Random MAR, as it can be defined as a sample of the population with a known precondition (Section 3.3.2. In our case, the precondition is being a hospitalized patient with a positive PCR test for COVID-19.

When grouping records with missing data by label, one can argue that missing data is occurring at a higher rate for patients with less severe cases, as shown in Table 4.2. This makes sense when further investigating for what variables the missing data occurs: After filling manually selected variables with zeros (compare Section 4.1), features that are missing at most are *blood_pH* with $N_{missing} = 1060$ and *A1C* with $N_{missing} = 875$. *BMI* information has been missing in $N_{missing} = 400$ records.

When we group those high-missing features by target variable, one can observe similar results that are shown in Table 4.3. A reason for the difference could be that it was not necessary in some cases to conduct laboratory tests since the illness was not deemed too dangerous and therefore the records are missing in the final dataset. *Blood pH* values

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

label	Outcome	Incompl. Rows	N per label	Proportion
0	Hospitalized (only)	901	948	0.95
1	Ventilated	9	11	0.82
2	ICU	38	46	0.83
3	ICU + Vent	89	100	0.89
4	Deceased	51	61	0.84
5	Deceased + Ventilated	2	2	1.00
6	Deceased + ICU	9	11	0.82
7	Deceased + ICU + Ventilated	78	100	0.78

Table 4.2: Patient records with missing data by label.

label	Outcome	pH	A1C	BMI
0	Hospitalized (only)	0.95	0.74	0.33
1	Ventilated	0.36	0.73	0.27
2	ICU	0.63	0.46	0.24
3	ICU + Vent	0.37	0.53	0.25
4	Deceased	0.79	0.62	0.20
5	Deceased + Ventilated	NaN	0.50	0.50
6	Deceased + ICU	0.09	0.55	0.18
7	Deceased + ICU + Ventilated	0.40	0.45	0.29

Table 4.3: Missing-ratio by label for variables with high occurring NA's. PH and A1C (measurement of blood sugar) values are results of blood tests. Non-standard pH blood values can occur due to acute kidney or lung function problems.

deviating from the norm occur during acute kidney or lung-related diseases [1]. Given that COVID-19 is a respiratory disease and comparing this information to the labels in Table 4.3, we can observe a lower proportion of missing values for *Blood pH* values in patients records that have been ventilated or admitted to ICU.

Comparing Imputation Strategies

Different imputation strategies are compared in the context of clustering the data, which have been mentioned in Section 3.3.2. Imputed data is explored and one method is chosen as the best fit after quantitative and qualitative comparison. This method is then also used prior to prediction (Section 4.3.2.

Figure 4.1 shows CH and Silhouette scores for Ward and K-means clustering. This time, different imputation strategies have been applied to the dataset prior to clustering. We can observe a reasonable improvement when using the indicator encoded as violet lines

[1] `https://www.healthline.com/health/ph-of-blood#blood-p-h-test`

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

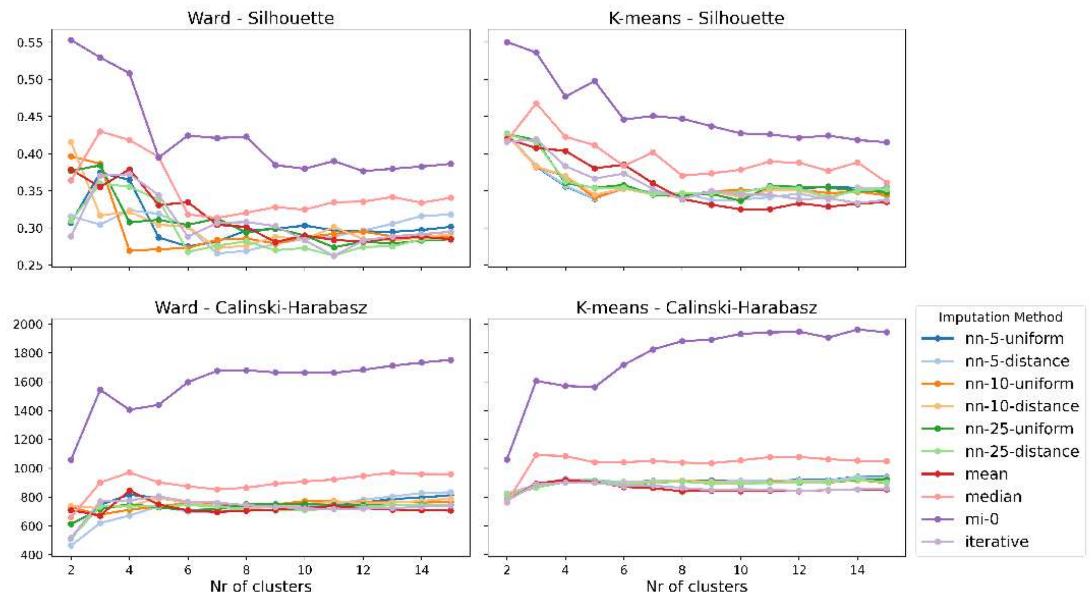


Figure 4.1: Comparing different imputation methods and their impact on clustering metrics. Indicator method is abbreviated with *mi* and zero-imputation was used here.

in Figure 4.1. In this case, the indicator method was used by imputing zeroes. As a runner-up, the metrics would suggest using kNN-imputation.

When having a detailed look at the impact of the different imputation strategies on the dataset, we can observe numerous things. Figure 4.2 shows three features, one binary feature indicating if the patient was experiencing *vomiting*, *blood lymphocyte* values, and *creatine kinase*, an enzyme, for which high values can be an indication for several health issues. As for iterative imputation, we observe an introduction of negative values through all three features. None of those features benefit from this artificial distribution as one cannot have a negative number of *creatine kinase* enzymes, lymphocytes or have a negative indication for *vomiting*. We introduced unrealistic values which yields the iterative imputation obsolete. The indicator method, that reported best metric scores in Figure 4.1, introduces another common problem. It vastly influences the original feature distribution towards zero, which can best be observed in *lymphocytes*. While the original data seems to be distributed somewhat normally around 0.7, the indicator method introduces another peak at zero. Comparable phenomena can be observed for *creatine kinase*, where the distribution is now highly skewed towards zero. The same problems occur for mean and median imputation. They do not keep the original distribution in place but introduce new artificial peaks.

For kNN imputation we observe the most realistic imputations of the presented methods. One disadvantage is the creation of in-between values for binary variables, hence *vomiting* takes values between 0 and 1. As proposed in Section 3.3.2, rounding would be an

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

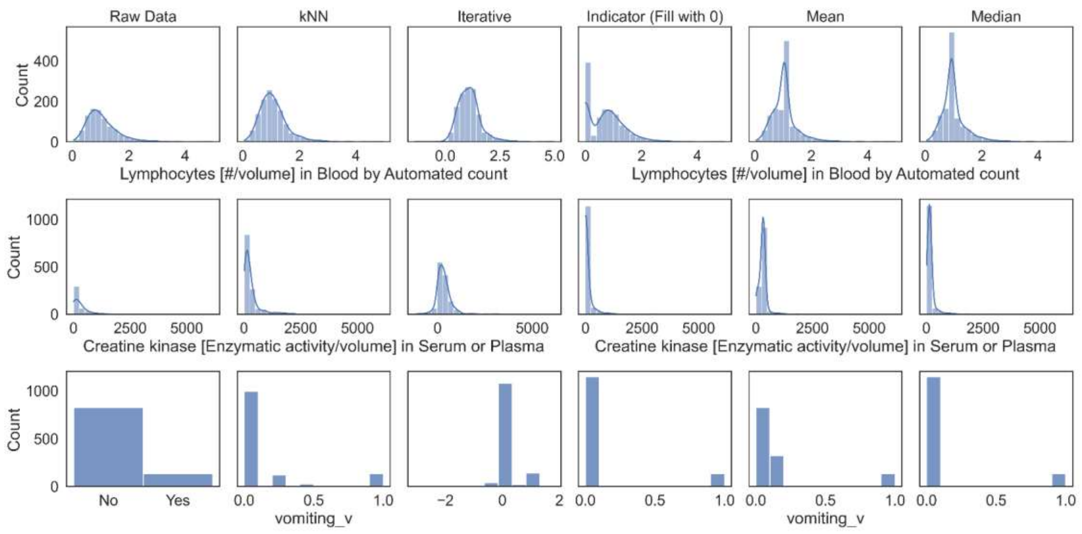


Figure 4.2: Imputation methods impact on feature distributions in the dataset shown for three example features, one binary and two numerical.

appropriate solution for this problem, but as Ake [2005] proposed, this might not be necessary when the data is used for analysis straight away, and thus this option is waived.

Conclusion and Further Implications

Overall kNN-Imputation gives the best results for the dataset given the classification metrics provided and the qualitative exploration of the imputed data. This is also on par with results from previous research [Jerez et al., 2010; Yenduri and Iyengar, 2007]. We thus use it as the basis for further analytical solutions built upon the imputed data such as clustering and prediction of hospitalization outcome.

4.2 Task & Requirement Analysis

As explained in Section 3.5.2 we use the taxonomy for requirement definition by Brehmer and Munzner [2013]. For each user group, multiple tasks are defined that have to be solved by the visualization and thus define the design. The tasks are enumerated in the following scheme: *User Group Incremented Number*. When defining the tasks we keep in mind that our designs are not created in contact to any medical experts or general user population test users. Therefore to stick to most agreed to design principles for information visualization as well as the concepts described in Section 3.5

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

4.2.1 Medical Experts

The first user group is medical experts. The goal is to provide an overview over the available data as well as methods to search for patients with specific attributes, filter for outcomes and give an overview over the disease progression during the hospitalization where applicable. Progression overview is only achievable for patients for consecutive observations during their hospitalization, namely patients with several CXR images available, as medical data is only available for one point in time. Medical experts should be able to use the application for lookup of similar patients in order to receive an indication for their outcomes and retrieve a prediction for new patients CXR images, which could be used as a tool for decision making support. Visual notation of the medical experts tasks are found in Figure 4.3.

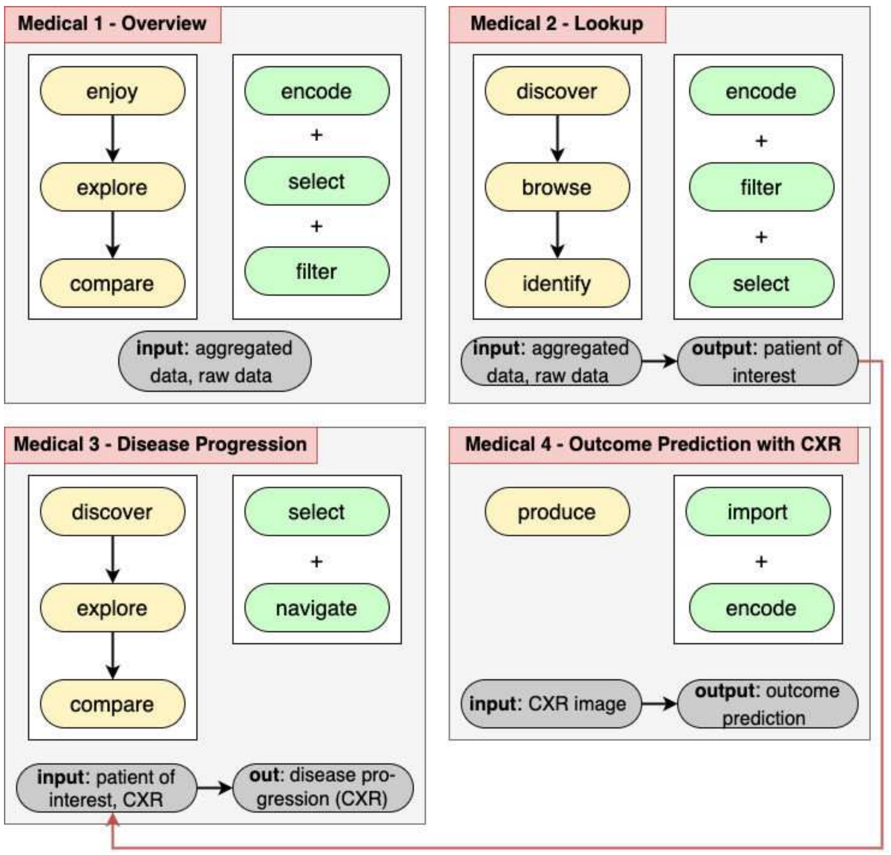

Medical 1 - Overview
enjoy
explore
compare
encode
+
select
+
filter
input: aggregated data, raw data
Medical 2 - Lookup
discover
browse
identify
encode
+
filter
+
select
input: aggregated data, raw data
output: patient of interest
Medical 3 - Disease Progression
discover
explore
compare
select
+
navigate
input: patient of interest, CXR
out: disease progression (CXR)
Medical 4 - Outcome Prediction with CXR
produce
import
+
encode
input: CXR image
output: outcome prediction

Figure 4.3: Abstract visualization task design for the medical experts in visual notation as proposed by Brehmer and Munzner [2013].

Medical 1: Overview

Given the application is not designed to solve a specific problem that was defined a-priori, for instance by conducting interviews with user groups [Sedlmair et al., 2012], the first

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

goal is to provide an overview about the data available in case of a *casual encounter*. In the multi-level typology this translates to *enjoy*. Targets of particular interest as well their locations are unknown and shall be *explored* in the first step. In order to give the user the possibility to dive deeper, different groups of patients should be highlighted in order to enable *comparison*. This should be achieved by *filtering*, *encoding* and *selection* of attributes of particular interest. The input data is the raw electronic health data that may be aggregated to some degree to enable compact, comparable representations of the data.

Medical 2: Lookup

After making itself familiar with the data, the user should be able to discover data for patients with particular characteristics, for example a decease status, existing pre-conditions or age. This might be the results of the previous task Medical 1 but it does not need to be so necessarily. Users should be able to *browse* in order to *identify* a patient of interest by *selecting* and *filtering* specific attributes. The input is again the raw data in addition to aggregated views. The output can be one specific patient of interest that can be *selected* to receive a visualization of the decease progression using the patients CXR images.

Medical 3: Discover Patients Decease Progression

Given a patient of interest has been selected as the result of prior tasks, the progression of the patients hospitalization stay should be shown in form of one patients CXR images in ascending time order. The goal is to *present* the images with the possiblity to select specific time points and *navigate* back and forth. The user is able to explore the patients hospitalization development and compares different points.

Medical 4: Predict Outcome for new Patient based on CXR Image

The user should be able to *produce* a prediction for a patients hospitalization outcome, as a best effort approach to support decision making. The prediction is based on the available CXR data. An image can be *imported* and is processed by the application similar to the data used as the prediction basis and produces a prediction for the outcome based on this particular image. Predictions from the model are then encoded and shown to the user.

4.2.2 Analytical Experts

Analytical experts are users with analytical background, including machine learning engineers, data scientists or mathematicians. Interest in the data stem from their ability to create predictive models. Analytical users should be familiar with many different visualization types and prediction models proposed in Section 3.4 and are most interested in seeing how hyperparameters affect models on the data at hand. In Figure 4.4, tasks for the analytical experts user group are presented in visual notation.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

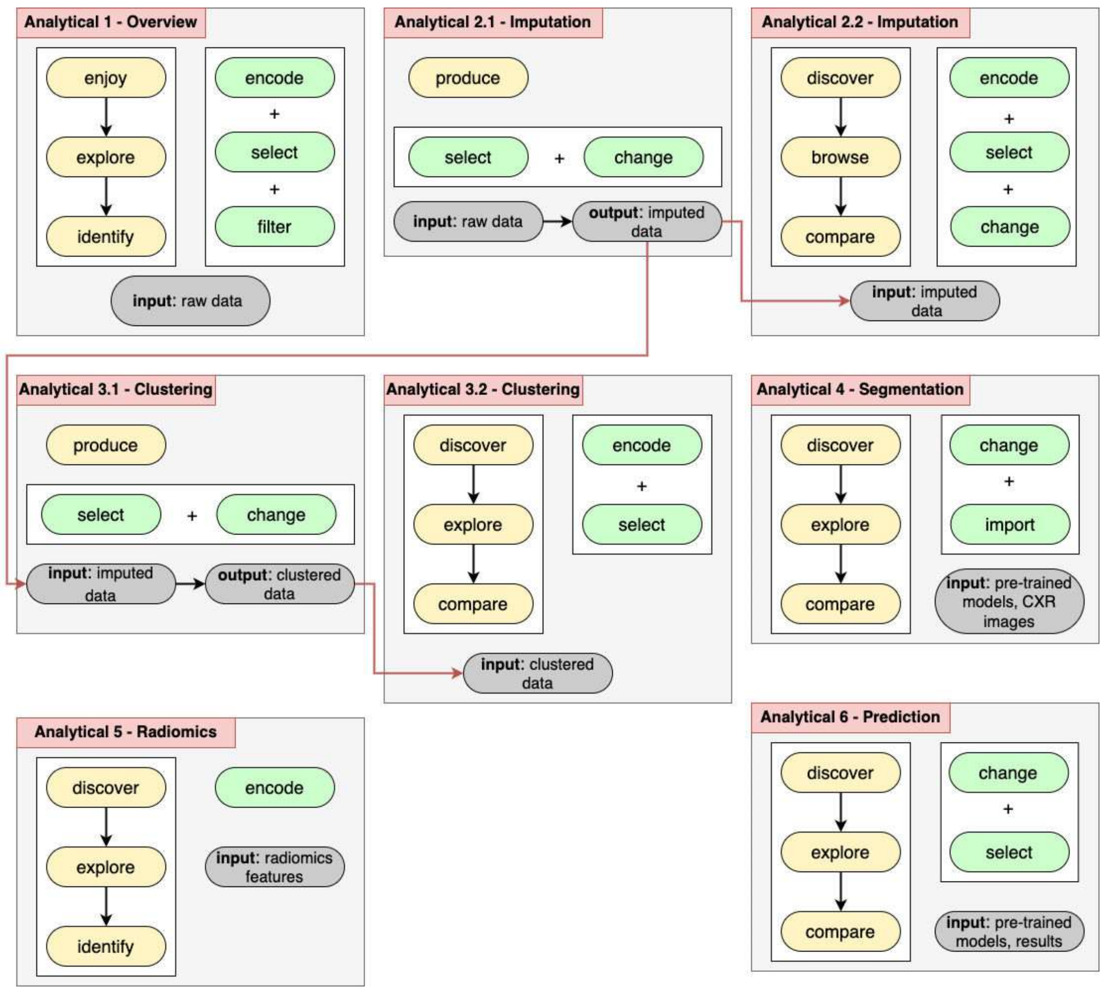


Figure 4.4: Abstract visualization task design for the user with analytical background in visual notation as proposed by Brehmer and Munzner [2013].

Analytical 1: Overview

Again we start with an overview. Users landing on the application are first engaged to *enjoy* the application. Therefore they should be enabled to explore the data in detail to *identify* variables and patients of interest. *Encoding* data to simple structures in the first step is crucial. From then, *selection* and *filtering* capabilities are used to enable exploration of the data. The input is raw data in the first place.

Analytical 2: Discover Imputation Methods

After getting an overview, analysts will recognize many missing data variables in the original dataset. Here, different imputation strategies should be implemented which can be *selected* by the analyst. By *changing* them, different imputation settings are applied

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

to the data. The input is again raw data which will be imputed and will be the basis for the upcoming tasks, hence the connections to *Analytical 2.2* and *Analytical 3.1* in Figure 4.4.

After changing, imputed data should be displayed in a way that lets analysts decide which might be the best fitting method. Analysts *discover* imputation effects on the data by *exploring* different models which are *compared*. Features will be affected differently so analysts should be able to *select* features of interest, where we use *encodings* to display the data in a informative way. This can be repeated numerous times.

Analytical 3: Discover Patient Clustering

Similar to the previous task, this is divided into two sub tasks. Again, first a clustering strategy should be *selectable* with hyperparameters that can be *changed*. This produces clustered data from the imputed data of task Analytical 2.

Next clustering outputs should be made *comparable* by *encoding* data and *selecting* different views on the data. This will enable *discovery* as well as *exploration* of different clustering methods.

Analytical 4: Discover Automatic Lung Segmentation

This task covers lung segmentation. Similar to the medical experts task, a CXR image should be *importable* which lets an analyst *explore* the performance of the automatic segmentation. Additionally analysts should be provided with options to influence the segmentation in order to identify a best strategy by *comparing* methods. The input are pre-trained models for lung segmentation as well as user provided images.

Analytical 5: Discover Radiomics Feature Extraction Results

The radiomics features extracted from the lung segmentaion data. Similar to the first task an overview should be given which enables *discovery* and *exploration* of the features to *identify* important features. This is done through *encoding*. The input here is solely consisting of radiomics features.

Analytical 6: Prediction

Electronic health data cleaned in previous tasks shall be used to enable analysts to train different prediction models in an interactive way [Sacha et al., 2017]. Again a user should be able to *explore* different models and *compare* them, by *selecting* models and data as well as *changing* hyperparameters. Results should be *encoded* in a way that comparison is easy and straightforward.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

4.2.3 General Population

Tasks for the general population group should be solvable by layman users without any specific background knowledge. This includes knowledge about data analysis and task-specific visualization types, medical topics or COVID-19 related characteristics aside from the information that could be received through public media coverage. Therefore it is of greatest importance to reduce complexity for the user group and make use of straightforward and easy-to-use embeddings and non-complex visualization types. Tasks for the general population are displayed in visual notation in Figure 4.5.

General 1: Overview

The first task of the dashboard is to give users an overview of the data in form of groups of patients and *present* it to them in a simplified way. The goal is to generate simple *encodings* that enables users to easily distinguish between clusters formed via clustering methods and understand differences between them without too much detail. Users should be able to engage with one or more of the clusters presented due to demographic attributes as well as life-style choices or medical pre-conditions that could be extracted from the raw data.

General 2: Hospitalization Outcome Prediction

The second task should be designed in order to enable interactive *discovery* of the users. Given users experiences after the presentative overview in the preceding Task, users should now be able to *explore* how different inputs change the prediction of COVID-19 hospitalization outcomes. In order to simplify this, only certain features should be provided as input which distinguish the clusters of patients, including information available to every possible user. Finally, users are able to influence prediction outcomes by *selecting* and *changing* features interesting to them in an interactive way. Prediction results generated from user inputs should be *presented* in a way that is easily understandable for the users. This should enable comparison of outcomes and *identifying* important values for the prediction. Simple *encodings* should be used that give a clear and transparent picture of the prediction without demanding too much background knowledge.

4.3 Clinical Data

This section describes the process of working with the medical data records which are of tabular form with the goal of clustering patient records into semantically reasonable groups as well as predicting COVID-19 hospitalization outcomes by using clinical data. Different clustering methods, as proposed in Section 3.4.4, are compared using metrics discussed in Section 3.4.2. Furthermore, classification models presented in Section 3.4.5 were trained on the clinical data in addition to using strategies to deal with imbalanced data.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

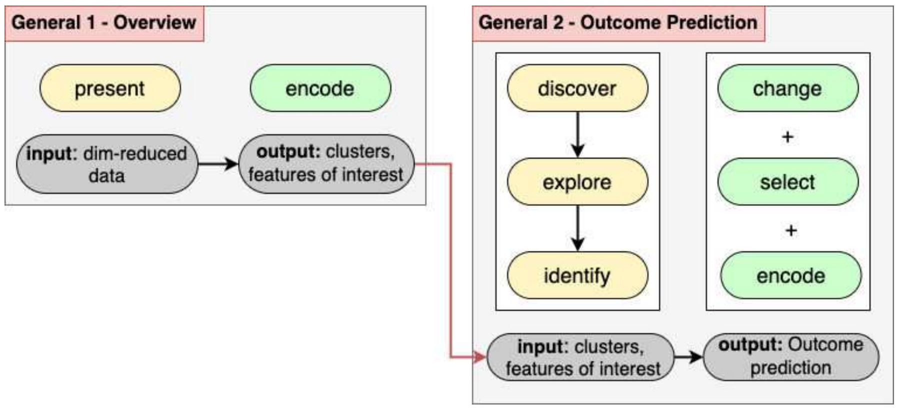

General 1 - Overview
present
encode
input: dim-reduced data
output: clusters, features of interest
General 2 - Outcome Prediction
discover
explore
identify
change
+
select
+
encode
input: clusters, features of interest
output: Outcome prediction

Figure 4.5: Abstract visualization task design for the general population.

4.3.1 Clustering

Before conducting any experiments, the data has been cleaned and pre-processed as explained in detail in Section 4.1. The first method investigated is K-means clustering. Starting with a distortions plot shown in Figure 4.6, which is showing the sum of squared distances of each observation from their closest respective cluster center. We observe that the development is not very satisfying. At best, one can already identify an optimal value of k-clusters by using the elbow method. In this case, there is no real drop-off available which already indicates that some kind of dimensionality reduction is necessary before achieving a reasonable clustering solution.

Figure 4.6 provides the comparison for two K-means initialization implementations in *sklearn*. One is *K-means++* which is more stable here, while a random initialization of cluster centers shows a less satisfying reduction in the target metric for larger k. Given the random initialization of clusters can change the optimization by quite some margin, the experiment was repeated 100 times and a 95-Confidence-Interval (CI) is shown. In further experiments we thus prefer the *K-means++* initialization.

Figure 4.7 shows Silhouette and CH metrics for different number of clusters in Ward hierarchical clustering, K-means and DBSCAN. For DBSCAN, ϵ values smaller than five did not yield more than two clusters. Silhouette scores for Ward and K-means clustering did not yield very good results on the data, not increasing much higher than 0.1, indicating overlapping clusters in nearly all settings, even for higher k. This is only supported by the linearly decreasing CH metric. For DBSCAN, we observe a desirable peak for CH, as well as increasing scores in the Silhouette metric. The problem though is that for the higher scoring solutions, only two clusters are generated by DBSCAN. Keeping in mind that one is reserved for outliers solely, results are not conclusive.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

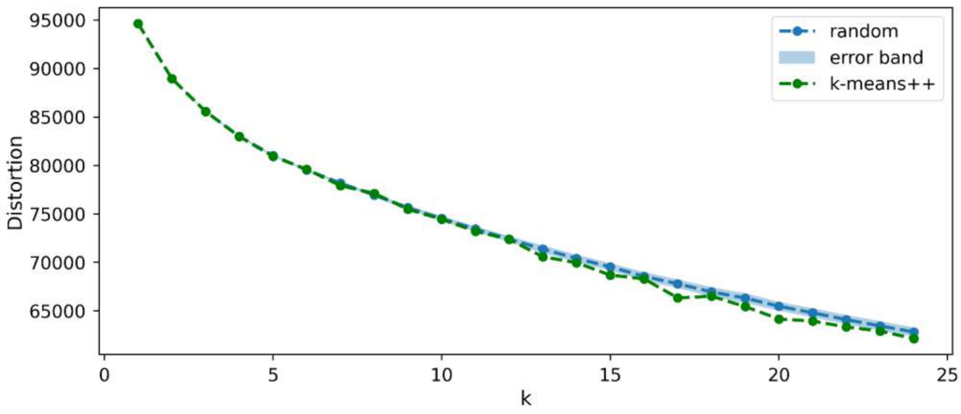


Figure 4.6: K-means distortions plot showing the sum of squared distances for each observation for $k\epsilon[2, 25]$. The error band is a 95% Confidence Interval for 100 repetitions with random initialization.

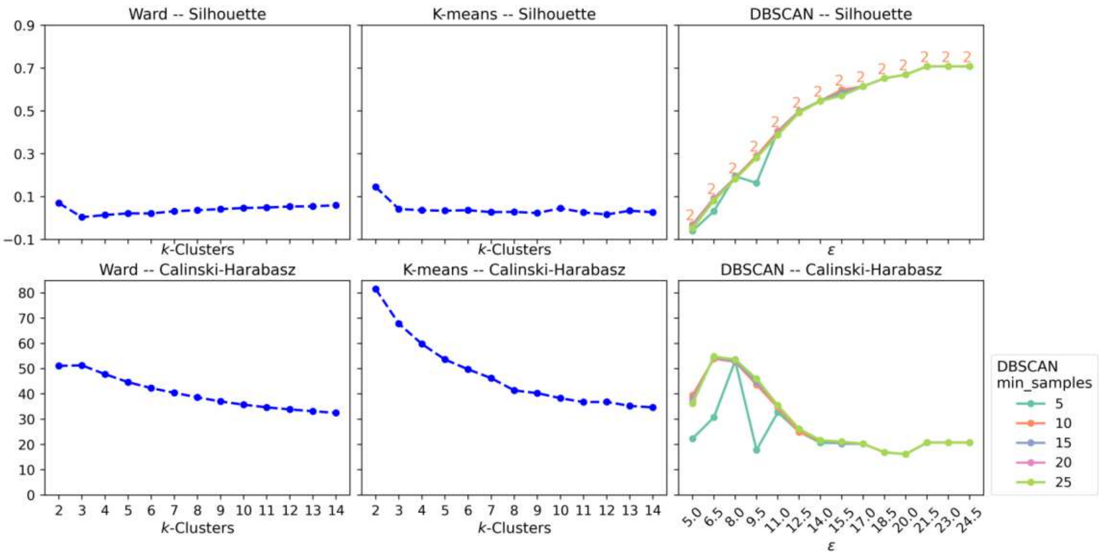


Figure 4.7: Clustering scores for the different methods with different number of k-clusters. Numbers in the DBSCAN Silhouette plot indicate the number of clusters created for the respective ϵ with $min_samples = 10$.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

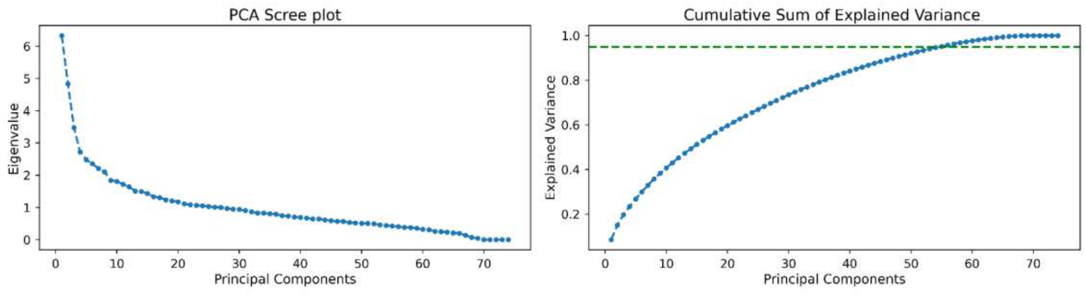


Figure 4.8: PCA scree plot and cumulative sum of explained variances for the electronic health data. The 95% explained variance is highlighted as the green dashed line.

Dimensionality Reduction

It becomes quite clear that some kind of dimensionality reduction is needed in order to obtain better results for patient clustering which also supports a visual assessment thereof. Figure 4.8 shows a scree plot with eigenvalues for each principal component of a PCA, as well as the respective cumulative sum of explained variances. Based on the results and the plot, and keeping in mind common ways of choosing the right value of principal components (PCs) as proposed in Section 3.3.3, we arrive at different values which we investigate:

- A 95% of variance is explained with 55 PCs;
- Kaisers stopping suggests including PCs with eigenvalues < 1, so 26 in this case;
- using the elbow method on the scree plot in Figure 4.8 leads to four PCs;
- using two PCs gives the possibility to plot in two-dimensions.

Those settings are used prior to clustering, leading to optimized results shown in Figure 4.9. It is obvious that a lower number of PCs leads to better scores overall since the reduced number of dimensions leads to a reduced number of possible distances in theory. Nonetheless, the differences are not that immense, but it becomes quite clear that higher dimensionality results in lower metrics with this dataset. In a reduced two-dimensional feature space, the best results are achieved over nearly all metrics and classification methods. K-means is generally slightly better than ward clustering, but cannot compete against DBSCAN for Silhouette scores. The issue is again, that DBSCAN is only outperforming other methods with a very low setting for ϵ, resulting in a low number of clusters.

Figure 4.10 shows the results for the best settings retrieved from the metrics of of Figure 4.9. For K-means a k greater than three is indicated by the peak in the CH plot. A larger value for k is only slightly decreasing the silhouette score. For Ward's, CH again

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

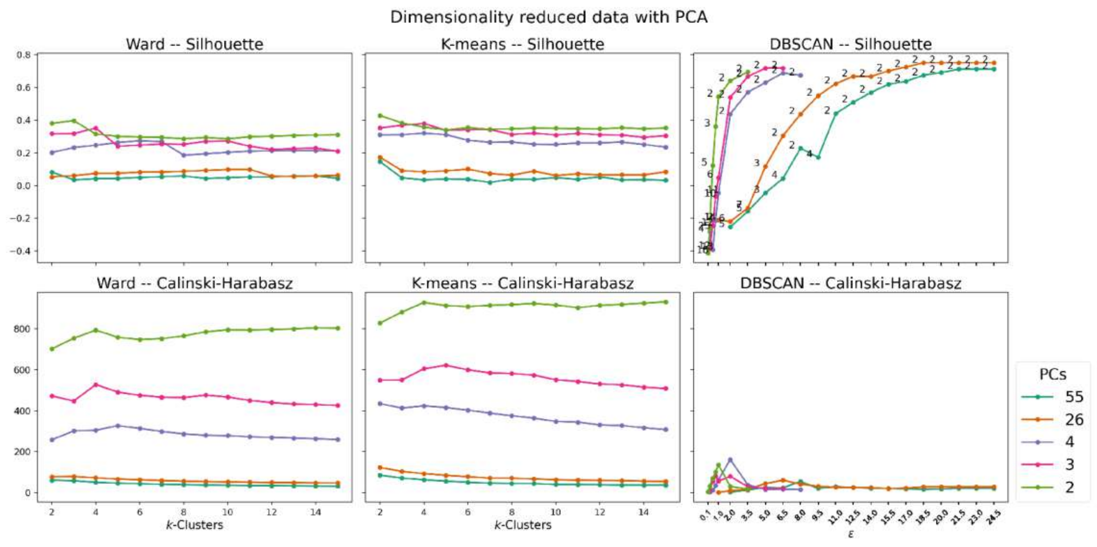


Figure 4.9: PCA metrics for different number of principal components. Number of clusters resulting in DBSCAN settings are again plotted next to the respective setting in the graph.

suggests selecting a k greater four. Metrics for DBSCAN again suggest results that will consist of two clusters only, so $\epsilon = 0.75$ is chosen which yields 4 clusters.

The results in Figure 4.10 show that DBSCAN is not a good fit for the underlying data. Results for DBSCAN also indicate that, while the method is able to detect outliers from the high-density space, those outliers tend not to have a semantically strong coherence. Not all of the patients in the outer areas of the two-dimensional space are patients which deceased. Due to the nature of DBSCAN being a density-oriented approach it is only natural that it will select an optimal centroid in the middle of the clutter.

Ward and K-means solutions do not differ much in their results. It is clearly visible how Ward clustering establishes smoother borders. Both methods identify a cluster in the left bottom region of the plot (red in Ward's solution, blue in K-means) which is mostly made up of patients that did not have a mortal hospitalization stay but were dismissed. Patients in the right upper space tend to be more probable to suffer a severe illness during hospitalization.

When applying t-SNE we hope to gain another helpful visual representation that prioritizes the local structure of the high-dimensional feature space. After testing different hyper-parameter settings for t-SNE (compare Figure A.6), we end up with a representation as depicted in Figure 4.11. K-means and Ward cluster affiliation is color encoded again. They show a very similar outcome once more. Interestingly, one can identify some smaller groups of patients at the bottom which mostly share the same outcome and cluster membership.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

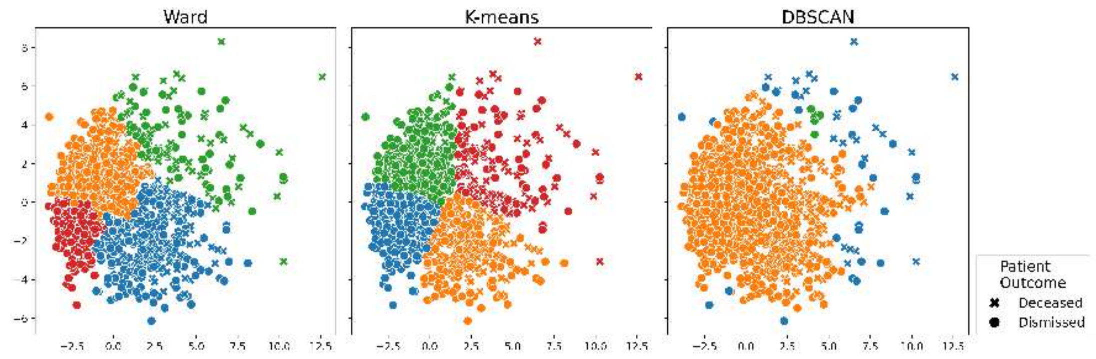


Figure 4.10: Clustering after two-dimensional PCA compared. Parameters for k and ϵ used according to best metrics in Figure 4.9.

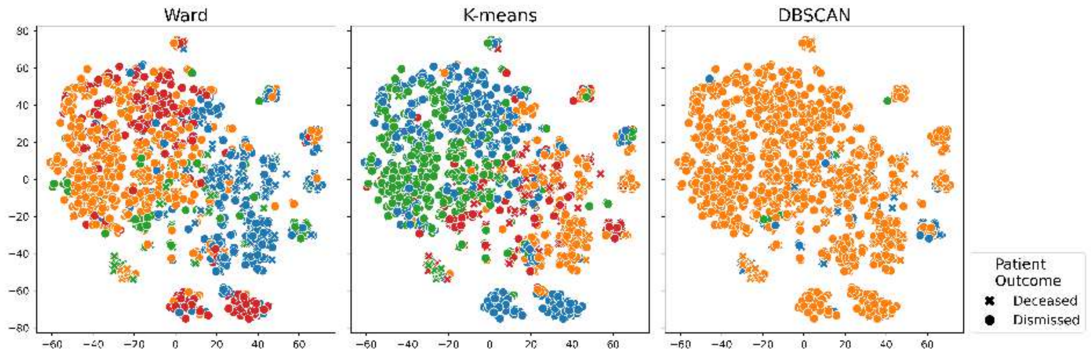


Figure 4.11: T-SNE visualization of the feature space with color coded cluster memberships for each respective method.

Conclusion and Further Implications

Given the results for CH and Silhouette scores, K-means seems to be a slightly better fit than Ward hierarchical clustering. To finalize the clustering, an *optimal* number for k has to be chosen. Repeating Silhouette scoring and plotting a scree plot for the within-cluster sum of squared distances for K-means presented in Figure 4.12, we do not recognize an obvious common choice. Given we aim to provide some choice for users and comparing our results to other results in the literature $k = 4$ seems a better choice. As a reference, Zhao et al. [2020] end up with six for ICU admission, and eight classes for risk for decease in their risk score model. [Dai et al., 2020] classify into three groups in their proposed risk score model.

4.3.2 Prediction

Again, we started with data preprocessing as described in Section 4.1. For imputation we used kNN with $n_{neighbors} = 5$. Imputations for categorical values have not been rounded.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

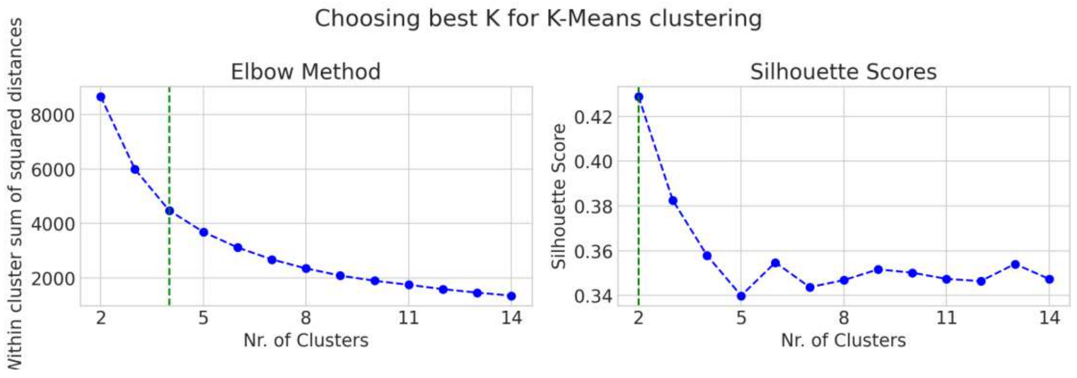


Figure 4.12: Choosing the best number of clusters k for K-means clustering. Silhouette would argue for $k = 2$ but the difference for higher choices is not that big. The elbow method would argue for $k = 4$.

We end up with 74 features after dropping irrelevant or highly numerical or semantically correlated columns.

Classification Pipeline

Next, a classification pipeline is created. Different strategies for dealing with imbalanced data are employed, namely using balanced weights where applicable, random oversampling, random undersampling, and SMOTE.

Data were split into 70% train and 30% validation and were not imputed prior to splitting to avoid information leakage. Training data was used within 8-fold cross-validation, as stratified folding was limited to the lowest number of records for minority classes, which was 8. LOOCV was not used in order to reduce computational effort. For each fold, data was scaled first, imputed using kNN imputation with $N_{neighbors} = 5$, scaled again to account for changes in the data distribution after imputation and used to train five classifiers: Random Forest ($n = 10$, $max\_depth = 5$, Logistic Regression ($max\_iter = 250$), SVM, XGB and MLP ($max\_iter = 500$, $lr = adaptive$). This has been repeated by using PCR ($n \approx 0.95var$) prior to training the classifiers. Imputing with kNN was the best option as shown in Section 4.1.1. Since kNN is distance based, data variables have to be scaled before to asses different scales in different variables. It is also necessary to scale data in numerically non-robust pipelines for instance when using SVM or MLP. After imputation, the underlying data distributions might be manipulated. Scaling after imputation mitigates those influences again.

We reused metrics proposed by other works to be able to compare our results directly [Dai et al., 2020; Zhao et al., 2020; Li et al., 2020b]. In addition to Accuracy, Recall and ROC–AUC (One vs. One) we added Balanced Accuracy. Accuracy is a non-optimal metric for imbalanced datasets, not taking class weights into account and thus tending to provide too optimistic results for classifiers that are biased towards the majority class.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Balanced accuracy takes this into account. Results are documented in Figure 4.13.

The results paint a clear picture. Undersampling did not have a good effect overall. Other undersampling strategies (Near-Miss) were tried here as well which did not lead to improvements, so they are not documented here. Since ROC–AUC is also reported by previous works, we follow that and choose the best classifier for that particular metric. In general, a high ROC–AUC is a good in-between metric. A higher balanced-accuracy score would indicate that we are most likely possible to get the classifier to predict the underrepresented classes. This is especially true when we also look at the recall score, that might be lower in that case, which would indicate less good performance in the majority class. For instance, notice SVM-balanced weights and SVM without - they have a comparable ROC–AUC score but they differ quite in balanced accuracy and recall/accuracy, as those accuracy and recall score better when the majority class got predicted better.

Summary

Overall, PCA only had a small positive impact on some of the results. The best classifier overall was SVM with balanced class weights, which had the highest ROC–AUC tied with XGBoost. SVM was superior though in balanced-accuracy and had a lower standard deviation over the training folds. SVM was trained again with k-fold cross-validation to tune for optimal hyper-parameters and is finally evaluated using the validation split. The results are presented in detail in Section 5.1.2.

4.4 Medical Image Data

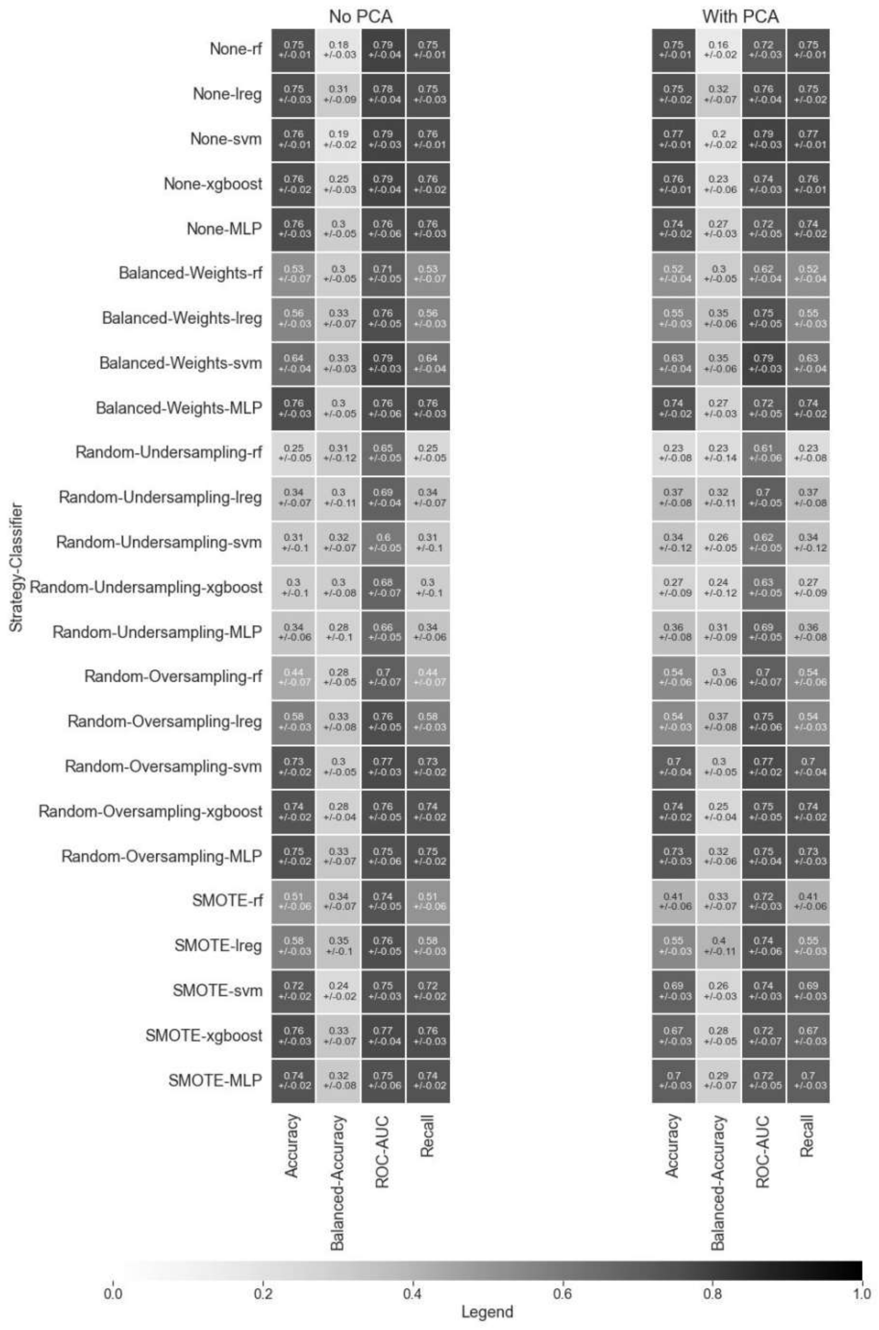
This section describes the process of working with the CXR data. First, we compare different encoder-decoder networks in form of U-net derivations for automatic lung segmentation by using transfer learning. We also compare different image preprocessing steps and evaluate them qualitatively and quantitatively, using a *reverse transfer learning* approach inspired by the *reverse accuracy* proposed by Valindria et al. [2017]. Automatically segmented images are used to extract radiomics features using *PyRadiomics*. Finally, multiple classifiers are trained similar to the clinical data approach presented in Section 4.3.2. The whole process is visualized in Figure 4.14.

The temporal aspects of the data have been ignored for the prediction and thus, images have been treated as individual data records rather than a series of records. Temporal aspects are only used in the context of visual analytics and data exploration later.

4.4.1 Transfer Learning

In order to train a segmentation network from publicly available masked CXR images including the lungs, the focus was set on finding and applying previously proposed architectures in the context of transfer learning. As mentioned in Section 3.2.3, two possible publications were found that are focused on transfer learning and have been

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



Strategy-Classifier	No PCA Accuracy	No PCA Balanced-Accuracy	No PCA ROC-AUC	No PCA Recall	With PCA Accuracy	With PCA Balanced-Accuracy	With PCA ROC-AUC	With PCA Recall
None-rf	0.75 +/-0.01	0.18 +/-0.03	0.79 +/-0.04	0.75 +/-0.01	0.75 +/-0.01	0.16 +/-0.02	0.72 +/-0.03	0.75 +/-0.01
None-lreg	0.75 +/-0.03	0.31 +/-0.09	0.78 +/-0.04	0.75 +/-0.03	0.75 +/-0.02	0.32 +/-0.07	0.76 +/-0.04	0.75 +/-0.02
None-svm	0.76 +/-0.01	0.19 +/-0.02	0.79 +/-0.03	0.76 +/-0.01	0.77 +/-0.01	0.2 +/-0.02	0.79 +/-0.03	0.77 +/-0.01
None-xgboost	0.76 +/-0.02	0.25 +/-0.03	0.79 +/-0.04	0.76 +/-0.02	0.76 +/-0.01	0.23 +/-0.06	0.74 +/-0.03	0.76 +/-0.01
None-MLP	0.76 +/-0.03	0.3 +/-0.05	0.76 +/-0.06	0.76 +/-0.03	0.74 +/-0.02	0.27 +/-0.03	0.72 +/-0.05	0.74 +/-0.02
Balanced-Weights-rf	0.53 +/-0.07	0.3 +/-0.05	0.71 +/-0.05	0.53 +/-0.07	0.52 +/-0.04	0.3 +/-0.05	0.62 +/-0.04	0.52 +/-0.04
Balanced-Weights-lreg	0.56 +/-0.03	0.33 +/-0.07	0.76 +/-0.05	0.56 +/-0.03	0.55 +/-0.03	0.35 +/-0.06	0.75 +/-0.05	0.55 +/-0.03
Balanced-Weights-svm	0.64 +/-0.04	0.33 +/-0.03	0.79 +/-0.03	0.64 +/-0.04	0.63 +/-0.04	0.35 +/-0.06	0.79 +/-0.03	0.63 +/-0.04
Balanced-Weights-MLP	0.76 +/-0.03	0.3 +/-0.05	0.76 +/-0.06	0.76 +/-0.03	0.74 +/-0.02	0.27 +/-0.03	0.72 +/-0.05	0.74 +/-0.02
Random-Undersampling-rf	0.25 +/-0.05	0.31 +/-0.12	0.65 +/-0.05	0.25 +/-0.05	0.23 +/-0.08	0.23 +/-0.14	0.61 +/-0.06	0.23 +/-0.08
Random-Undersampling-lreg	0.34 +/-0.07	0.3 +/-0.11	0.69 +/-0.04	0.34 +/-0.07	0.37 +/-0.08	0.32 +/-0.11	0.7 +/-0.05	0.37 +/-0.08
Random-Undersampling-svm	0.31 +/-0.1	0.32 +/-0.07	0.6 +/-0.05	0.31 +/-0.1	0.34 +/-0.12	0.26 +/-0.05	0.62 +/-0.05	0.34 +/-0.12
Random-Undersampling-xgboost	0.3 +/-0.1	0.3 +/-0.08	0.68 +/-0.07	0.3 +/-0.1	0.27 +/-0.09	0.24 +/-0.12	0.63 +/-0.05	0.27 +/-0.09
Random-Undersampling-MLP	0.34 +/-0.06	0.28 +/-0.1	0.66 +/-0.05	0.34 +/-0.06	0.36 +/-0.08	0.31 +/-0.09	0.69 +/-0.05	0.36 +/-0.08
Random-Oversampling-rf	0.44 +/-0.07	0.28 +/-0.05	0.7 +/-0.07	0.44 +/-0.07	0.54 +/-0.06	0.3 +/-0.06	0.7 +/-0.07	0.54 +/-0.06
Random-Oversampling-lreg	0.58 +/-0.03	0.33 +/-0.08	0.76 +/-0.05	0.58 +/-0.03	0.54 +/-0.03	0.37 +/-0.08	0.75 +/-0.06	0.54 +/-0.03
Random-Oversampling-svm	0.73 +/-0.02	0.3 +/-0.05	0.77 +/-0.03	0.73 +/-0.02	0.7 +/-0.04	0.3 +/-0.05	0.77 +/-0.02	0.7 +/-0.04
Random-Oversampling-xgboost	0.74 +/-0.02	0.28 +/-0.04	0.76 +/-0.05	0.74 +/-0.02	0.74 +/-0.02	0.25 +/-0.04	0.75 +/-0.05	0.74 +/-0.02
Random-Oversampling-MLP	0.75 +/-0.02	0.33 +/-0.07	0.75 +/-0.06	0.75 +/-0.02	0.73 +/-0.03	0.32 +/-0.06	0.75 +/-0.04	0.73 +/-0.03
SMOTE-rf	0.51 +/-0.06	0.34 +/-0.07	0.74 +/-0.05	0.51 +/-0.06	0.41 +/-0.06	0.33 +/-0.07	0.72 +/-0.03	0.41 +/-0.06
SMOTE-lreg	0.58 +/-0.03	0.35 +/-0.1	0.76 +/-0.05	0.58 +/-0.03	0.55 +/-0.03	0.4 +/-0.11	0.74 +/-0.06	0.55 +/-0.03
SMOTE-svm	0.72 +/-0.02	0.24 +/-0.02	0.75 +/-0.03	0.72 +/-0.02	0.69 +/-0.03	0.26 +/-0.03	0.74 +/-0.03	0.69 +/-0.03
SMOTE-xgboost	0.76 +/-0.03	0.33 +/-0.07	0.77 +/-0.04	0.76 +/-0.03	0.67 +/-0.03	0.28 +/-0.05	0.72 +/-0.07	0.67 +/-0.03
SMOTE-MLP	0.74 +/-0.02	0.32 +/-0.08	0.75 +/-0.06	0.74 +/-0.02	0.7 +/-0.03	0.29 +/-0.07	0.72 +/-0.05	0.7 +/-0.03

Figure 4.13: Medical data prediction results for multiple metrics and pipelines. Confidence interval (±) is given by standard deviation over cross validation folds. XGBoost classifier implementation did not have a balanced-weights setting and has no results in this case.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

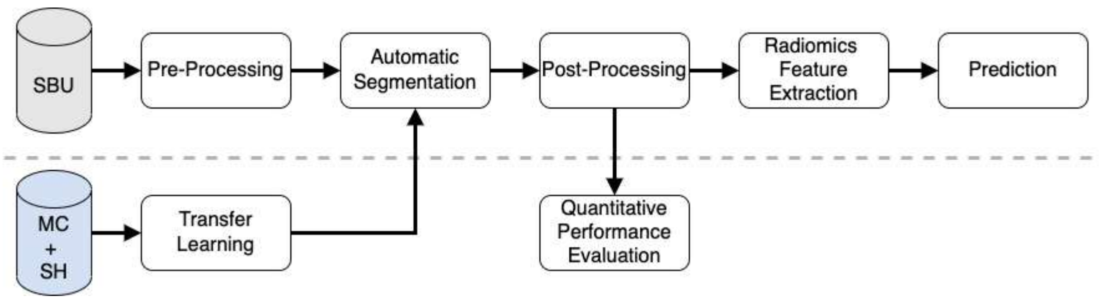


Figure 4.14: Visual representation of working with the medical image data.

used in the context of lung segmentation [Islam and Zhang, 2018; Iglovikov and Shvets, 2018]. *TernausNet* by Iglovikov and Shvets [2018] was originally designed for background segmentation of car images but has been used for other applications like segmenting lungs [Ovcharenko, 2019]. It can be initialized with a pre-trained network in a warm-start scenario. The authors show that using a VGG11 model that was pre-trained on ImageNet showed promising results. Islam and Zhang [2018] make use of the openly available MC and SD datasets [Jaeger et al., 2014] to train a U-Net network specialized for lung segmentation.

We trained five models based on the MC and SD data solely. Given that the dataset used in this thesis is less clean than the data used for transfer learning, as it was taken using mobile radiography systems with patients admitted to ICU (compare Figure 2.2), we added additional rotational data augmentation during the training phase, with the goal to make the model more robust. In addition, random cropping, zooming, and shifting have been applied to the data during training to increase variance. We also tried *AdamW* optimizer rather than the proposed Adam to see if it has any positive implications. Additionally, an early stopping criterion has been implemented. Specific model characteristics are found in Table 4.4 and their respective training process in form of loss and metric curves can be found in the appendix (Figures A.1, A.2, A.3, A.4, A.5).

From the loss curves and metrics curves during training of the networks found in the appendix, we observe multiple findings. Early stopping was a good tool to decrease training time without hindering training quality. *AdamW* optimizer had a negative impact on the training, resulting in a less stable optimization flow Figure A.2 of the loss. Rotation augmentation had a serious impact on the validation metrics and the training time for training on the TernaususNet design, hence Figure A.1 versus Figure A.4 for instance.

4.4.2 Image Preprocessing

As shown in Figure 2.2, images from the training datasets differ in terms of style and contrast from the ones in the SBU dataset. The datasets used for training the transfer learning model are way cleaner. In order to make the available X-ray data more compatible with the training data, some traditional image preprocessing techniques are applied. The

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Name	Design	Rotation	Optimizer	Early Stopping
Model A	TernausNet VGG11	None	Adam	None
Model B	TernausNet VGG11	+/- 30	AdamW	10%
Model C	TernausNet VGG11	+/- 30	Adam	10%
Model D	TernausNet VGG11	+/- 25	Adam	20%
Model E	U-Net blueprint	+/- 25	Adam	20%

Table 4.4: Deep learning models trained on Montgomery County and Shenzen Lung data with various additions. TernaususNet architecture as proposed by Iglovikov and Shvets [2018], U-Net architecture proposed by Islam and Zhang [2018].

resulting segmentation model should be able to ignore artifacts such as tubes and cables that are present in the images.

Multiple approaches are documented in the literature on improving contrast in chest X-ray images. Norval et al. [2019] measure different techniques prior to training a neural network. They included contrast enhancement, histogram, and adaptive histogram equalization in their work, besides region of interest selection, with the latter being the best performing one for the purpose of training a network. Udeshani et al. [2011] use median filtering prior to histogram equalization in order to enhance contrast. They also continue with feature extraction techniques also used in PyRadiomics on the preprocessed images, namely using GLCM (see Section 3.2.4).

Figure 4.15 shows three preprocessing strategies prior to segmenting and the respective segmented outcome using transfer learning model A (Table 4.4). The results show smoother edge detection around the lungs for adaptive histogram equalization, while histogram equalization was not always able to detect two separate lungs, hence the results for Sample 3. Most robust results towards handling cables and tubes are experienced on adaptive histogram equalization as well. Compare Sample 4 in Figure 4.15 for example, with multiple cables exiting the lung area. The segmented region is increased without preprocessing of the image, and adaptive histogram equalization is able to reduce this effect. But this also results in additional artifacts in some cases. Segmented masks are thus cleaned from additional segmented regions using contour detection and only keeping the two biggest regions. This reduces the impact of additionally segmented regions outside of the lungs, as long as two different big regions are still recognized. This process is explained in more detail in Section 4.4.5

Further enhancements were added to the adaptive histogram equalization which is illustrated in Figure 4.16. Adding median and blur filtering before adaptive histogram equalization furthermore increased robustness and reduced artifacts. Different hyper-parameter settings were tried for the Gaussian blur, median filtering as well as the adaptive histogram equalization from the *skimage* library [Van der Walt et al., 2014]. Experiments showed that the median filtering was approximately two times slower than the gaussian blur filter without considerable improvements when tested on a small

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

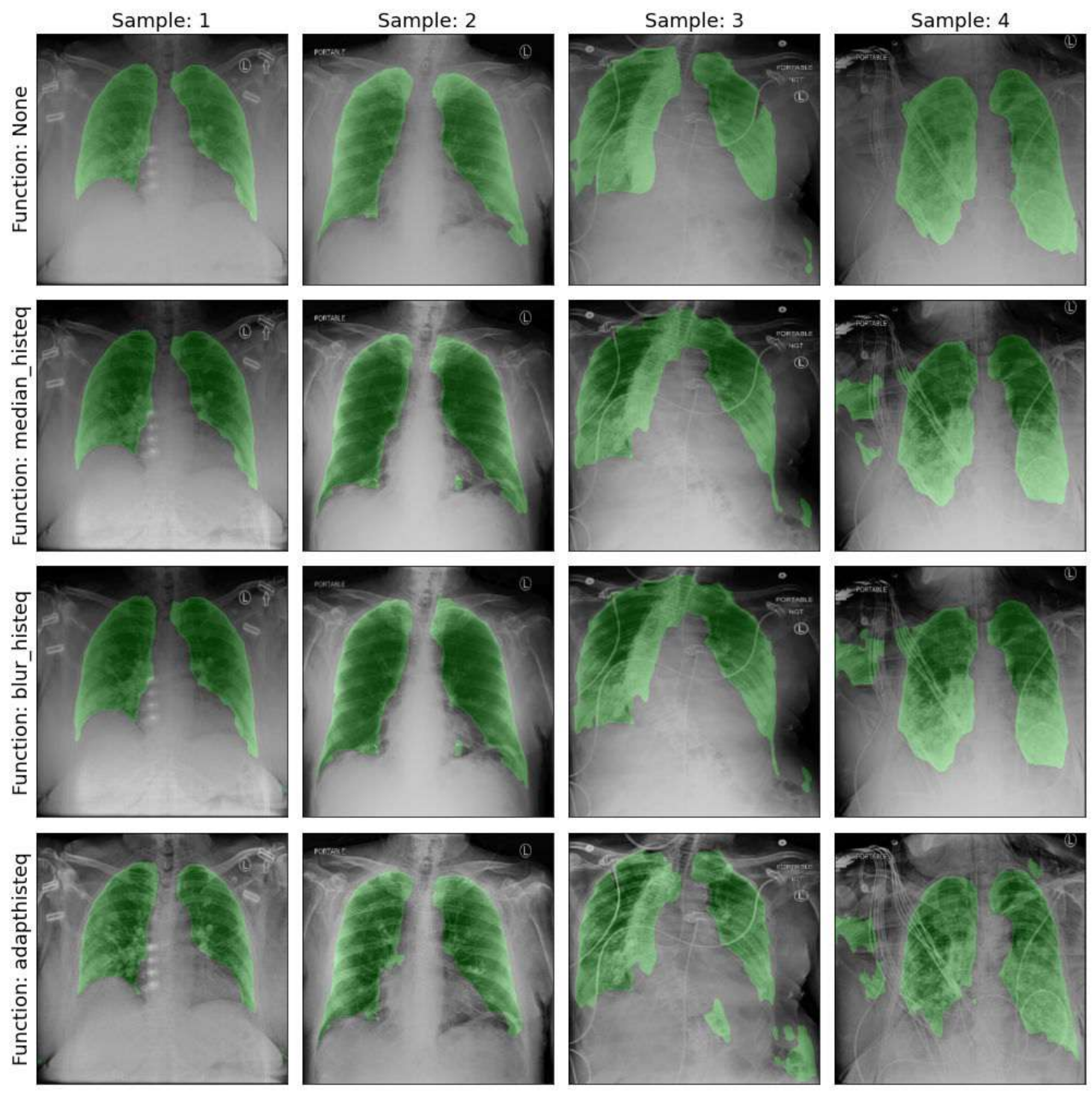


Figure 4.15: Using different image preprocessing techniques prior to segmentation: No preprocessing, median filtering + histogram equalization, Gaussian blur + histogram equalization, adaptive histogram equalization. Ideally the segmentation is not utterly confused by cables as observed in sample 3 and 4, but is able to ignore the artifacts without increasing the segmentation region. Additional smaller regions outside are removed in a post-processing step, only keeping the two largest areas. Green indicates segmentations.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

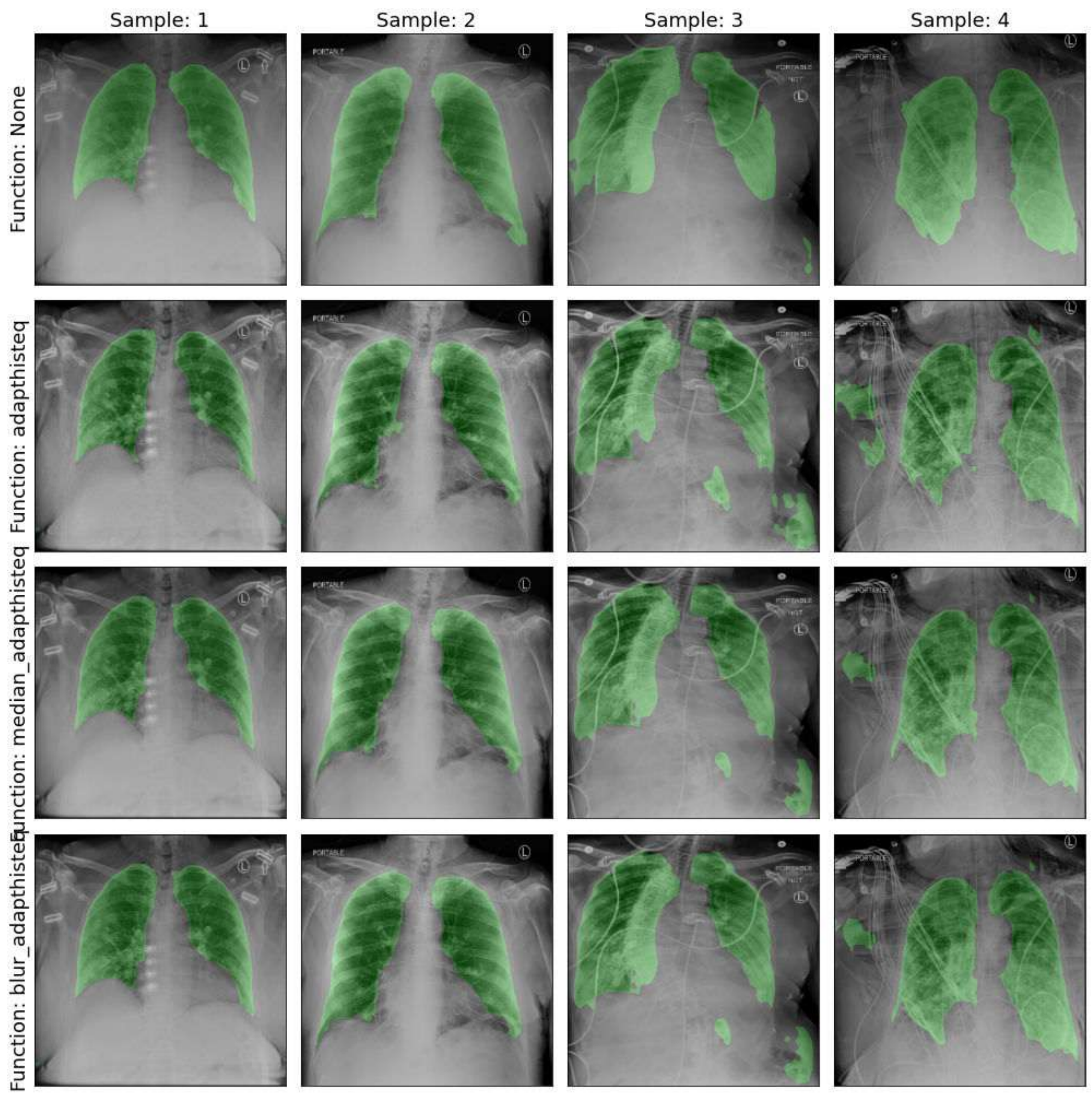


Figure 4.16: Different image preprocessing techniques continued. Tuning adaptive histogram equalization by using local averaging methods. Green indicates segmentations.

sub-sample, therefore the Gaussian blur was deemed to be the better choice as the segmentation is meant to be carried out in an interactive way where longer processing speeds may influence user experience in a negative way [Munzner, 2009].

4.4.3 Model Evaluation - Qualitative

Finally, a model has to be chosen as the best-performing one given the available data. Figure 4.17 shows results for the five models described in Table 4.4 without image preprocessing. Using the *AdamW* optimizer did not have a high positive impact in this

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

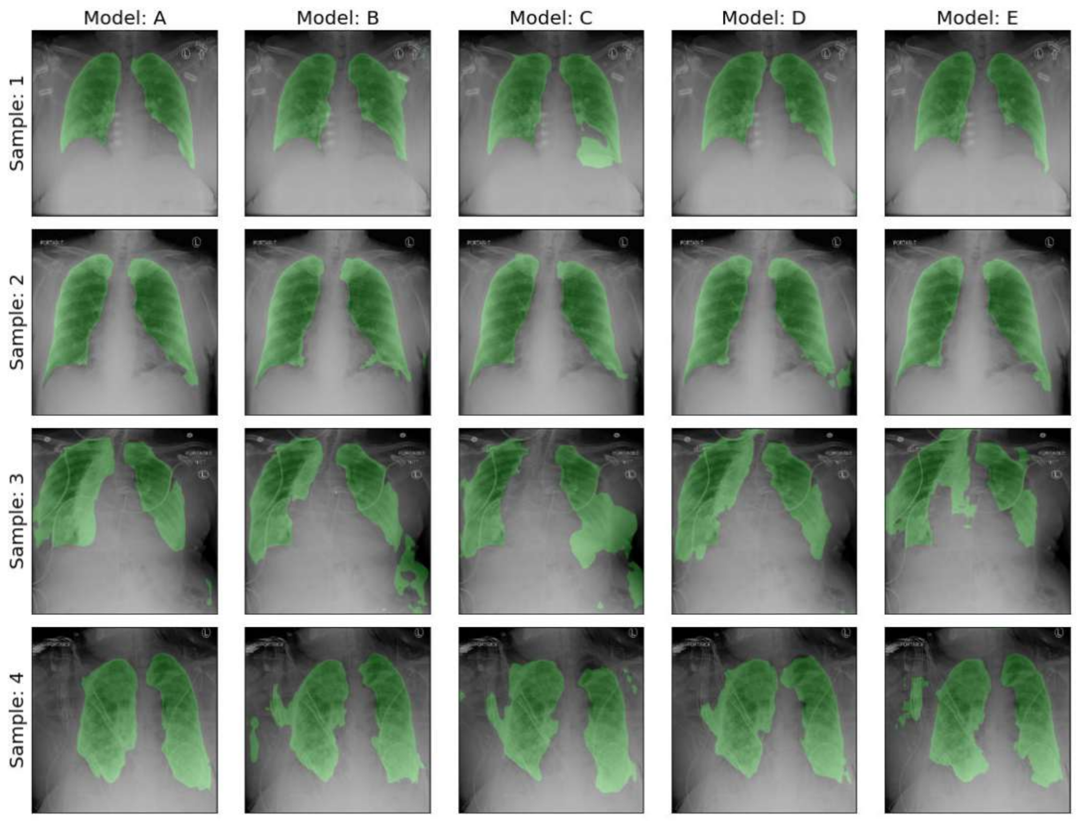


Figure 4.17: Comparison of segmentation results for models described in Table 4.4 without image preprocessing. Green indicates segmentations.

particular case [Loshchilov and Hutter, 2017]. The training loss was also observed to be unstable, as documented in Figure A.2 in the appendix section. Rotations of ±25 degree had a more robust effect on the outcome of the segmentation (Model D, E) when compared to ±30 (Model B, C). Model E arguably had the most problems separating the lung from the neck (Sample 3) but was promising in other areas with smooth regions for more clearly distinguishable samples (Sample 1+2).

Results after adding the previously proposed image processing techniques are shown in Figure 4.18. Generally, all models are more robust when handling cable artifacts while introducing increased issues separating the lung from the neck on one sample. Sample 3 gives a good overall feeling on how models would perform on the highly variable data. Model B, C, and D are unable to reliably tolerate cable artifacts, while Model E tends to include the neck area, which can also be seen in Sample 1. *Model E* and *Model A* record the smoothest detection in general.

Repeating this qualitative assessment for all samples in the dataset would be too much work and it is also already visible that there is no perfect solution at hand, ultimately

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

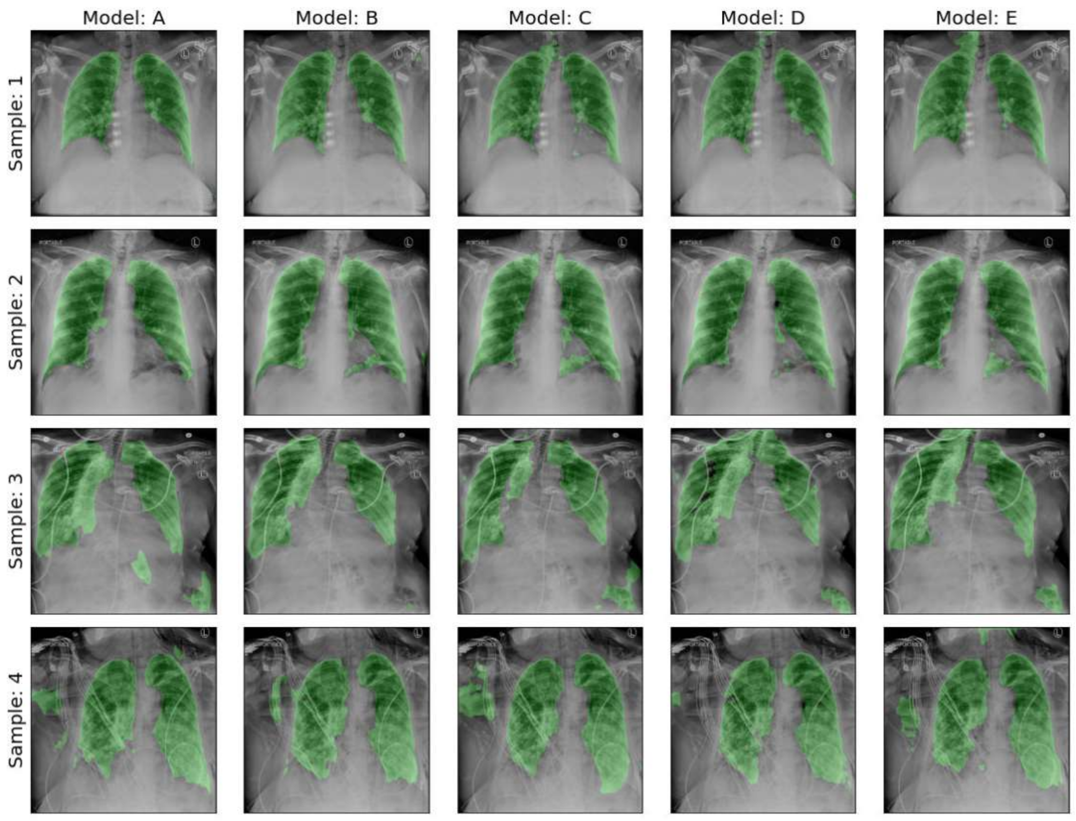


Figure 4.18: Comparison of segmentation results for models described in Table 4.4 with Gaussian blur and adaptive histogram equalization. Hyperparameters used: Gaussian blur $\sigma = 1$, histogram equalization *clippling_limit* $= 0.01$. Green indicates segmentations.

leading to a trade-off at some point between robustness towards artifacts and keeping the regions as small as possible. It is also worth noting here that, in order to achieve reproducible results over multiple runs for the segmentation, it is of the highest importance to use manual random seeds everywhere possible, most importantly for PyTorch, to make results comparable over consecutive segmentations.

4.4.4 Model Evaluation - Quantitative

In order to make a well-founded decision about the performance of the numerous segmentation models and preprocessing strategies employed, a quantitative assessment was designed. Inspiration for the approach used in the thesis stem from the concept proposed by Valindria et al. [2017] already described in section 3.4.2. While the reverse accuracy they proposed is based on a model trained on only one sample that is presented to an online-segmentation system, the data available in this thesis allows for the creation of many reference segmentations to train a new model. Using parts of the original training datasets MC and SD with ground truth available as a validation set, the usage of standard

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

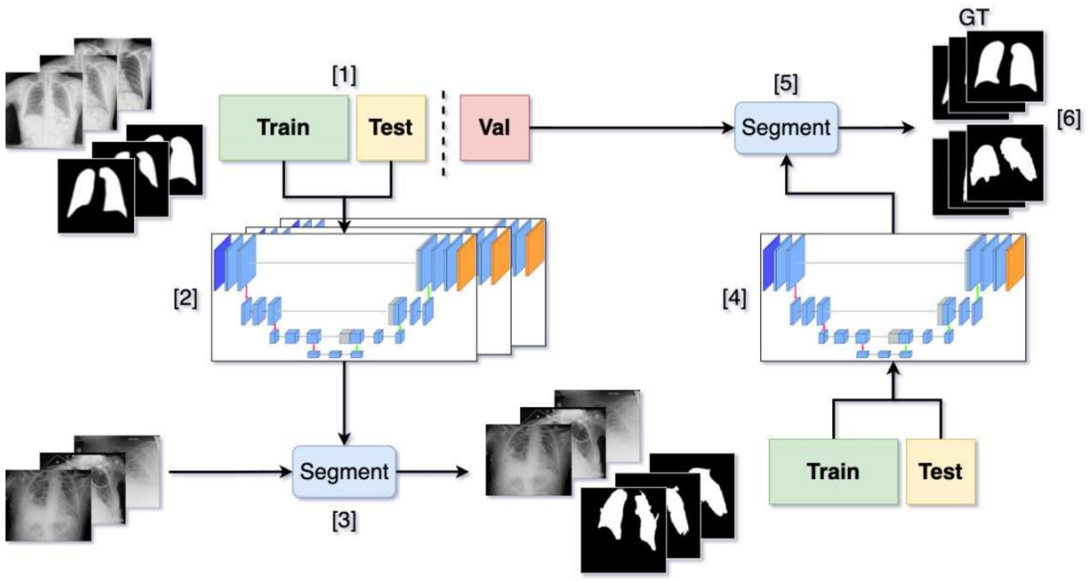


Figure 4.19: Model evaluation strategy inspired by Valindria et al. [Valindria et al., 2017]. [1] MC and SD dataset is divided into train, test, and validation split. [2] Train and test split are used to train multiple CNN models shown in Table 4.4. [3] Models are used to generate segmentation masks for the SBU COVID-19 dataset. [4] Created masks and raw images are used to train new networks, using the design from Model A. [5] Remaining validation split from the original training data is used which was ignored in the training process of the first bunch of models in [2]. Models from [4] are used to create segmentation masks for the validation split, where ground truth data in form of segmentation from human experts is available. [6] Since ground truth data is available here, usage of standard metrics (Jaccard, Dice) is possible again.

metrics like Dice and Jaccard is enabled again. Figure 4.19 visualizes this process. We use the term *reverse transfer learning* for this process.

Summary

The models deemed as most promising from qualitative assessment — models A and E — were used with and without preprocessing (gaussian blur combined with adaptive histogram equalization) to create segmentation masks for the SBU dataset. Based on the masks and images from the SBU dataset, another CNN was trained. Only a random subset of images from the SBU dataset was used due to the high amount of disk space necessary to store multiple versions of all segmentations. Again data augmentations such as cropping and rotation are used for training. The architecture chosen for this approach was the TernausNet (Model A). Models have been validated on the validation set ($n = 141$) that has been left out during the whole original transfer learning process. Final results are presented in Section 5.1.3 in detail.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

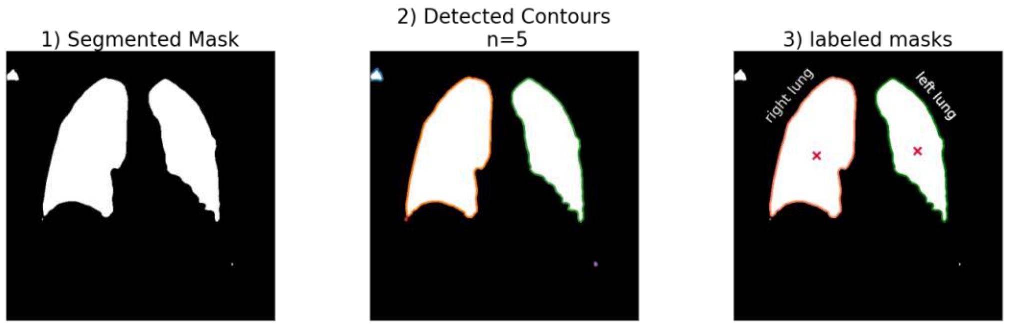


Figure 4.20: [1] Results from automatic segmentation which may include artifacts. [2] Separate areas are detected using contour detection. [3] The two biggest areas are selected and the center is calculated using the contour points coordinate information. The left area is defining the right lung and vice versa.

4.4.5 Feature Extraction

In the next step, *PyRadiomics* is used to extract radiomics features (Section 3.2.4) from the chest X-rays and their respective segmentation masks. First, the masks are cleaned, removing additional obsolete segmented areas by the automatic segmentation. Contour detection from *skimage* is used on binarized gray scale masks. The two biggest contours are selected, based on the number of points defining the contours, where a higher number indicates a bigger contour. Pixel-space point coordinates defining the outline of one contour are averaged for simple estimation of the center location of the respective contour, in order to associate if the area defines the right or left lung, where the left contour determines the right lung and vice versa. This procedure is illustrated in Figure 4.20. Combined images of the respective lungs are saved as well for quick visualization as shown in Figure 4.21. Those results have also been used for the modified reverse accuracy approach presented in Section 4.4.4.

Dealing with Multiple ROIs

Following mask creation, *PyRadiomics* can be used in several ways to extract features. According to PyRadiomics developer guidelines, only one ROI should be used per extraction run [van Griethuysen, 2020a,b]. While PyRadiomics is able to deal with masks containing multiple labels in an image, in form of different organs, for instance, it only extracts one label per run. When dealing with different organs and differently labeled ROIs, it would make sense to use different runs for feature extraction since the organs might appear different in the original image due to different tissues and thus, ultimately influencing feature extraction. When using the same organ type though, in our case two separate lungs, it is not completely clear how one should deal with this and there are multiple possibilities now:

(i) Extract features for the left and the right lung in separate runs. Those extracted

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

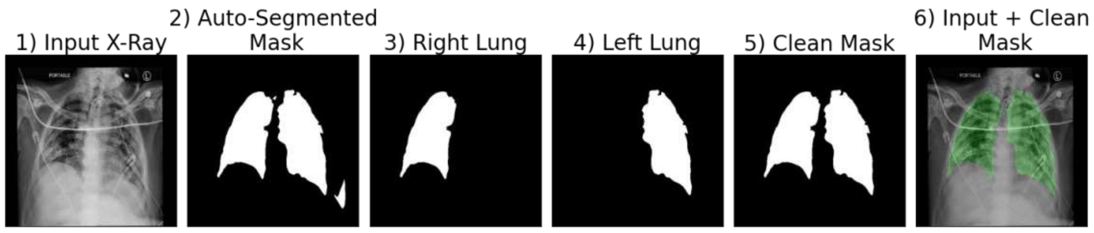


Figure 4.21: 1) Resized input X-ray image. 2) Automatically segmented image using transfer learning. Potentially resulting in more than two areas, as shown in this example. 3+4) Right and left lung is separated from the mask as shown in Figure 4.20 5) Left and right lung are combined to define the final mask. 6) Resized input X-ray with the cleaned segmentation is blended on top.

features could be used to train a predictor for each lung instead of each patient's image. This would make the prediction and the presentation of the results more complex.

(ii) Another way of handling this is extracting features per lung and stacking them for each image, doubling the dimensionality. This would be in line with developer guidelines and the prediction process would not be more complex.

(iii) Finally one can also run the feature extraction with the complete mask including both lung masks. *PyRadiomics* does not throw any errors or warnings in this case despite the guidelines. When testing this behavior on First Order features and comparing it to the outcome of separate runs for each lung respectively, one could identify somewhat of a weighted average being returned, depending on the size of the ROI. This makes sense for First Order statistics, as they are only statistical features such as the mean or median of gray-level pixel values in the mask. Repeating the process for other feature classes relying more on the contextual information of one region, such as shape features, one could not experience such simple effects.

In any case, there is no information in related papers on how this was dealt with when working with this type of data [Tamal et al., 2021; Han et al., 2021]. For the sake of completeness, we tried multiple ways. Results are improving from per-lung extraction, to combined features and finally are best with stacked features. This is shown in Table 4.5. Results were created using an SVM-classifier with a 70/30 train-test split without hyper-parameter tuning. High correlated features ($corr > 0.95$) were removed. As it also makes the most sense to stack features and complies with the developer guidelines we continue with using the stacked approach going forward.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

	Accuracy	Balanced-Accuracy	ROC-AUC	Recall
Separated	0.399	0.292	0.723	0.399
Stacked	**0.483**	0.330	**0.766**	**0.483**
Combined	0.436	**0.341**	0.754	0.436

Table 4.5: Results for the different methods of extracting radiomics features.

4.4.6 Radiomics Feature Exploration

Radiomics feature extraction with stacked features for all feature classes used leads to a total of 204 features with 102 for each lung. When comparing to the official docs and previously mentioned feature classes available in Section 3.2.4, 104 features are available. In fact, two are disabled by default as they are correlated with others. This is the case for standard deviation (correlated to variance) in First-Order Statistics and spherical disproportion (correlated to sphericity) in 2D-Shape [Radiomics, 2022].

Figure 4.22 displays the correlation between radiomics features before removing highly correlated features. Feature classes are color-coded on the left and top borders. The heat map is divided into four visible quadrants. The left-upper one depicts the features for the left ROI in the mask, and the right-lower one depicts correlations for the right ROI. Features classes can be grasped by strong correlations inside their feature class making up smaller quadrants in the heat-map as well, indicated by the colors on the top and the left side of the plot. Here, First Order as well as shape based features seem to translate between the ROIs, yielding medium correlations. 2D-shape features seem less correlated to other feature classes in general, whereas the latter classes are showing intra-class correlations at a high degree, hence GLCM and GLRLM for instance.

4.4.7 Prediction using Radiomics Features

Highly correlated features that can be defined as redundant have been removed. This was done using a threshold of $corr > 0.95$ as proposed by Chen et al. [2017]. Data were split into train and validation sets. The training data split was used in 10-fold cross-validation, as multiple pictures are available for patients, especially for those with worse disease progression. The same classifiers and metrics were used as reported in Section 4.3.2.

In Figure 4.23 radiomics features correlation with clinical outcomes are displayed. Features are sorted ascending in regards to a less severe case, the hospitalization stays without ventilation, ICU admission or mortal outcome. Again, features negatively correlated with a less severe disease progression are indicators for a severe progression in most cases. Outcomes with lower N correlations are not that clearly visible, and also not clearly distinguishable between the classes. Many negative and positive correlated features for a less severe hospitalization share their correlation tendencies with ICU admissions.

Figure 4.24 reports the classifier performances on radiomics features. The best performing classifier was again SVM with balanced weights as the imbalanced learn strategy. Random

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

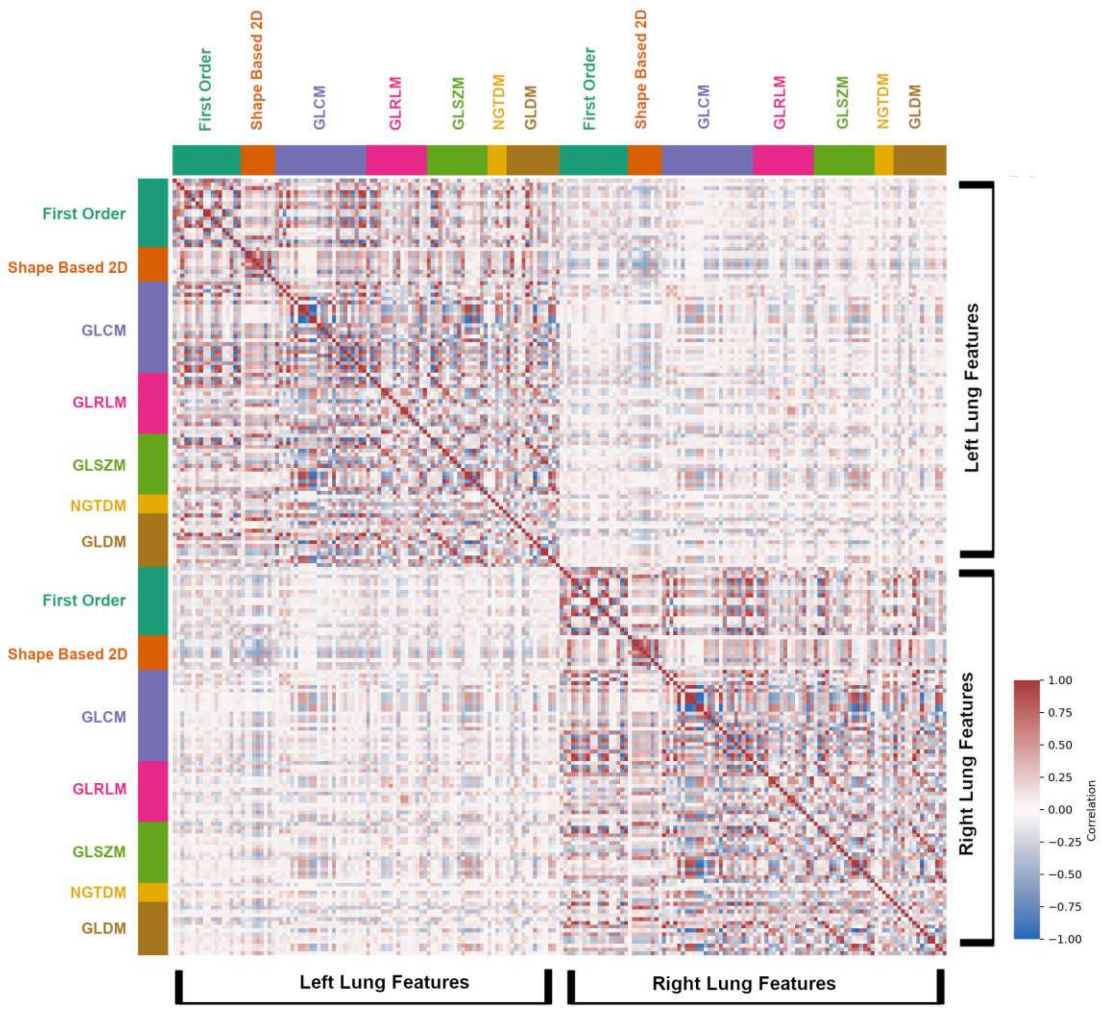


Figure 4.22: *PyRadiomics* features correlation heat map showing both ROIs of the images, the left and the right lung. Feature classes are highlighted on the top and left side of the map.

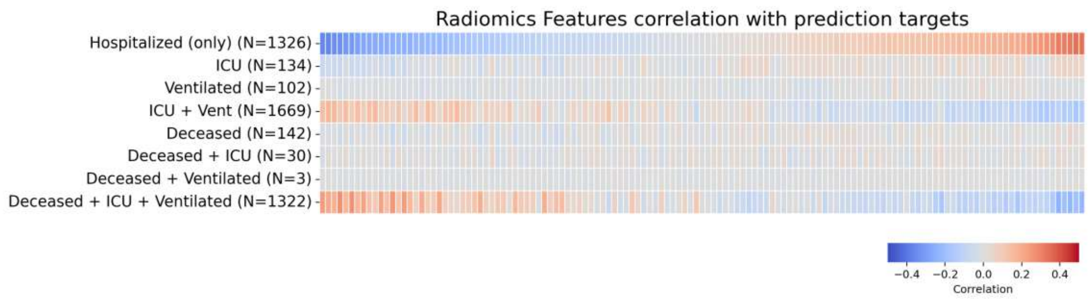


Figure 4.23: Radiomics features correlation heat map with hospitalization outcomes.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

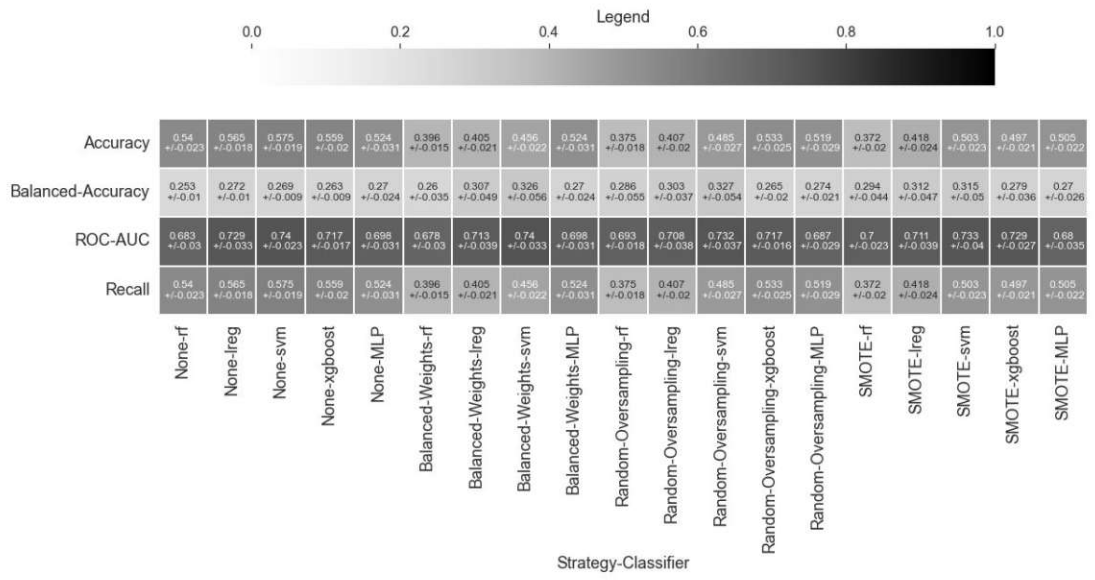


Figure 4.24: Classifier performances on 10-fold cross-validation for radiomics features.

undersampling yielded the worst results again and thus is not shown in the figure. When compared to results from medical data prediction (Figure 4.25, results are slightly worse. Hyper-parameter tuning for SVM was done again using cross-validation. Final results are presented in Section 5.1.3.

4.5 Combining Features

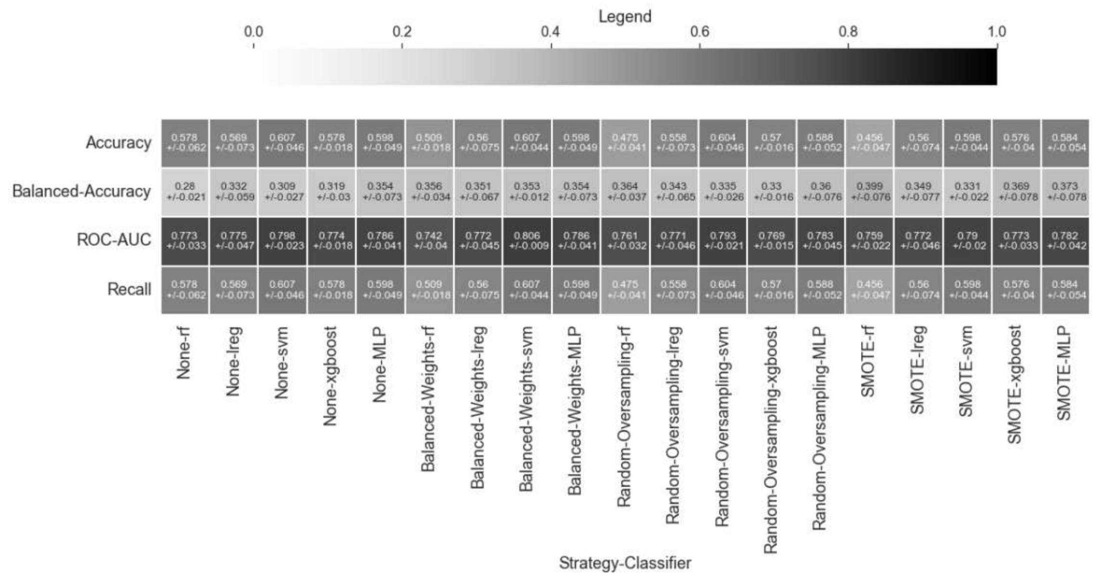
The most simple approach to combining features is to add the per-patient medical health records to the extracted image data. For each feature set the same preprocessing was used as before when handling the feature sets separately. For radiomics features, high-correlated ones ($c_{pearson} > 0.95$) were removed.

Data splitting is not as straightforward as before. Randomly dividing into separate datasets as done for the previous prediction tasks is not feasible since it would lead to information leakage. Medical data is present in the combined feature set several times for patients with several images. If not prevented, patients will be present inside training and test splits which will lead to information leakage, as the model most likely will remember training records. This leads to overfitting models and therefore, instead of randomly sampling from the complete dataset, the dataset has to be divided by patients instead.

To deal with that, we used a grouping K-Fold split instead, with one patient being

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



	None-rf	None-lreg	None-svm	None-xgboost	None-MLP	Balanced-Weights-rf	Balanced-Weights-lreg	Balanced-Weights-svm	Balanced-Weights-MLP	Random-Oversampling-rf
Accuracy	0.578 +/-0.062	0.569 +/-0.073	0.607 +/-0.046	0.578 +/-0.018	0.598 +/-0.049	0.509 +/-0.018	0.56 +/-0.075	0.607 +/-0.044	0.598 +/-0.049	0.475 +/-0.041
Balanced-Accuracy	0.28 +/-0.021	0.332 +/-0.059	0.309 +/-0.027	0.319 +/-0.03	0.354 +/-0.073	0.356 +/-0.034	0.351 +/-0.067	0.353 +/-0.012	0.354 +/-0.073	0.364 +/-0.037
ROC-AUC	0.773 +/-0.033	0.775 +/-0.047	0.798 +/-0.023	0.774 +/-0.018	0.786 +/-0.041	0.742 +/-0.04	0.772 +/-0.045	0.806 +/-0.009	0.786 +/-0.041	0.761 +/-0.032
Recall	0.578 +/-0.062	0.569 +/-0.073	0.607 +/-0.046	0.578 +/-0.018	0.598 +/-0.049	0.509 +/-0.018	0.56 +/-0.075	0.607 +/-0.044	0.598 +/-0.049	0.475 +/-0.041

	Random-Oversampling-lreg	Random-Oversampling-svm	Random-Oversampling-xgboost	Random-Oversampling-MLP	SMOTE-rf	SMOTE-lreg	SMOTE-svm	SMOTE-xgboost	SMOTE-MLP
Accuracy	0.558 +/-0.073	0.604 +/-0.046	0.57 +/-0.016	0.588 +/-0.052	0.456 +/-0.047	0.56 +/-0.074	0.598 +/-0.044	0.576 +/-0.04	0.584 +/-0.054
Balanced-Accuracy	0.343 +/-0.065	0.335 +/-0.026	0.33 +/-0.016	0.36 +/-0.076	0.399 +/-0.076	0.349 +/-0.077	0.331 +/-0.022	0.369 +/-0.078	0.373 +/-0.078
ROC-AUC	0.771 +/-0.046	0.793 +/-0.021	0.769 +/-0.015	0.783 +/-0.045	0.759 +/-0.022	0.772 +/-0.046	0.79 +/-0.02	0.773 +/-0.033	0.782 +/-0.042
Recall	0.558 +/-0.073	0.604 +/-0.046	0.57 +/-0.016	0.588 +/-0.052	0.456 +/-0.047	0.56 +/-0.074	0.598 +/-0.044	0.576 +/-0.04	0.584 +/-0.054

Figure 4.25: Classifier performances after merging the feature sets.

one group. Folding is not done on a per row level but per patient. Again, a train and validation split was created in a stratified way, keeping the target labels at the same distribution. The train data was then split with Group K-Fold with $k = 4$.

Figure 4.25 shows the prediction results for classifiers on the test set. We observe a small increase from the results for medical prediction reported in Figure 4.13 after merging feature sets. Final results after Hyper-parameter tuning are presented in Section 5.1.3.

4.6 Environment

The implementation was mainly done in Python. For the deep learning part we used *PyTorch* while utilizing CUDA on a NVIDIA GTX 1060Ti 6GB. For classifier training, *scikit-learn* was used in addition to *imbalanced-learn* for working with imbalanced data. For working with medical image data, we utilized *pydicom* and *skimage* in addition to *PyRadiomics*. Random seeds are used in various libraries and implementations, where algorithms are randomly initialized, for example as laid out in Sections 3.4.4 and 3.4.5. Random seeds were always set to 42 if not specified otherwise. A conda environment dump containing the exact versions used is available online, for cross-platform availability without any build-number constraints [2] and, for the sake of complete

[2]No builds: `https://gist.github.com/oStritze/c9c05767ba4f42712341740888cf7a81`

TU WIEN Bibliothek Your knowledge hub

transparent documentation, also with build hashes where available [3].

A short overview of possible tools for the implementation of visual analytics solutions is given next.

Data Driven Documents D3: D3[4] is an extremely flexible and open-source JavaScript (JS) library. It is designed to modify and manipulate static images on the fly using Scalable Vector Graphics (SVG) by using them in an embedded way in HyperText Markup Language (HTML), styling them via Cascading Style Sheets (CSS). It is necessary to have some backend server which can be a Python Flask or NodeJS server. Advantages are, apart from its flexibility, the use of constantly manipulating SVG files which are flexibly scaled and platform- or browser-independent. Disadvantages include data registration and moving data from the back-end to front-end, computational limitations during running workload via JS in the browser, and additional complexity that comes with the huge landscape of opportunities given via the library. Interactivity is not provided by default for visualization types. This leads to more flexible solutions while being more work-intensive and complex.

Dash: Dash[5] is a visualization platform that combines React with Plotly[6]. It is available in multiple languages such as R, Python or Julia. By default, it is running on a self-managed Flask backend server. It abstracts away many tasks that accrue when working with D3 combined with a self-managed backend, such as data transportation between front and backend. Plotly is used as the main visualization library in dash, which includes a variety of interactive plots for different use-cases. Interactivity is defined by the library and some customization options are available. Specialized optional packages are available that provide visualization types for special research areas such as biological use-cases (dash-bio). Library pre-defined HTML and CSS objects can be created in python itself, including user-input widgets such as sliders or select-boxes. The python implementation makes working with typical machine learning libraries, such as sklearn, straightforward. Plotly already implements most common visualization best practices by default which saves time when implementing visualizations for developers. As a downside, much of the freedom that is available in D3 is missing, as visualization is limited to plot types available in Plotly. Those can be customized to a high degree though. One also has to deal with CSS and HTML-specific positioning and error handling when designing dashboards.

Streamlit: Streamlit[7] is another open-source library that is built on Typescript, Javascript, CSS and Python. It lets developers create dashboards using pure Python code without any HTML or CSS pre-knowledge. It naturally fits into the machine learning ecosystem, providing pre-defined interaction forms and wrappers for most popular plotting libraries, such as Bokeh, Altair, or previously mentioned Plotly. It is built for

[3]https://gist.github.com/oStritze/6e0146d682c2880e1d4116d9dfe947f1
[4]www.d3js.org
[5]https://plotly.com/dash/
[6]https://plotly.com/
[7]www.streamlit.io

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

interactive training, tuning, validation and exploring of datasets, not only in a tabular form but also in video and images, rendering data on the fly in the browser. Interaction in visualizations is provided by utilizing supported libraries (Plotly), as well as pre-defined widgets including popular user input options (select boxes, sliders, ...). Custom elements can also be created via HTML and CSS by using self-written components. Cross-visualization data manipulation is not straightforward but can be established. It is the least flexible and complex, but the most high-level and straightforward solution of the libraries presented here.

Given the manifold of tasks that are to be done and their analytical background, we choose streamlit as the front-end as it is the most natural way of working with the data science libraries used in the backend. It exceeds in building simple, interactive dashboards for just that without a steep learning curve. It also excels at handling image data in a simple way which will be an important part when displaying segmentation results. Given that it is also able to render Plotly charts, it was not seen as necessary to use dash. Since interactive data handling and manipulation of underlying data is a big part of the design, we also preferred streamlit over D3 given the simple python implementation.

4.7 Visual Analytics

In this section we will discuss how our visual analytics solution is designed and how design choices are affected by the task and requirements defined in Section 4.2. We discuss intermediate results as well as unused drafts while proposing alternative visualization techniques. Final dashboard results are presented in Section 5.2 of Chapter 5.

4.7.1 Task Implications

Visual analytics design choices are influenced by the task design discussed in Section 4.2. The general typology definition, as depicted in Figure 3.19, already indicates which parts of the design are influenced by which parts of the typology. Visualization and interaction techniques are influenced by *"how"* users are expected to use the visualization to achieve predefined tasks. How those tasks are solved is roughly defined with *"why"*, pointing out high and low level tasks that are to be solved by the visualization. The data that is to be visualized is also already defined by *"what"*.

User groups tasks are designed roughly in the same order. We start with an overview in all cases (*Medical 1*, *Analytics 1*, *General 1*), which is followed by more dedicated tasks where dedicated designs are considered. Later on, designs may be shared again, hence *Medical 4* and *Analytics 4* both focused on importing CXR images, or *General 2* and *Medical 4* both designed around encoding prediction presentation.

Therefore, we will present our task designs ordered by the high level task that was to be solved while focusing on specialities for each user group separately. We will argue which visualization techniques may fit the desired tasks by also mentioning potential

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

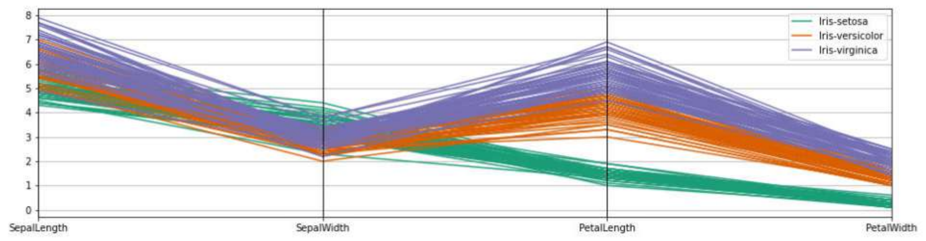


Figure 4.26: Parallel Coordinates plot for iris dataset.

alternatives, with respect to discussing why some alternatives might not be as suitable as others.

Overview First

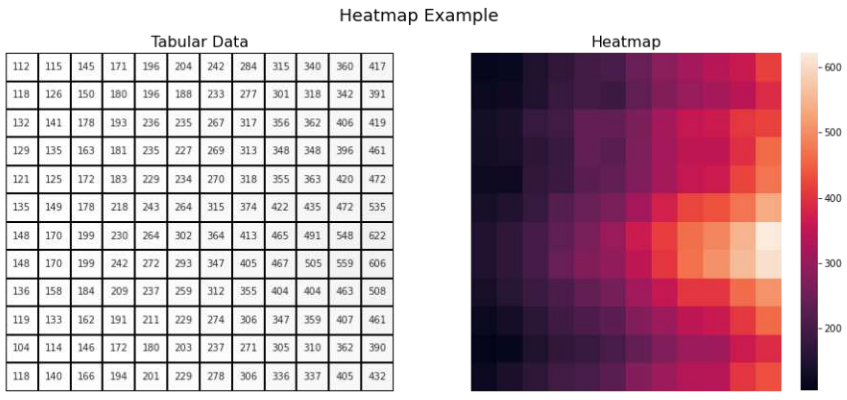
Medical experts require overview strategies to delve into the multivariate electronic health data, while analysts demand interactive solutions to customize views which enable them to *identify* features of high impact, also for high-volume radiomics features. Layman users need a condensed, highly abstracted version to gain a quick overview, without too much detail that might be otherwise confusing. This is fined in Section 4.2 as *Medical 1 + 2*, *Analytical 1 + 5* and *General 1.*

Interaction possibilities are key to enable *selection* and *filtering* (*Medical 1+2*, *Analytical 1*) which will make identifying outliers and patients of interest easier. Here the most capable user group should be given a lot of freedom. Also they might be interested in a more detailed view by providing options for *zooming.* For the general population, the data should not be presented in its raw form to avoid too complex views which could overburden potential users without medical or analytical knowledge.

The most common ways to gain an overview is to display data in form of scatterplots. Sarikaya and Gleicher [2017] argue that its relative simplicity and flexibility enables the scatterplot as *an ideal sandbox for early information visualization.* Furthermore, scatterplots let users solve a number of high to low level tasks, including identification, object location, comparison, exploration, search as well as characterization of distributions and identifying correlations, many of which are defined in the respective task definitions in Section 4.2. Scatterplots encode each observation as a point, often in a 2-D coordinate system. Points can be furthermore encoded by color or shape to make differentiation easier. On the contrary scatterplots are known to be prone to overfitting with many data records leading to overlapping points which may be counted with interaction possibilities such as brushing and zooming [Sarikaya and Gleicher, 2017].

Heatmaps are another common choice when visualizing connections in a dataset. They have been used in many applications focusing on multivariate data analysis, especially in high dimensional biological datasets used in medical research [Fernandez et al., 2017; Metsalu and Vilo, 2015]. Heatmaps provide designers with the possibility to display

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



112	115	145	171	196	204	242	284	315	340	360	417
118	126	150	180	196	188	233	277	301	318	342	391
132	141	178	193	236	235	267	317	356	362	406	419
129	135	163	181	235	227	269	313	348	348	396	461
121	125	172	183	229	234	270	318	355	363	420	472
135	149	178	218	243	264	315	374	422	435	472	535
148	170	199	230	264	302	364	413	465	491	548	622
148	170	199	242	272	293	347	405	467	505	559	606
136	158	184	209	237	259	312	355	404	404	463	508
119	133	162	191	211	229	274	306	347	359	407	461
104	114	146	172	180	203	237	271	305	310	362	390
118	140	166	194	201	229	278	306	336	337	405	432

Figure 4.27: An example of a heatmap, showing the data as tabular form and in the heatmap representation, illustrating how color coding helps in detecting areas of interest.

non-reduced data. Thus they strive in representation of dense datasets as 2-D matrices, such as correlation matrices, where correlation values can be encoded as colored tiles. Since static images with high dimension datasets tend to be hard to grasp, it is useful to design interactive versions allowing users to zoom, pan, search or select areas of interest [Fernandez et al., 2017]. With that they fit the task requirements introduced in Section 4.2 for *Medical 1* and *Analytical 1 + 5*, including *browse*, *compare*, *summarize* and *locate*, where *brushing* is known to further enhance those capabilities [Nusrat et al., 2019]. Figure 4.27 shows an example of a heatmap, as well as the tabular data which is the basis for the visualization.

Parallel coordinates [Inselberg, 1985] would be another popular option. In parallel coordinates, features are aligned on the x-axis while for each feature an axis is created. Each record in the dataset is then encoded as a line and is positioned according to its numerical value on each feature axis. An example is shown in Figure 4.26 using a minimal example for the iris dataset with three different classes. The classes can be color coded additionally, so each record can be associated with its label. Already with the example dataset in Figure 4.26, which only inherits 150 data records, overlapping can be noticed. Thus, without interaction techniques such as brushing or linking, parallel coordinates are deemed unable to successfully display a high amount of features [Heinrich and Weiskopf, 2013]. As the dataset considered for this tasks is consisting of hundreds of features (130 for EHD, 200 for radiomics features) and records (1279), parallel coordinates is not considered as an appropriate option here. Furthermore, the plot type is limited by different data types that need to be mapped to a common data scale [Heinrich and Weiskopf, 2013]. Also, parallel coordinate plots have a rather high learning curve and thus are not always clear to people that have never seen or worked with them before, as many different versions exist [Heinrich and Weiskopf, 2013].

For medical experts and analytical users we decided to keep the visualization design

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

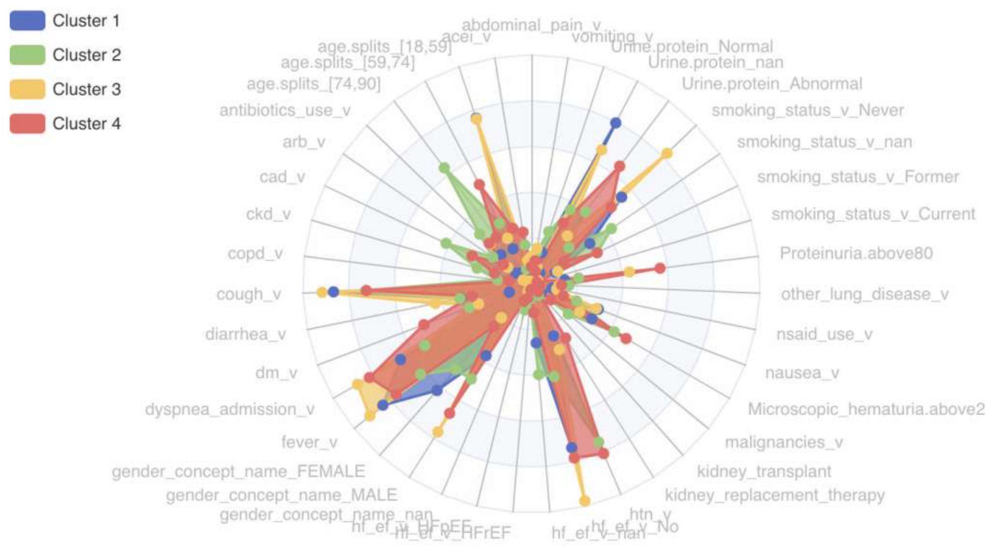


Figure 4.28: Radar plot for cluster comparison with mean features values per cluster. Too many variables or clusters lead to overplotting.

simple and opted for scatterplots and heatmaps, visualization types that should be well known to this user groups. This is once more not recommendable for the general population though, as those users might be able to read a simple scatterplot but it will not provide enough information to abstract enough data. It would not make much sense to include detailed medical domain variables, like blood pH values or C-reactive protein masses, which would be too specific for layman users. We instead chose to communicate the information as concisely as possible. The overview for the general population is thus done by communicating cluster differences, which is described in the next section.

Compare Clusters and Groups of Patients

Comparing clusters and groups of patients is a goal in tasks *Medical 1 + 2*, *Analytics 3.2* and *General 1*, again with different foci on the outcome. User with medical background are interested in groups of patients that are semantically considered groups by their preconditions or disease outcome. Identification is focused on certain patients of interest. Analytical users are interested in the comparison of clusters after unsupervised learning. The question here is in which ways the clustering is affecting features and how those clusters ultimately differ. For the general population, the clustering is done in the background and should help them grasp differences between the groups of patients by abstracting information in a way that makes understanding those differences fast and easy.

Medical Experts: For medical experts, we did not design tasks that compare the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

different clusters formed by unsupervised learning methods. Instead, the focus is set on grouping here done by manually selecting groups of patients with shared preconditions or disease outcomes (*Medical 1 + 2*). This is done by providing tables of raw data for interactive data exploration providing support for filtering of specific values and preconditions in addition to using interactive possibilities on a scatterplot. By enabling users to make simple queries with the use of input widgets. Individual patients are then shown in a scatterplot encoded as points and colored by the patient's outcome, with the final results presented in Chapter 5.

Analytical Experts: Analytical experts have the ability to change the clustering process, which asks for detailed aspects of information that need to be visualized. This is defined in *Analytics 3.2*. Natural ways of presenting clusters on high dimensional (more than 3) are already shown during Section 4.3.1, with dimensionality reduced scatterplots using PCA or t-SNE. In addition to that, differences between features in the cluster can be communicated. This is also related to displaying proportions, which is talked about in the next Section.

One way of comparing data of this type is a radar or spider plot [Chambers et al., 1985], which has been popular in the medical context [Saary, 2008]. A draft with boolean features (ranging from $[0, 1]$) is shown in Figure 4.28. Features are aligned on a radial axis using angles while values are encoded on the respective features axis. With too many features the radar plot becomes too crowded and leads to overplotting, meaning that elements in the plot are not recognizable anymore due to overlapping elements. Additionally, for numerical features, values were very high and needed to be transformed to logarithmic form. This lead to negative values, which is not something the radar plot is designed to communicate and fails silently in highly crowded circumstances.

A more simple way of showing this information is a bar chart, which is one of the most preferred visualization types for analysts [Spence and Lewandowsky, 1991]. Because of the high number of features, to save space in one dimension, it makes sense to use a stacked bar chart here. Rather than encoding bars next to each other, stacked bar charts stack them on top of each other. This saves space with the cost of being less easy to read in cases were different scales are used. The result is shown in Figure 4.29. Differences between clusters are easier to grasp and can be done very quickly when compared to the radar plot solution in Figure 4.28.

General Population: First, the users should be given an overview which is done for the general population by giving insights to the different clusters of patients in the dataset. A first draft is shown in Figure 4.30 which shows an approach on how to communicate the hospitalization outcomes of each cluster using a jitter plot [Chambers et al., 1985; Trutschl et al., 2003]. Jitter plots are a deviation of simple scatter plots that introduce small deviations of points from the actual positions. This is especially useful for categorical data or data that tends to overlap. Jittering makes points more readable, and thus, the plot is easier to interpret allowing to better understand distributions in the dataset. Still, when inspecting Figure 4.30, one can observe clutters and overlapping points for less severe hospitalizations where more observations are present. Also we used a very

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

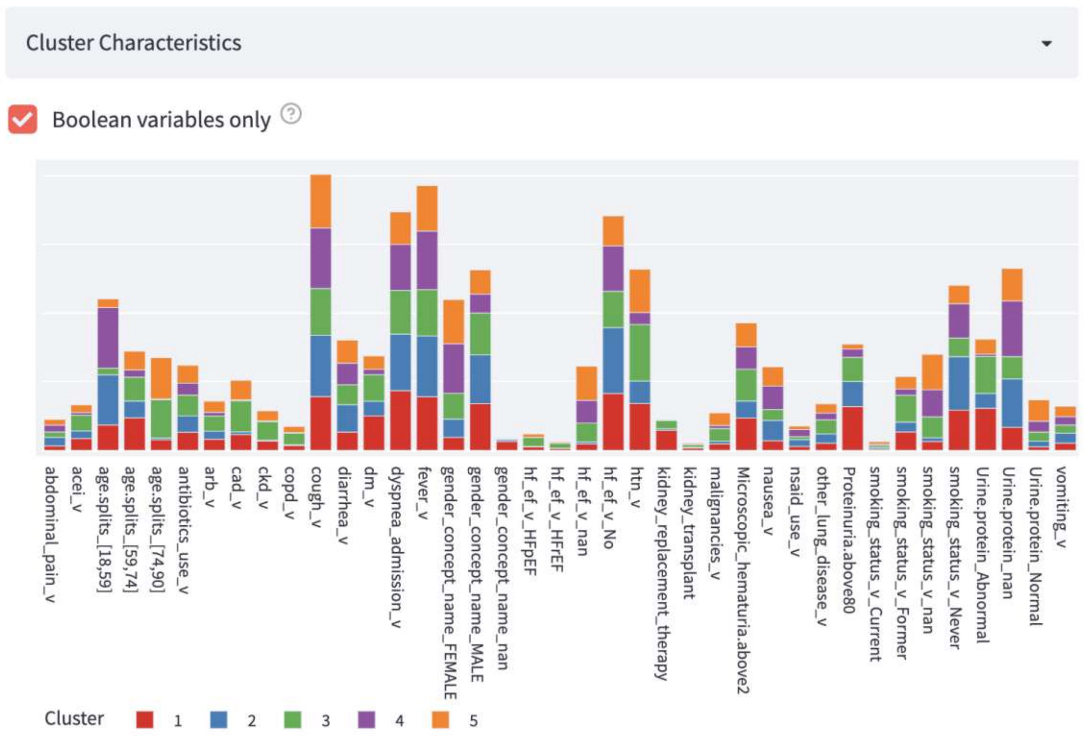


Figure 4.29: Stacked bar chart for comparing clusters of patients with a high number of variables.

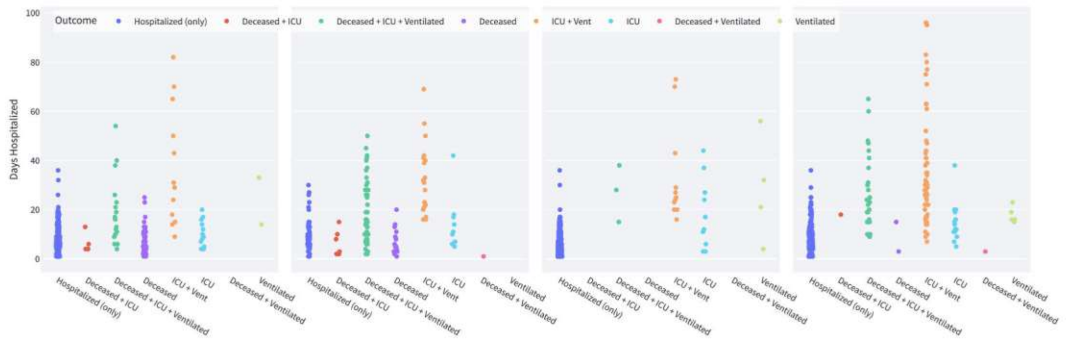


Figure 4.30: Displaying hospitalization outcomes per cluster in a jitter plot.

detailed way to encode the x-axis which does not make good use of the space, hence some variables like Deceased + Ventilated occurred at a low number or not at all for some clusters.

In another effort shown in Figure 4.31 using jitter plots, to overcome unequal distributions in the groups, we sampled 100 patients per group to reduce clutter and combined outcomes in a more natural way. We used symbols (or glyphs) instead of points and colors to

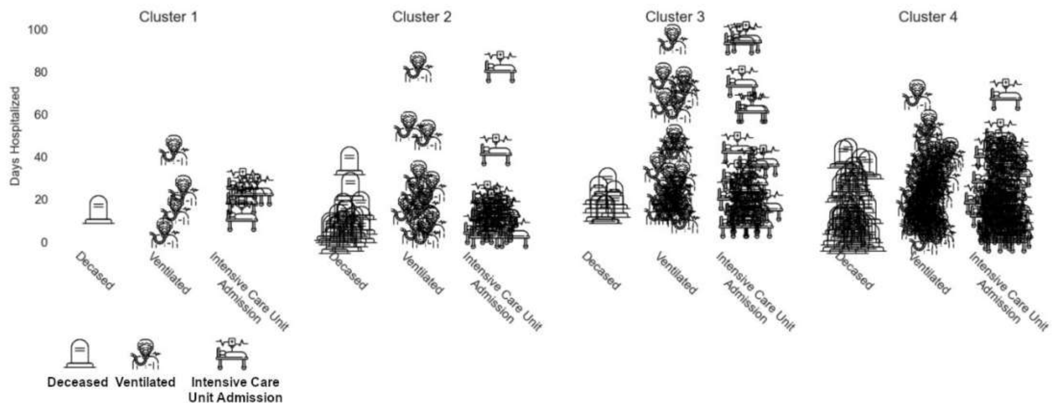


Figure 4.31: Displaying hospitalization outcomes per cluster in a jitter plot with 100 patients sampled per cluster, encoding markers with symbols instead of points.

encode the outcomes again. Unfortunately, the detailed symbols were introducing even more overlapping objects given their size needs to be big enough to be readable, which made the plot really hard to read. In addition it took a long time (+10s) to render, as each glyph is drawn independently. A progressive approach would be used as a solution to this, but we do not pursue glyph plots further due to the overplotting issue.

We finally chose to generate two dedicated plots in order to give the general population a way to compare groups of patients. Therefore, we choose a scatterplot to display the distribution of outcomes per group as well as creating an infographic that inherited most important characteristics of the clusters defined by unsupervised clustering methods. Infographics [Smiciklas, 2012] combine visualization with design principles. They are particular useful to improve engagement and memorability while also being quick to process [Harrison et al., 2015] and have been used for centuries to display complex datasets in an understandable way. Figure 4.32 shows one of the most famous examples of an infographic, the french army march in the Crimean war, combing multiple visualization types in one graphic. The detailed results of our work are shown in Section 5.2.3.

Displaying Distributions

Displaying distributions of data is key for understanding the effects on data imputation that is done by users with analytical backgrounds defined in tasks *Analytical 2.1 + 2.2*.

Data distributions may be visualized by scatterplots using raw or dimensionality reduced data. Given that imputations are done on raw data as defined in *Analytical 2.1*, scatterplots will struggle when categorical data is displayed, resulting in overlapping points. Jitter plots, as already defined before would be an option to enhance this but would most probably not solve all issues.

Instead, the most common way to display distributions is to use a histogram [Pearson, 1895]. Histograms bin the data and encode the occurrences of records, termed frequencies,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

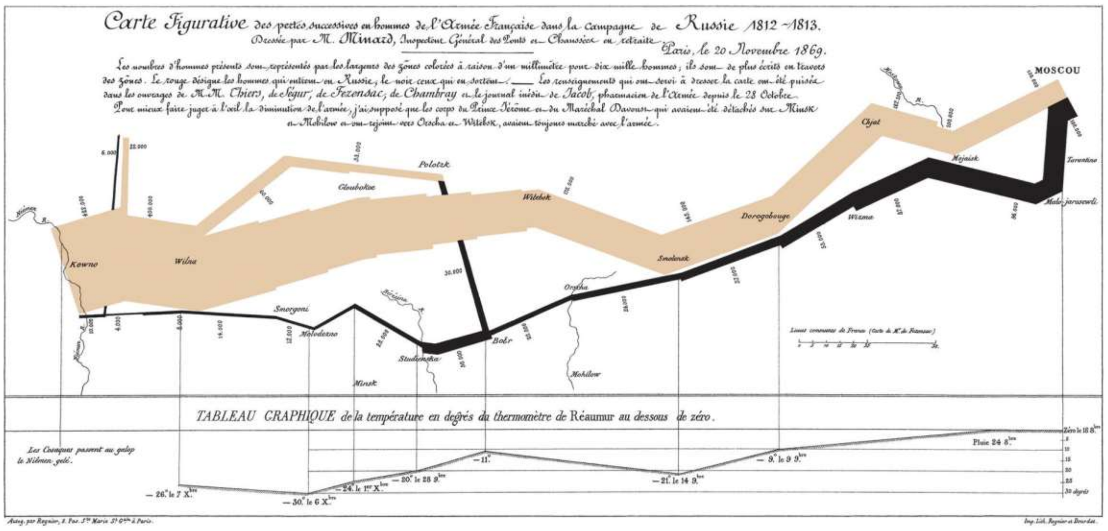

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
MOSCOU
Kowno
Wilna
Polotzk
Witebsk
Smolensk
Dorogobouge
Mojaisk
Tarantino
Malo-jarosewli
Bobr
Studienska
Minsk
Mohilow
Orscha
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Figure 4.32: One of the most famous infographics by Charles Minard from 1861 [Tufte, 1985], showing the march of french army troops from Kaunas to Moscow and back. The amount of living soldiers is displayed as orange and black bars being located on a abstract map of the eastern european region, with highlighted cities and rivers on the way. The bottom line plot shows temperatures during the winter causing massive death tolls.

in a vertical bar chart. The number of bins created is defined by a specific algorithm, of which many can be chosen today. Histograms help understanding the distribution of data on a per variable level. They are, if binning is chosen incorrectly, prone to miss out on specific information as outliers could be missed by inclusion to high frequency bins [Freedman and Diaconis, 1981].

Displaying Proportions and Percentages

Communicating model predictions for the different user groups is a way of displaying proportions and percentages to a user, which is important for solving tasks *Medical 4*, *Analytical 6* and *General 2*. A very popular way of showing proportions is a pie chart, where each proportion is mapped to a slice of a pie. This is a very controversial topic in visualization. The effects of pie charts and their inaccuracies in certain settings have been discussed widely [Siirtola, 2019; Spence and Lewandowsky, 1991]. Figure 4.33 shows how pie charts struggle to display smaller differences, while simpler bar charts are able to communicate them.

Despite their unpopularity in the academic context, pie charts are quite important in business settings and thus present a promising technique for layman users [Kosara and Skau, 2016]. Kosara and Skau [2016] show that arc length and area are the most important parts, while also proving that donut charts are being as accurate as traditional pie charts [Skau and Kosara, 2016]. With this in mind we chose pie donut charts for the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

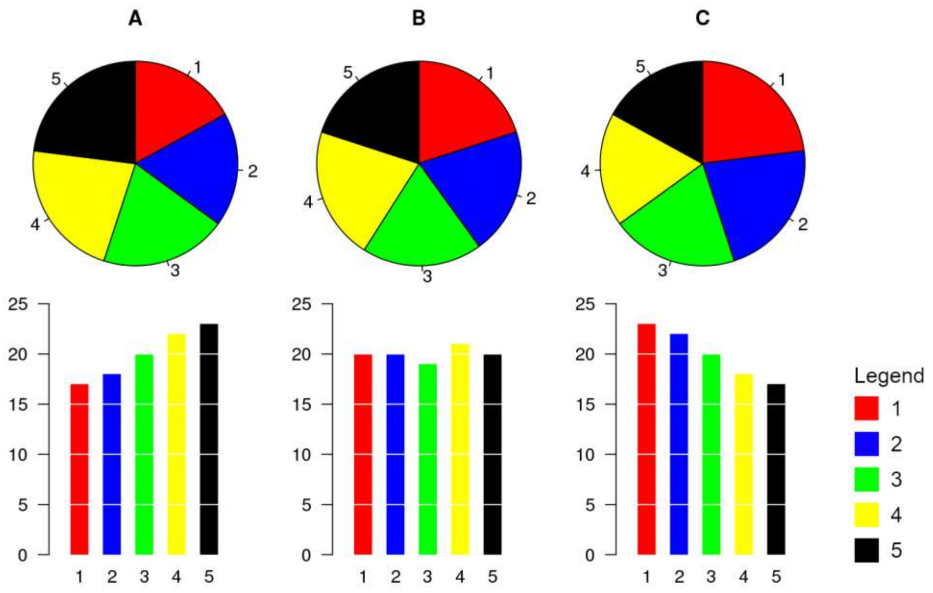


Figure 4.33: Illustration of pie charts versus bar charts. The differences between the lables become quite clear in the bar chart whereas are hard to grasp in a pie chart. Modified from original [User:Schutz, 2007].

general population while using bar charts for medical experts and analytical users, as shown in the final results in Section 5.2.

For analytical users *Analytical 6*, we provide a more detailed overview which is common in machine learning by using tables in form of confusion matrices (Section 3.4.2). In addition to that, for comparing the classification metrics, we used simple line plots with color encodings for different metrics. They have the advantage of needing less space than bar plots.

Segmentation and Disease Progression

Display of automatic image segmentation is important for medical experts as well as users with analytical background. Tasks focusing on this data are *Analytics 4* and *Medical 3*, but with different foci set on the outcome of the analytical process. Medical experts are more interested in the disease progression for successive images of a patient while analytical experts or ML engineers are interested in the automatic segmentations performance.

The easiest way to display segmentation of lungs is to encode the respective area in the image with color, already shown in numerous examples in Chapter 4. Visualizing the disease progression would be possible by displaying the progression in some form of encoded visualization with features, in our case radiomics features, extracted from the images. Those features are several hundred though and multivariate data of this volume

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

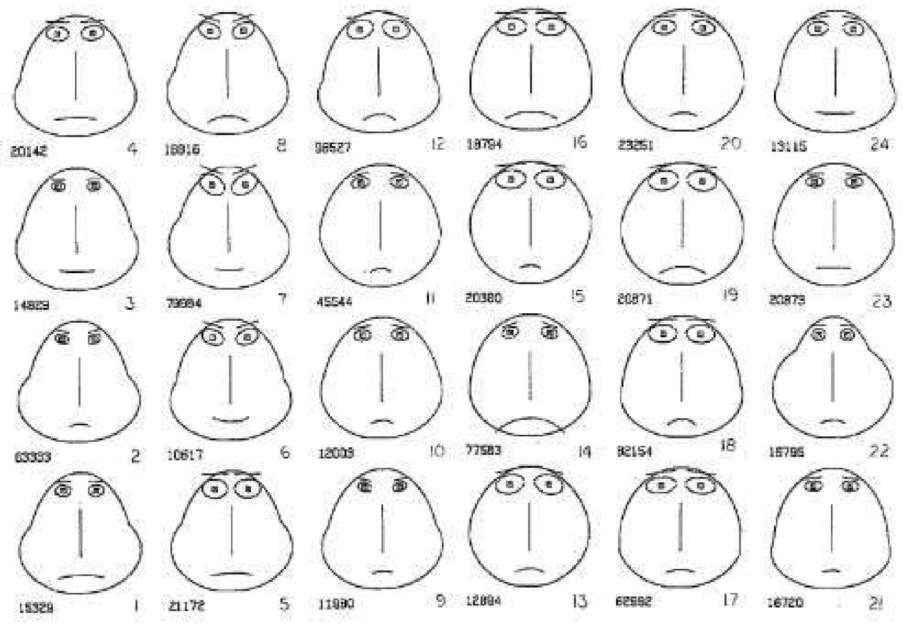


Figure 4.34: Chernoff faces example from Chernoff [1973]. Features are mapped to facial characteristics.

is hard to grasp alone, especially when thinking about visualizing trends or differences between several time steps.

Possible visualization techniques that would be eligible for disease progression display include chernoff faces or radar plots. Chernoff [Chernoff, 1973] faces are a way to represent multivariate data of $k \leq 18$ dimensions in form of cartoon faces, where each feature is represented by facial characteristics, for instance mouth angle, size of eyes or cheeks [Chernoff, 1973]. An example is shown in Figure 4.34. When visualizing a high number of features chernoff faces might not be ideal, while they also require memorization (what is the eye for example).

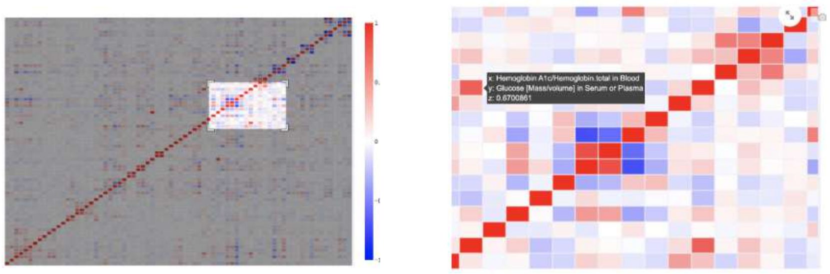
Thus we thought of something else, namely a carousel display. This is ultimately inspired by the infamous carousel slide projector[8] initially used to display slide photographs. We expand this by showing not only one but five images, starting with the first image available as a reference and enabling selection through a slider. The final design is presented in Chapter 5.

4.7.2 Interaction & Interface

The main idea in the visualization design of this thesis is to follow known best practice principles. First focus on overview was already mentioned by works of Shneiderman [2003].

[8] `https://en.wikipedia.org/wiki/Carousel_slide_projector`

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.


x: Hemoglobin A1c/Hemoglobin.total in Blood
y: Glucose [Mass/volume] in Serum or Plasma
z: 0.6700861

(a) Rectangular select

(b) Zoomed in after selection and mouse-hover tooltip

Figure 4.35: Plotly interaction shown on a heat map. The rectangular selection in (a) gets zoomed in and shows the selected area in a higher resolution. Double-clicking resets the view.

This follows the believes of Card [1999], where visualizations enable viewers to see objects of primary interest presented in full detail, while showing the surrounding information. This is defined as *focus+context*, where overview (context) and detailed information (focus) work simultaneously within a visualization design, and where different levels of information are shown in different circumstances. This can be established through showing details on hovering over data objects for example.

Another important principle is *linking and brushing* [Becker and Cleveland, 1987]. While many dedicated visualization techniques exist that have several pros and cons on different types of datasets, the idea in linking and brushing is to combine different visualization techniques to combine their pros and limit their cons [Keim, 2002]. It is done by encoding information between different visualizations and following interactions by the user. This is exceptionally useful in highly multivariate datasets, as in this thesis. Several tasks have been following this idea, namely *Medical 1* combining tables and scatterplots, or *Analytical 3*, using multiple visualization types to give users insights into different detail levels of information.

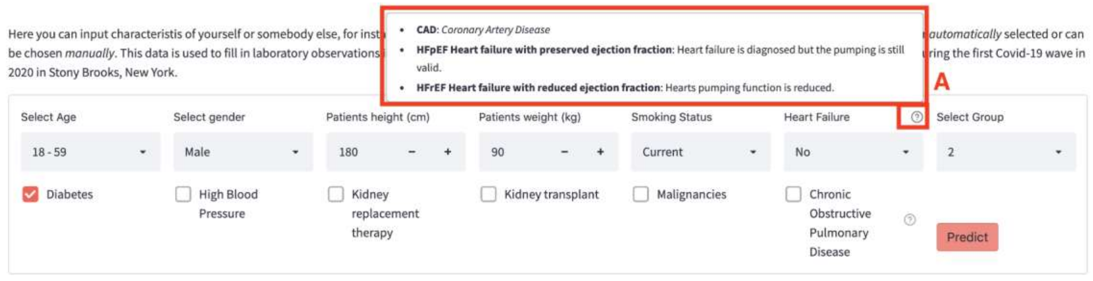
Visualizations are built using *Plotly* and *Apache Echarts*[9]. Both libraries by default implement filtering, brushing, mouse hover tooltips and picture exports for all plots, shown in Figure 4.35.

Streamlit has built in integration with Plotly which enhances interaction possibilities even more. Streamlit by default implements many input widgets[10]. From buttons, checkboxes, radio select buttons, sliders and camera inputs or file uploaders. In addition to that, most widgets can be made more explainable by defining help areas within widget creation as seen in Figure 4.36. This lets designers be more compact in what information is shown at first glance. Every figure, picture or table element is also expandable to full screen as depicted in Figure 4.37.

[9] echarts.apache.org/examples/en/index.html

[10] https://docs.streamlit.io/library/api-reference/widgets

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.


Here you can input characteristis of yourself or somebody else, for inst… be chosen *manually*. This data is used to fill in laboratory observations… 2020 in Stony Brooks, New York.

- **CAD:** *Coronary Artery Disease*
- **HFpEF Heart failure with preserved ejection fraction:** Heart failure is diagnosed but the pumping is still valid.
- **HFrEF Heart failure with reduced ejection fraction:** Hearts pumping function is reduced.

…*automatically* selected or can …ring the first Covid-19 wave in

Select Age	Select gender	Patients height (cm)	Patients weight (kg)	Smoking Status	Heart Failure	Select Group
18 - 59	Male	180	90	Current	No	2
[✓] Diabetes	[ ] High Blood Pressure	[ ] Kidney replacement therapy	[ ] Kidney transplant	[ ] Malignancies	[ ] Chronic Obstructive Pulmonary Disease	Predict

Figure 4.36: Illustration of the help signs that pop out on every *?* symbol either on click or mouse hover, explaining inputs in multiple locations throughout the whole application.

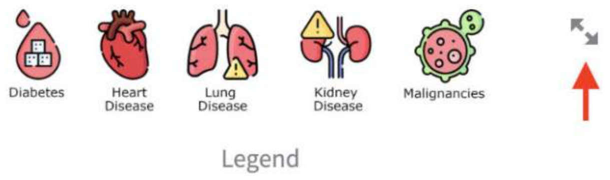


Figure 4.37: Illustration of full screen expansion which comes with every figure, image or table element in streamlit, highlighted by the arrow.

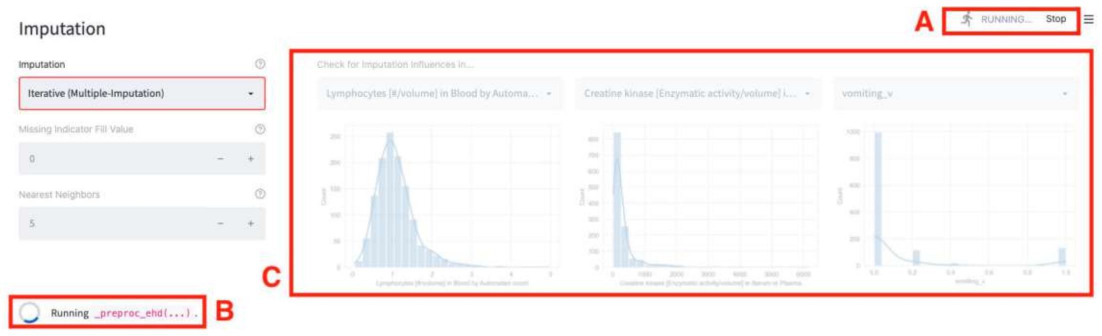

Imputation
Imputation
Iterative (Multiple-Imputation)
Missing Indicator Fill Value
0
Nearest Neighbors
5
Running _preproc_ehd(...)
B
C
A
RUNNING... Stop
Check for Imputation Influences in...
Lymphocytes [#/volume] in Blood by Automa...
Creatine kinase [Enzymatic activity/volume] i...
vomiting_v

Figure 4.38: Illustration of wait time indication in streamlit. In the top right corner a running animation is displayed indicating the application is computing (**A**). In addition, a temporary text field is displayed indicating which actual method is running at the very moment (**B**). All succeeding visualizations are made semi-transparent which furthermore indicates processing (**C**).

Wait times during computations in streamlit are indicated by several objects out of the box, shown in Figure 4.38. For the time a computation runs, a running animation is displayed in the top right corner (Figure 4.38**A**), a temporary textbox is shown which displays the method that is computing at the moment (Figure 4.38**B**) and all succeeding visualizations are made semi-transparent during the computation (Figure 4.38**C**). When the computation is finished, the indications are removed again automatically.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 5

Results and Evaluation

This chapter presents the results and evaluates them quantitatively and qualitatively based on the proposed research question and following detailed implementation described in Chapter 4.

First, quantitative evaluation is carried out in regards to the defined research questions RQ1 and RQ2 defined in Section 1.2. This includes presenting and evaluating clustering and prediction of COVID-19 outcomes based on the electronic health data available. Furthermore, outcomes are predicted based on medical image data. We measure the performance of automatic segmentation of CXR by applying a novel approach of transfer learning back to a domain where GT segmentations are available. Different approaches on how to deal with feature extraction in PyRadiomics are compared. Finally, we evaluate merging the feature spaces. Results are then compared to previous publications.

Visual analytics results are evaluated qualitatively next using methods described in Section 3.5.3. We conduct usage scenarios by mimicing user behaviour and proving usability and task fulfilments for medical experts and analytical users as well as providing case studies with $N = 6$ users for the general population, where users were asked to test the application.

5.1 Quantitative Evaluation

Figure 5.1 shows Pearson's correlation values for all features with the eight hospitalization outcomes, sorted in ascending order for less severe cases in which patients have only been hospitalized without dying, need for ventilation, or ICU admission. Variables with a negative correlation with hospitalization outcome tend to have a higher correlation with worse disease progression. Note that all values are in the range $[-0.5, 0.5]$.

Highest negative correlation values for a less severe infection occur for *C reactive protein* values. It is an indicator of an inflammatory disease, like open injury or infection. The

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

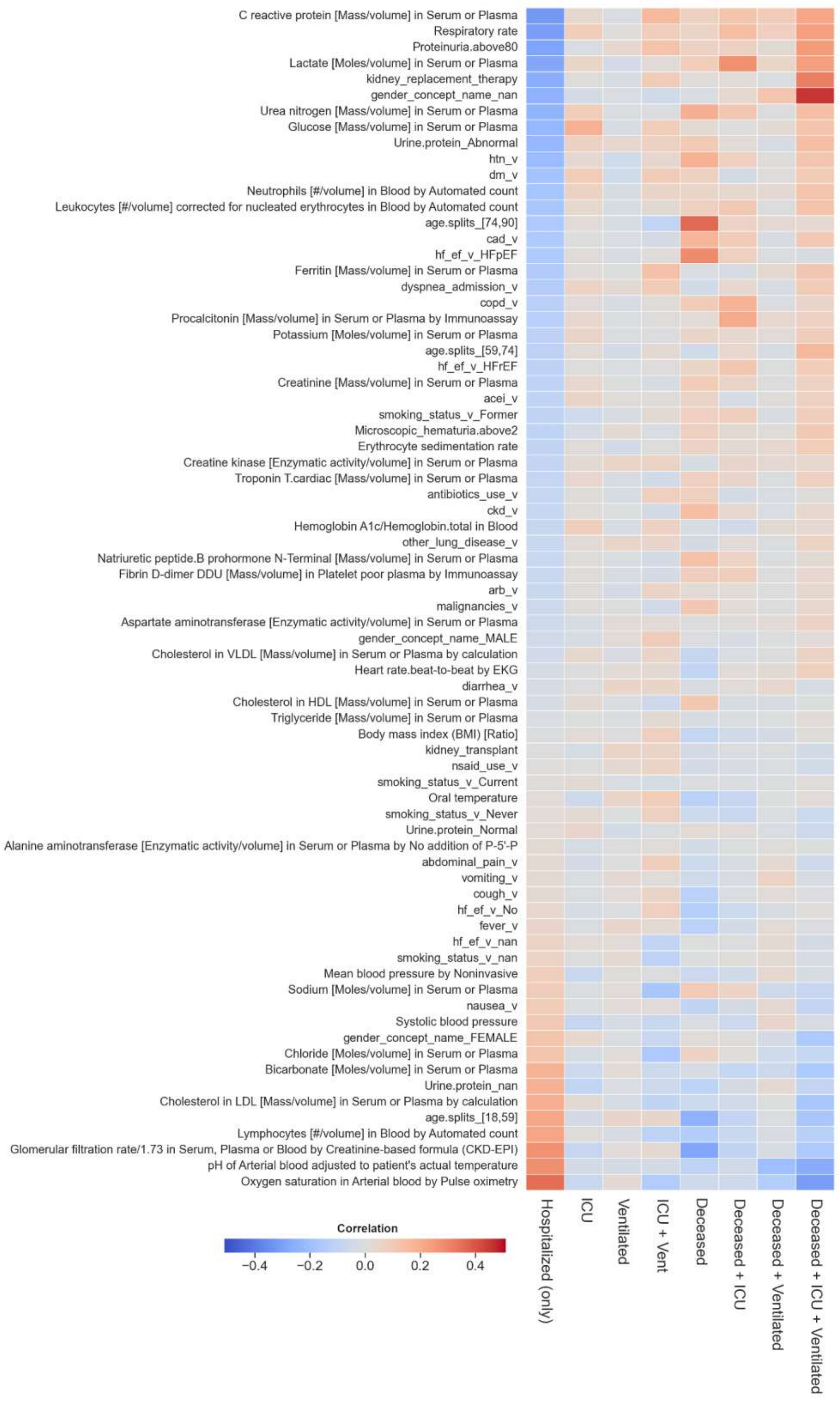


Figure 5.1: Correlation heat map for variables and prediction targets. Ordered ascending by correlation for patients that were hospitalized only.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

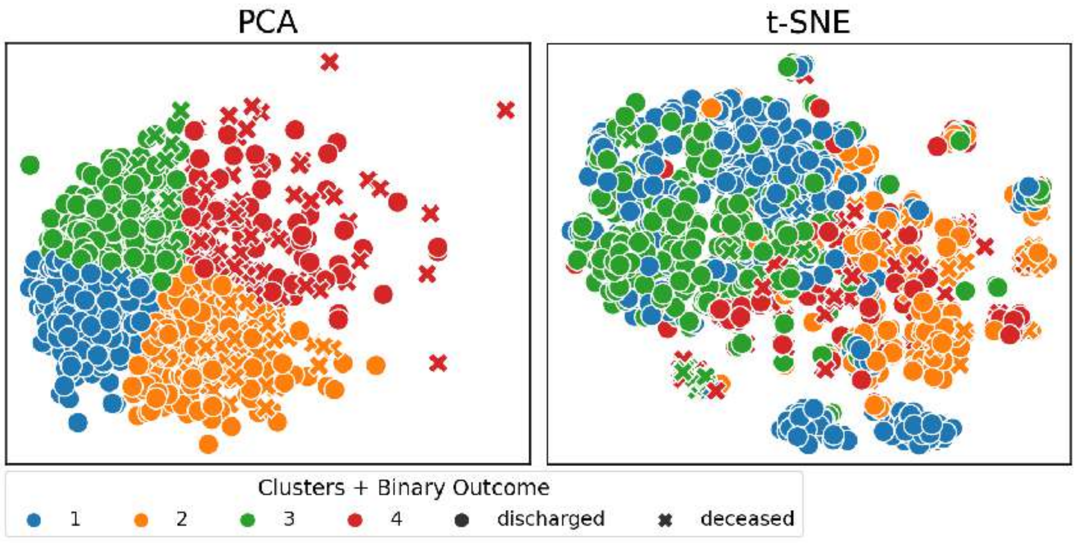


Figure 5.2: Final K-means clustering in PCA and t-SNE dimensionality reduced representation. Clear clutter of low risk patients are visible in the t-SNE representation at the bottom.

respiratory rate measures the number of breaths, usually per minute. Proteinuria indicates high numbers of protein in the urine. It is common in elderly patients with reduced kidney functions. *C-reactive protein* has also been found as an important biomarker for COVID-19 by previous works as reported in Section 3.1.3. Ponti et al. [2020] only hypothesize that chronic kidney disease may be an influence towards severe cases. This argument can be proven to some degree with the dataset available here, as patients with kidney replacement therapy have a higher correlation towards a more severe case and vice versa.

5.1.1 Clustering

The final clustering is displayed in Figure 5.2 for K-means with $k = 4$ clusters in PCA and t-SNE representation. How patient outcomes are distributed between the clusters is shown in Figure 5.3, with total values in Table 5.1. There are some major differences between the clusters here in terms of outcome. Cluster 1 is almost solely consisting of patients that did not decease. Cluster 2 is made up of patients that had a high chance of having a mortal outcome during their hospital stay, with only a few of them getting ventilation support. Cluster 3 is made up of a high number of hospitalized only patients while it has the highest proportion of patients that got ventilated and got admitted to ICU during their stay. Cluster 4 is mostly made up of patients that arguably had a high chance for a serious progression of the disease during their stay, with a high number of deceased or ventilated patients.

When testing for significant differences between the clusters in feature variables, we would

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Hospitalized (only)	462	164	269	53
ICU	10	12	15	9
Ventilated	4	2	5	0
ICU + Vent	11	12	59	18
Deceased	0	43	2	16
Deceased + ICU	0	4	1	6
Deceased + Ventilated	0	0	1	1
Deceased + ICU + Ventilated	3	17	26	54
N	490	378	254	157

Table 5.1: Outcome distribution per cluster.

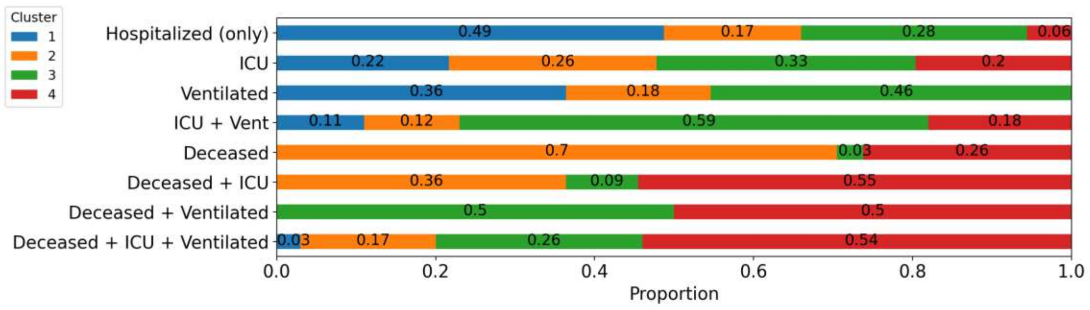


Figure 5.3: Distributions of outcomes per cluster.

expect non-significant distributions between cluster 1 and cluster 3 as well as clusters 2 and 4 given the outcomes observed in Figure 5.3. Testing with Mann-Whitney U for each variable (74) and each combination of clusters (6) we get significant differences for 355 out of 444 tests. 89 tests did not yield a significant difference. 21 of those are between clusters 1 and 3. and 18 are between clusters 3 and 4. Between cluster 2 and 4, 15 tests were not significantly different (Cluster 1 vs 4: 13; Cluster 1 vs. 2: 12; 2 vs. 3: 10). So indeed clusters 1 and 3, as well as 2 and 4, are similar also based on the variables.

In Figure 5.4 each boolean variable is averaged for each respective cluster. Again, a clear difference between clusters 1+3 and 2+4 can be distinguished. Clusters 1+3 are consisting of younger patients. Together with cluster 4, they inherit patients that are less likely to have heart failure as a pre-existing disease (*hf_ef_no*). Since this correlates with smoking, they also have a higher proportion of self-reported never-smokers overall. In general, there was no difference between the clusters for current smokers. Cluster 4 inherits most of the oldest patients, with only a marginal number in Cluster 1+3. Heavier preconditions like coronary artery disease (*cad*), chronic kidney disease (*ckd*), chronic obstructive pulmonary disease (*copd*) or malignancies are more often found in clusters 2 and 4. Diabetes (*dm*) is less likely to be found in cluster 1. Lighter symptoms like

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

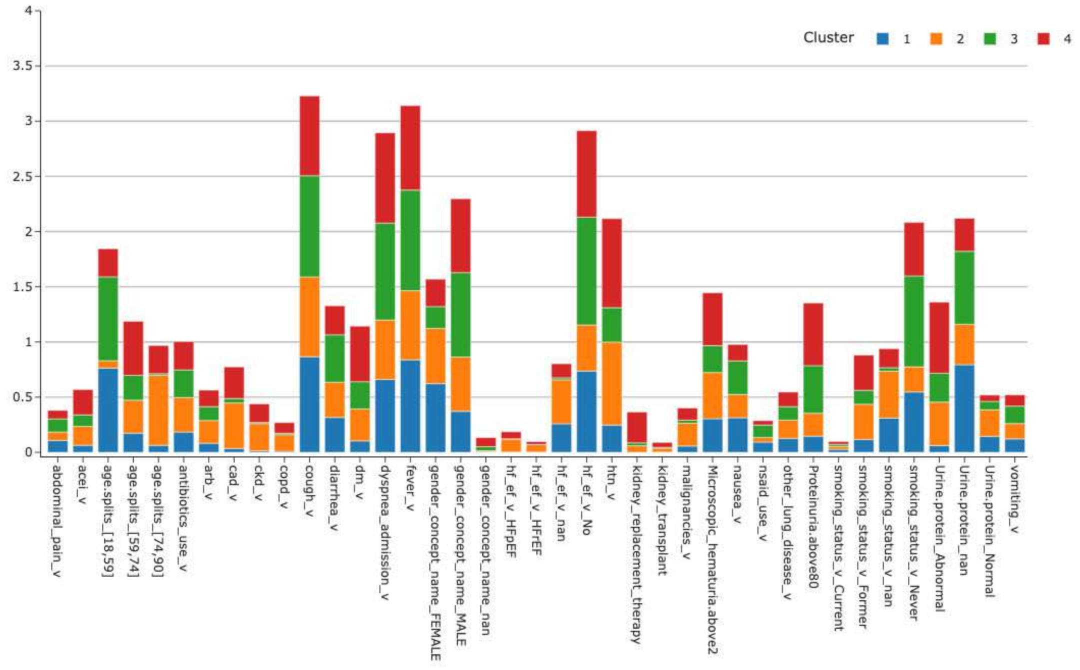


Figure 5.4: Mean values of boolean variables for each respective cluster.

coughing, fever or diarrhea are present in all groups. Kidney replacement therapy is mostly found in patients of cluster 4.

Figure 5.5 shows averages for features of continuous numerical type. Values are shown in log-transformed form. The differences here are not as easy to spot as before. If focus is laid upon the higher correlated features for critical hospitalization outcome displayed in Figure 5.1 some artifacts can still be spotted. For C reactive protein, cluster 1 shows the lowest number on average, hence they are the least probable patients to suffer from a serious hospitalization stay. For urea nitrogen values, cluster 2+4 show higher values as the other consisting of more healthy patients.

We thus define the four clusters as the following groups:

- **Cluster 1 — Healthy Young to Middle aged**: Young to Medium age $[18, 59]$; Unlikely to have preconditions; Lowest decease rate; Shortest hospital stay on average ($median = 2.0 \pm 5.0 IQR$); Arguably low rate of ventilation overall; More female patients on average.
- **Cluster 2 — Elderly High Risk**: "Oldest Group" with medium to high age $[59, 74] - [59, 90]$; Likely to have preconditions (Diabetes mellitus *dm*; Heart issues *htn*, *ht_ef*; Malignancies); Lowest Body Mass Index $28.1 \pm 5.2 std$; Lowest rate of never-smokers; Highest decease rate with many dying quickly (second lowest hospital stay $7.0 \pm 8.0 IQR$).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

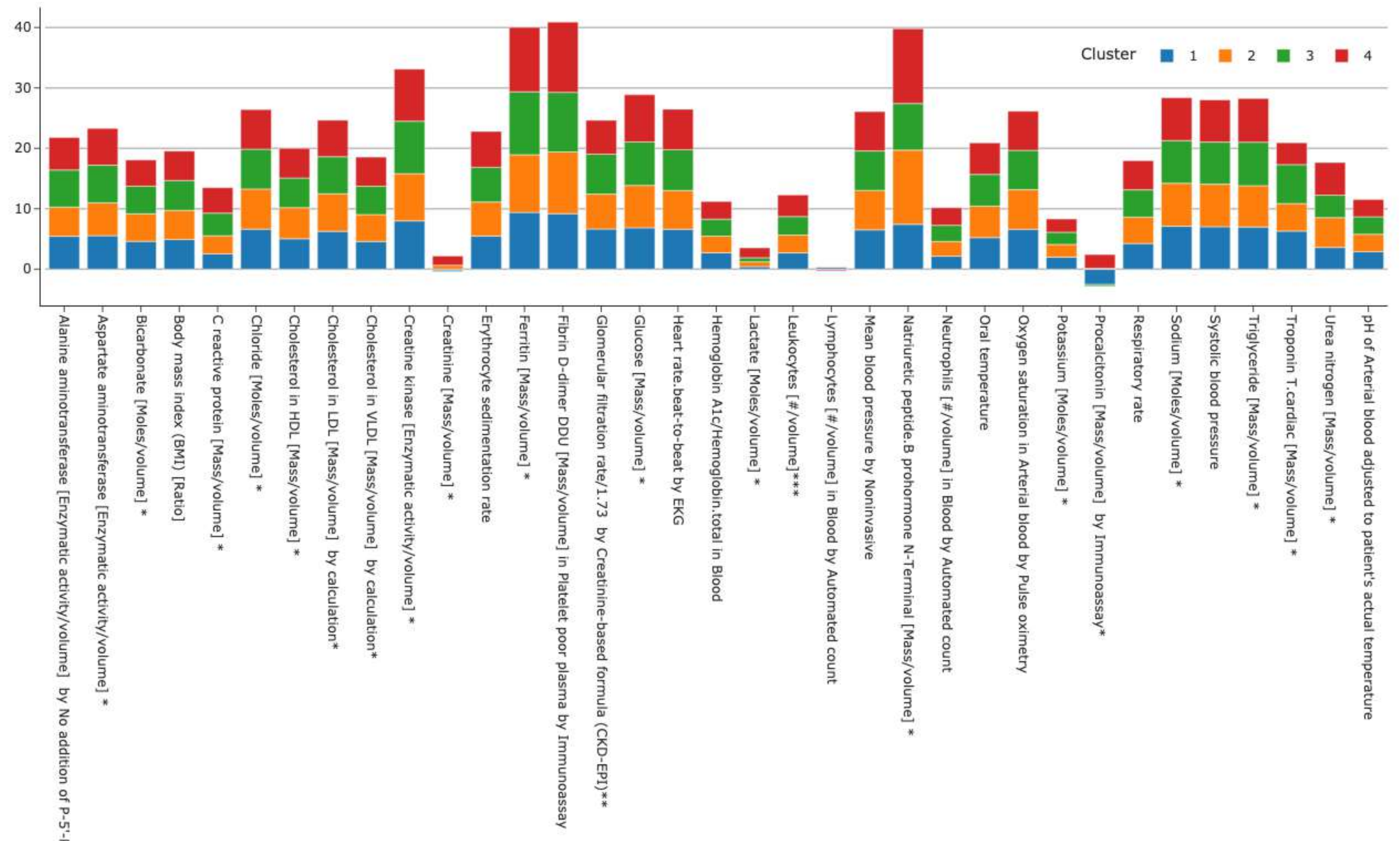


Figure 5.5: Mean values of numerical variables for each respective cluster. Y-axis in logarithmic scale. Variables names shortened for easier display: $*$ in Serum or Plasma; $**$ in Serum, Plasma or Blood. $***$ Leukocytes values have been inverted (multiplied by-1) for easier display.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

	TPR-Sensitivity (recall)	TNR-Specificity
Hospitalized	0.975	0.265
Ventilated	0.000	1.000
ICU	0.000	0.994
ICU + Ventilated	0.040	0.976
Deceased	0.067	0.997
Deceased + ICU	0.000	1.000
Deceased + ICU + Ventilated	0.423	0.983

Table 5.2: Per class Sensitivity and Specificity for the best performing classifier.

- **Cluster 3 — Less Healthy Young to Middle aged**: Also Young to Medium aged, not significantly different to Cluster 1 but significantly more male patients; Unlikely to have preconditions; Significantly larger hospital stay time $9.0 \pm 12.0 IQR$; High chance of ICU admission and ventilation. Overall Cluster 3 is not significantly different from Cluster 1 in 21/74 variables. They differ by having more male patients while containing a similar BMI. Male patients tend to have a lower BMI on average. There is also a higher amount of patients with diabetes in Cluster 3 which would be an indicator of more obese patients in that group.
- **Cluster 4 — High Risk with pre-existing Conditions**: Patients from all age groups, with the highest proportion of patients between $[59, 74]$; Largest hospital stay duration $11.0 \pm 15.0 IQR$; Highest proportion of patients with a severe hospitalization stay, mostly with patients that were admitted to ICU, ventilated and deceased.

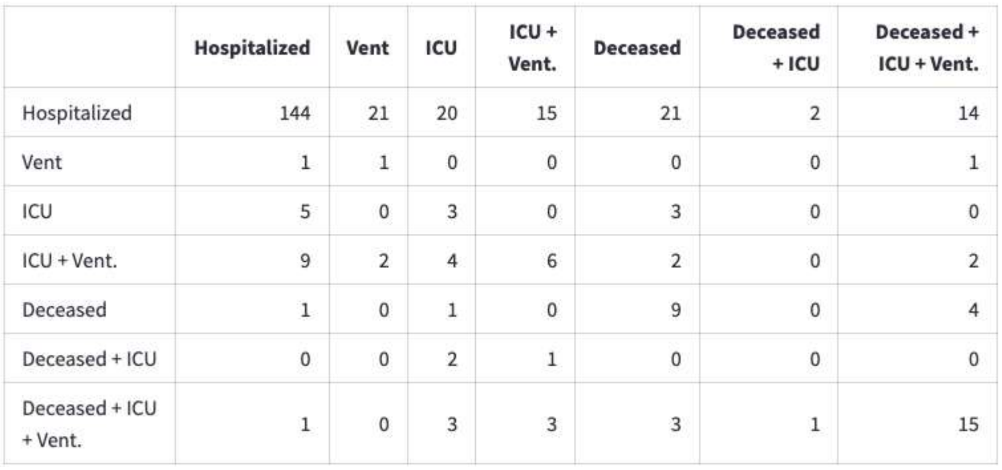
5.1.2 Clinical Data Prediction

The best performing model identified in Section 4.3.2, SVC with balanced weights, was used to train the final classifier. Hyperparameter tuning was performed in a grid-search cross-validation approach. The best metrics in terms of ROC-AUC where found for $C = 25$ with a linear kernel and seed 0, with a ROC of 0.795 ± 0.063 on the training set and 0.780 on the validation set. Per class sensitivity and specificity are found in Table 5.2.

When comparing those results to other publications, we get similar results. Specificity for ICU admission and mortality are equally high when compared to Zhao et al. [2020] for instance, with sensitivity and specificity of the risk scores were 10.5% and 99.2% for predicting ICU admission, and 7.1% and 100% to predict mortality. Others, using a risk model achieve 28.7 sensitivity and 98.2 specificity for identifying high risk patients [Dai et al., 2020].

Different refit strategies can be used in order to optimize the classifier towards a certain prediction target and penalize wrong predictions. The model presented previously was

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub



	Hospitalized	Vent	ICU	ICU + Vent.	Deceased	Deceased + ICU	Deceased + ICU + Vent.
Hospitalized	144	21	20	15	21	2	14
Vent	1	1	0	0	0	0	1
ICU	5	0	3	0	3	0	0
ICU + Vent.	9	2	4	6	2	0	2
Deceased	1	0	1	0	9	0	4
Deceased + ICU	0	0	2	1	0	0	0
Deceased + ICU + Vent.	1	0	3	3	3	1	15

Figure 5.6: Confusion Matrix of the used model for the general population dashboard.

retrained to maximize recall/sensitivity. Higher sensitivity scores are more important to doctors for instance, who do not want to miss patients with risks. This model is finally trained to be used for the general population in a way that they are able to get a risk assessment for individual inputs. Here we only want to correctly identify people without risks, focusing on *specificity*.

For the sake of completeness, the model with the best specificity was trained by focusing on balanced accuracy. The confusion matrix for that model is shown in Figure 5.6. Most miss-classifications are observed for the hospitalized class, which makes sense given its the majority class. Patients that had a mortal outcome following ventilation and ICU admission were often falsely classified to ICU, ICU + Ventilated and Deceased, proving the classifier was able to detect high risk outcomes pretty good while being too pessimistic on many hospitalized only patients.

5.1.3 Image Data

In this Section prediction results using CXR images and radiomics feature extraction are presented. Results are shown for the automatic image segmentation after evaluating a best performing model through qualitative and quantitative evaluation shown in Section 4.4.

Automatic Image Segmentation

Results of the reverse transfer learning on the MC and SD validation split, as shown in Figure 4.19 are presented in Table 5.3. Preprocessing had a slightly positive impact on Dice and Jaccard scores as well as lower standard deviations, leading to more stable results. Also, the results are not extremely far off from validation metrics of the original *Baseline Model A* architecture trained on the MC and SD dataset, as reported in Table 5.3. The

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

	loss	loss-std	Jaccard	Jaccard-std	Dice	Dice-std
Baseline Model A	0.0531	-	0.9268	-	0.9611	-
Model A	**0.0905**	**0.0560**	0.8858	0.0570	0.9384	0.0342
Model E	0.1062	0.0763	0.8794	0.0655	0.9344	0.0402
Model A**	0.0955	0.0664	**0.8884**	0.0596	**0.9398**	0.0353
Model E**	0.0931	0.0681	0.8846	**0.0553**	0.9378	**0.0325**

Table 5.3: Results for model evaluation on the validation set from the original training dataset MC and SD as shown in Figure 4.19. ** depicts the usage of preprocessing. The baseline model shows the metrics for the Model A architecture trained on the MC and SD dataset and validated on the validation set. The remaining results refer to validation metrics that were achieved by learning from segmentation masks from the respective model architecture and preprocessing (blur and adaptive histogram equalization). Best results are in bold.

best results in terms of evaluation metrics are observed for Model A with preprocessing, which was ultimately selected as the segmentation method.

Segmentation results of the chosen model for some examples from the validation dataset are shown in Figure 5.7. Most of the resulting segmentation look good, with a high amount of overlapping areas. Some artifacts are visible that occur in visually similar regions. The artifact on the finger may be caused by more regularly overlapping tubes in the SBU dataset that are dividing the lungs, which leads the segmentation to respect those areas as well, but this is not the case for all samples (compare mid picture). The low score on some images might be due to the less opaque structure in the sample which lead to the worst result in the sample provided here. Generally the automatically segmented area tends to be smaller than the original.

Radiomics Features as Biomarkers

In Figure 5.8 the highest negatively and positively correlated features in regard to a less severe disease progression during the hospitalization are displayed. Given the correlation is only between -0.4 and +0.3 this can only be considered as a weak to medium correlation between the outcome and the radiomics features.

Figure 5.9 and Figure 5.10 show examples of lower and higher values for features that showed a the highest correlations for either a less severe or more severe outcome. The highlighted regions in the images are only shown for the respective lung used for feature extraction. The values reported in the figures as part of the image titles were used in a centered and scaled representation for computation of the correlation heat map depicted in Figure 5.8.

The radiomics feature Informational Measure of Correlation in right lungs (Imc_r) showed a the highest negative correlation for a less severe stay at the hospital, so

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

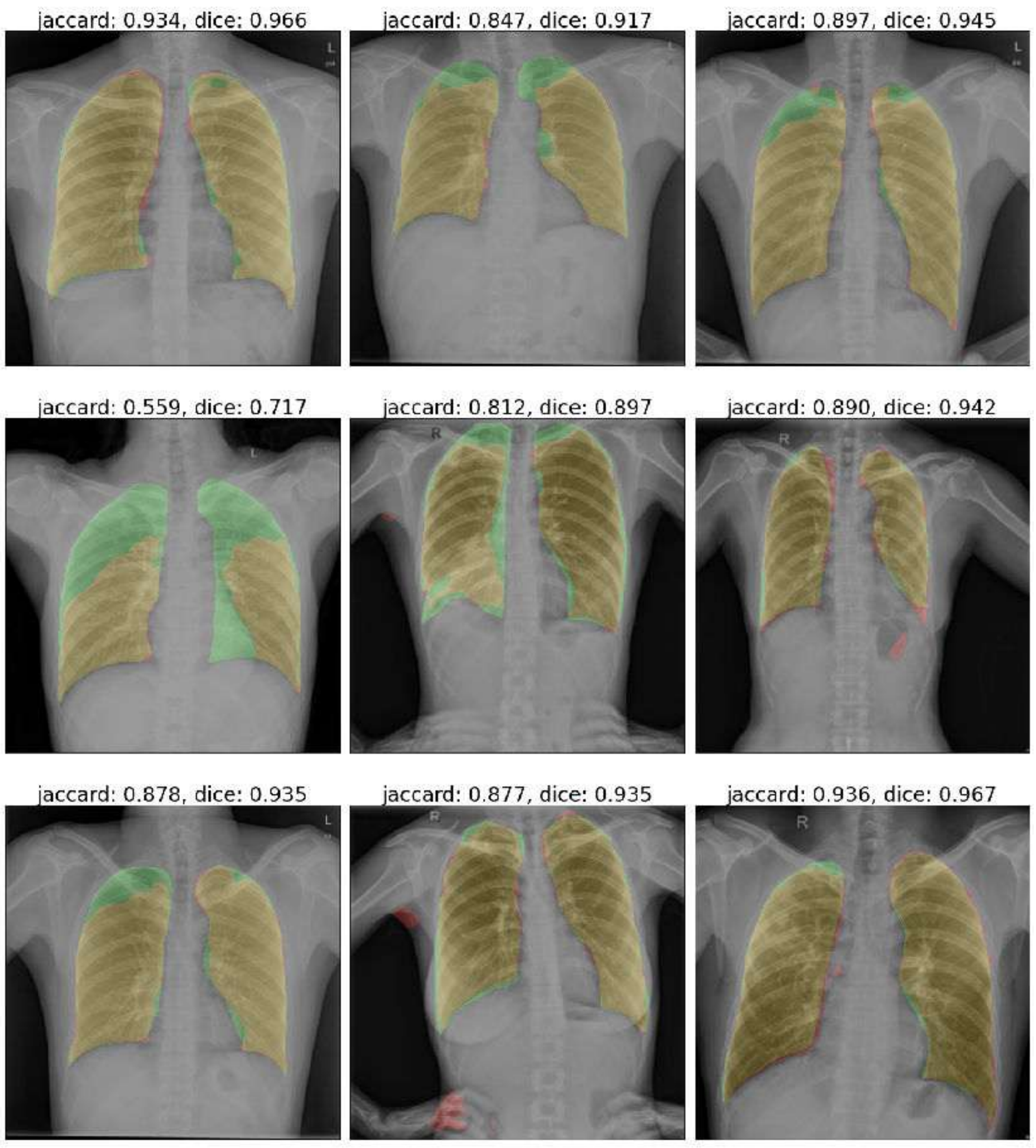
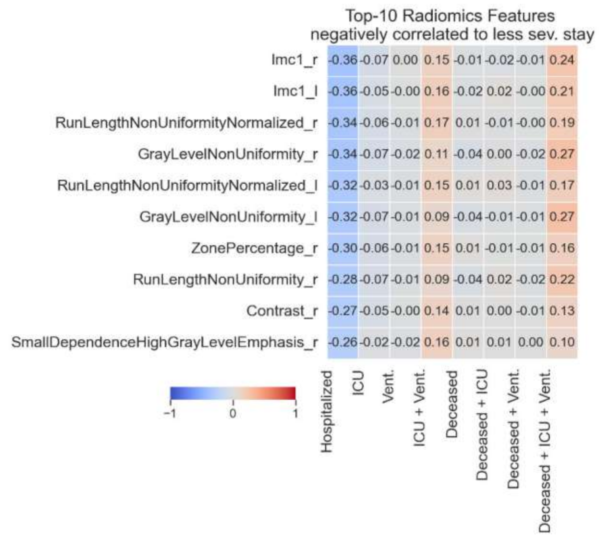

jaccard: 0.934, dice: 0.966
jaccard: 0.847, dice: 0.917
jaccard: 0.897, dice: 0.945
jaccard: 0.559, dice: 0.717
jaccard: 0.812, dice: 0.897
jaccard: 0.890, dice: 0.942
jaccard: 0.878, dice: 0.935
jaccard: 0.877, dice: 0.935
jaccard: 0.936, dice: 0.967

Figure 5.7: Results for the reverse transfer learning approach on MC and SD validation data. Training data created from Model A with Gaussian blur and adaptive histogram equalization. Green Area: GT; Red: Automatic segmentation; Yellow: Overlap.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



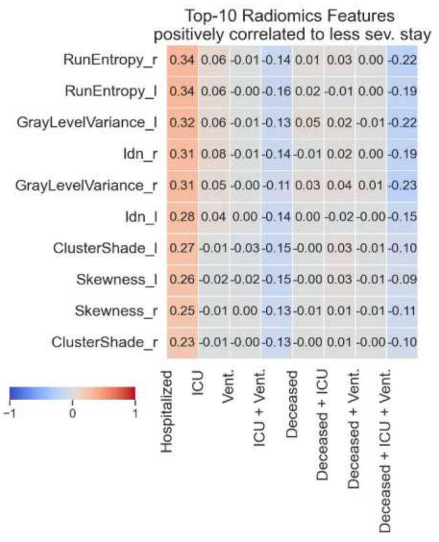


Figure 5.8: Radiomics features correlation heat map with hospitalization outcomes. Top-10 negatively and positively correlated features towards less severe disease progression.

that lower values would indicate a less severe case in CXR. The feature quantifies the complexity of the texture in the ROI by utilizing the Gray Level Size Zone Matrix (GLSZM). The examples in Figure 5.9 show that lower values are results of cleaner regions with less disturbances in the image. The texture is more smoother when gray values are more similar. For higher values, we observe that those occur in lungs that are pervaded by tubes which probably indicate the CXR was taken during ICU admission. The disturbances followed by those artifacts result in more complex textures. Keeping in mind the negative correlation between Imc_r and less severe cases, the results are reasonable.

In Figure 5.10, examples with lower and higher values of the Gray-Level Non Uniformity in left lungs (GLN_l) are shown. This feature measures the variability of gray-level intensity values in the image, with lower values indicating more homogeneity [1]. Again, lower values here indicate a less severe disease outcome (negative correlation) while the feature has shows the highest correlation for patients that were Ventilated, commissioned

[1]`https://pyradiomics.readthedocs.io/en/latest/features.html#module-radiomics.glszm`

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

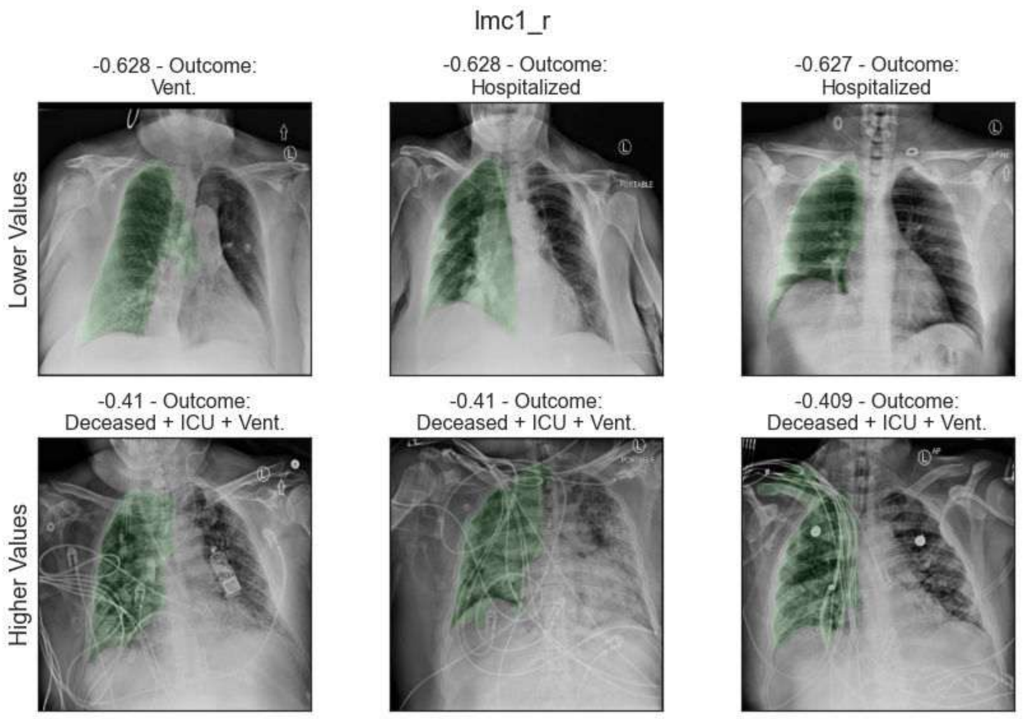

Imc1_r
-0.628 - Outcome: Vent.
-0.628 - Outcome: Hospitalized
-0.627 - Outcome: Hospitalized
Lower Values
-0.41 - Outcome: Deceased + ICU + Vent.
-0.41 - Outcome: Deceased + ICU + Vent.
-0.409 - Outcome: Deceased + ICU + Vent.
Higher Values

Figure 5.9: Results with lower and higher values for Informational Measure of Correlation in right lungs (Imc_r), quantifying the complexity of the texture. Higher values occur when lung regions are pervaded by tubes, indicating ICU stay. Lower values are visible in more clean examples. Green areas show the results of the automatic segmentation.

to the ICU and had a mortal outcome. Again lower values are observed for clean images. Higher values are experienced for images that may be accompanied by artifacts such as cables, but also are reported in images that show a higher proportion of smaller white areas which could indicate inflamed areas in the respective ROI.

Prediction on Radiomics Features

Again, the best model was trained using grid-search hyperparameter tuning for SVC with balanced weights. The best model was found for $C = 3$ with a radial basis function kernel and seed 420. Best results were found when tuning for ROC with a score of 0.747 ± 0.063 on train and 0.776 on validation. Sensitivity and specificity respective to each class are reported in Table 5.4. The only comparable work found to the best of our knowledge is from Varghese et al. [2021]. They report AUC 0.71 for predicting death and 0.61 for ICU admission for two binary classification settings.

The confusion matrix shown in Figure 5.11 highlights a few issues. The model ultimately not able to change predictions affection towards majority classes, in this case hospitaliza-

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

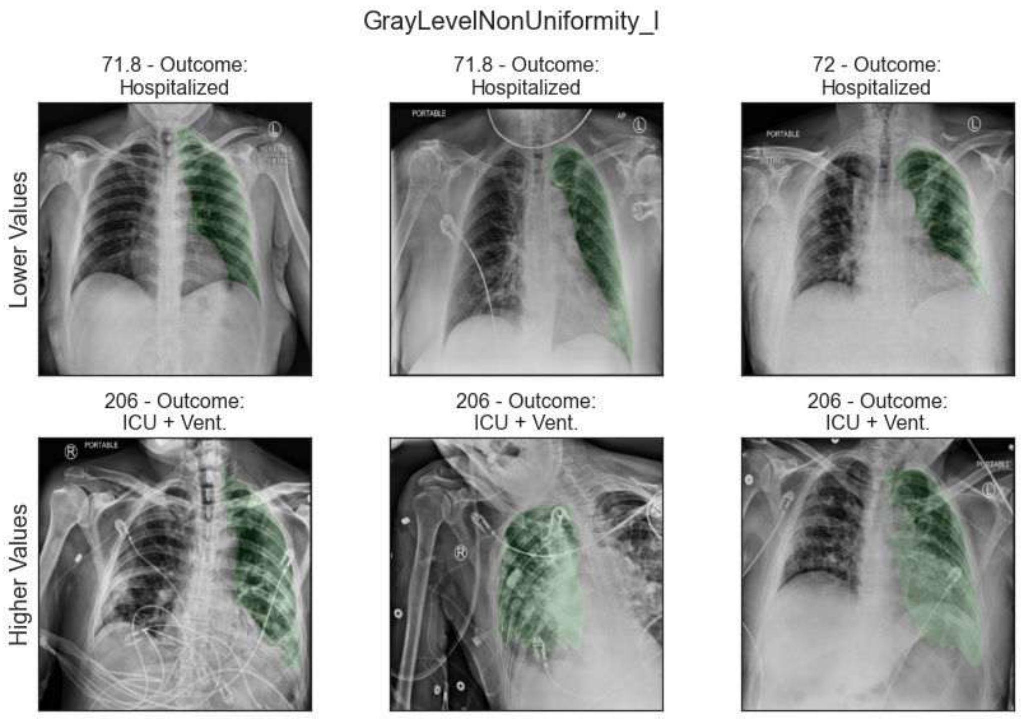

GrayLevelNonUniformity_l
71.8 - Outcome: Hospitalized
71.8 - Outcome: Hospitalized
72 - Outcome: Hospitalized
Lower Values
206 - Outcome: ICU + Vent.
206 - Outcome: ICU + Vent.
206 - Outcome: ICU + Vent.
Higher Values

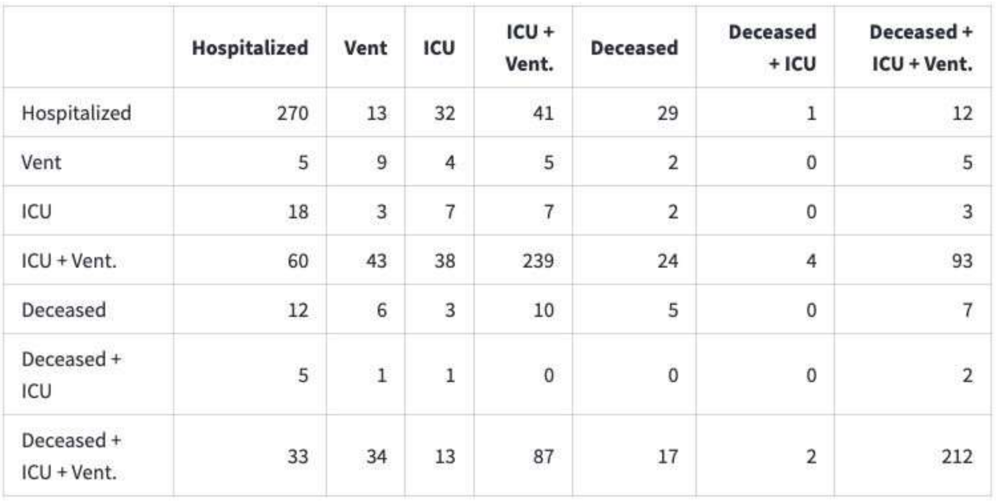
Figure 5.10: Results with lower and higher values for Gray-Level Non Uniformity (GLN) in left lungs. Higher values are visible for regions pervaded by tubes and in images with more inflamed regions. Lower values are visible in more clean examples.

	TPR-Sensitivity (recall)	TNR-Specificity
Hospitalized	0.678	0.870
Ventilated	0.300	0.928
ICU	0.175	0.934
ICU + Ventilated	0.477	0.837
Deceased	0.116	0.946
Deceased + ICU	0.000	0.995
Deceased + ICU + Ventilated	0.533	0.881

Table 5.4: Class sensitivity and specificity for prediction based on radiomics features.

tion, ICU + Ventilation and Deceased + ICU + Ventilation. Minority classes are thus misclassified often, hence Deceased + ICU often classified as Hospitalized.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



	Hospitalized	Vent	ICU	ICU + Vent.	Deceased	Deceased + ICU	Deceased + ICU + Vent.
Hospitalized	270	13	32	41	29	1	12
Vent	5	9	4	5	2	0	5
ICU	18	3	7	7	2	0	3
ICU + Vent.	60	43	38	239	24	4	93
Deceased	12	6	3	10	5	0	7
Deceased + ICU	5	1	1	0	0	0	2
Deceased + ICU + Vent.	33	34	13	87	17	2	212

Figure 5.11: Confusion matrix for best performing classifier on radiomics features.

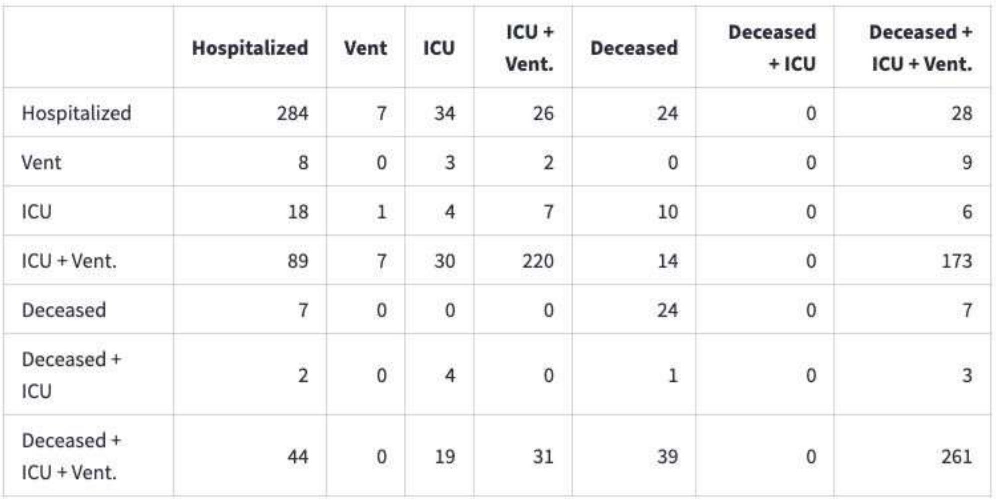
	TPR-Sensitivity (recall)	TNR-Specificity
Hospitalized	0.705	0.839
Ventilated	0.000	0.989
ICU	0.087	0.936
ICU + Ventilated	0.413	0.928
Deceased	0.632	0.938
Deceased + ICU	0.000	1.000
Deceased + ICU + Ventilated	0.662	0.785

Table 5.5: Class sensitivity and specificity for prediction based on merged feature sets.

5.1.4 Combined Prediction

The best SVC classifier using grid search ($C = 0.25$, $seed = 64532$, radial basis function kernel) scored 0.812 ± 0.024 on train and 0.785 on validation set. This indicates a slight improve from electronic health data features after merging with radiomics features. Sensitivity and specificity are again shown in Table 5.5, with the respective confusion matrix reported in Figure 5.12.

TU WIEN Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



	Hospitalized	Vent	ICU	ICU + Vent.	Deceased	Deceased + ICU	Deceased + ICU + Vent.
Hospitalized	284	7	34	26	24	0	28
Vent	8	0	3	2	0	0	9
ICU	18	1	4	7	10	0	6
ICU + Vent.	89	7	30	220	14	0	173
Deceased	7	0	0	0	24	0	7
Deceased + ICU	2	0	4	0	1	0	3
Deceased + ICU + Vent.	44	0	19	31	39	0	261

Figure 5.12: Confusion matrix for best performing classifier on merged feature sets predictions.

5.2 Qualitative Evaluation

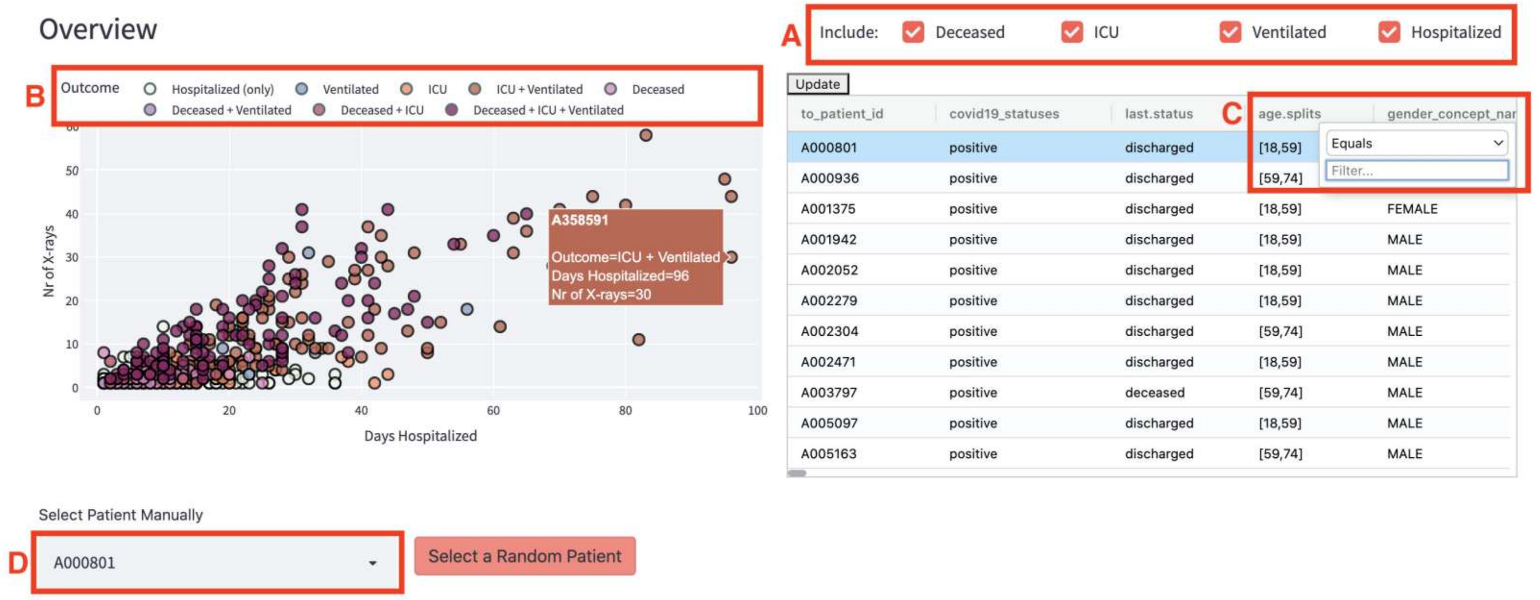
Here we present the final solutions of our VA application in form of screenshots. The solutions are validated following the Nested Model proposed by Munzner [2009] as presented in Section 3.5.3. We focused on the inner two levels of evaluation — namely the algorithm validation and the encoding and interaction techniques. Algorithm validation is done by logging runtime of our solutions. Encoding and interaction techniques are validated by conducting interviews with users from the general population group. Therefore we prepared two tasks that were to be solved in the general populations space of the application. Finally we gathered open feedback. Furthermore we used usage scenarios for all tasks defined in Section 4.2 to prove their usability [Isenberg et al., 2013].

5.2.1 Medical Experts

Medical 1: Overview

Figure 5.13 shows the first part of the medical experts page. The first task was about giving medical experts an overview of the data that may arrive at the application via a casual encounter. As the dashboard is conceptualized for analysis of medical health data as well as CXR images, an overview is given using two of the more easily understandable variables in the dataset. Therefore an aggregated view is generated that shows patients in a 2D scatterplot with number of hospitalized days and a count of CXR images (deduplicated) which lets users *explore*. Hospitalization outcomes are *encoded* via a mixed continous and categorical colormap. Ventilation gives blue elements, ICU admission red elements and mortal outcomes are darker. Outcomes can be quickly *filtered* either via selectboxes or the legend in the scatterplot itself (Figure 5.13**A** and Figure 5.13**B**). This

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



to_patient_id	covid19_statuses	last.status	age.splits	gender_concept_nar
A000801	positive	discharged	[18,59]	
A000936	positive	discharged	[59,74]	
A001375	positive	discharged	[18,59]	FEMALE
A001942	positive	discharged	[18,59]	MALE
A002052	positive	discharged	[18,59]	MALE
A002279	positive	discharged	[18,59]	MALE
A002304	positive	discharged	[59,74]	MALE
A002471	positive	discharged	[18,59]	MALE
A003797	positive	deceased	[59,74]	MALE
A005097	positive	discharged	[18,59]	MALE
A005163	positive	discharged	[59,74]	MALE

Figure 5.13: First part of medical experts dashboard. Used for tasks Medical 1+2. Simple queries for patient outcomes are available via checkboxes (**A**). An interactive legend in the scatterplot can be used to additionally filter outcomes (**B**). The interactive options in the table are highlighted as **C**. If a user of interest is found, it can be selected using a dropdown (**D**).

TU WIEN Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

lets a user compare different groups of patients easily.

In a tabular view on the right side shows the raw data. Here the user is able to additionally *filter* or select for certain attributes given the data. For example a user can specify a specific age range as highlighted in **C** or *select* only male patients with diabetes. The scatterplot will adjust to only conclude the current actively chosen data from all inputs and thus enables for repeatable interactions and *comparisons*.

Medical 2: Lookup

Given a user is familiar with the visualization or comes back to use the application, focus is set on the *discovery* now. Raw data can be *browsed* either in the table consisting of raw data or in the scatterplot to *identify* patients of interest. Those patients may be of a certain subset of patients. One specific task here could be to find similar patients to one that is currently at treatment at a medical facility or hospital. A doctor *selects* clinical data such as preconditions and demographic information using the *filter* tools highlighted in Figure 5.13**C** in Figure 5.13. Hovering over points of interest in the scatterplot shows additional information of a patient which helps *identifying* patients of interest. Those patients can then be *selected* in Figure 5.13**D**. This selection is defined as the output of this task and is the input of the next part of the medical experts page. The current selection is also highlighted in the data table (light blue background).

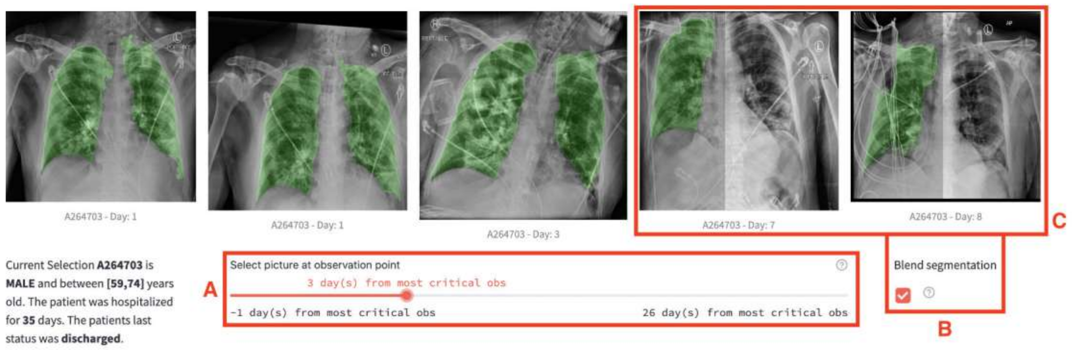
Medical 3: Decease Progression via CXR

Given a patient of particular interest has been selected in the previous task, a user has the ability to check the patients hospitalization stay by viewing available CXR data for this patient. This is shown in Figure 5.14. Image are not encoded in any special way, only the automatic image segmentation is colorized in the respective images. Users may *navigate* through available records by using a slider, highlighted in Figure 5.14**A**. Special days of interest can be selected via this slider as well, for instance the beginning or end of the hospitalization stay. Additionally the users have the option to disable the automatic lung segmentation which may cover interesting areas for a clinician via a checkbox (Figure 5.14**B**). This is indicated in the last two images which, are highlighted as Figure 5.14**C**.

Medical 4: Predict Outcome with CXR Image

The last task defined for the medical experts group was focused on giving an outcome prediction and as a form of decision making support. Here the user should have the ability to *produce* a prediction outcome based on *importing* a new image. The resulting page section in the VA application is depicted in Figure 5.19.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.


A264703 - Day: 1
A264703 - Day: 1
A264703 - Day: 3
A264703 - Day: 7
A264703 - Day: 8
C
A
Select picture at observation point
3 day(s) from most critical obs
-1 day(s) from most critical obs
26 day(s) from most critical obs
Blend segmentation
B

Current Selection **A264703** is **MALE** and between **[59,74]** years old. The patient was hospitalized for **35** days. The patients last status was **discharged**.

Figure 5.14: Medical experts page part two with image carousel. Users are able to discover the hospitalization stay and progression via CXR images for a patient of interest. The selected image is in the middle while the two previous and succeeding observations are shown as well, where selection is done using a slider **A**. Segmentation can be toggled on and off with a checkbox (**B**). **C** shows the difference between toggled segmentation displays.

5.2.2 Analytical Experts

Analytical 1: Overview

The overview for the analytical experts is provided by using scatterplots and heat maps as well as raw data tables. The final part of the page is shown in Figure 5.15. The scatterplot is modifiable by providing X and Y-axis variables through drop-down menus. Patients are encoded with color by hospitalization outcome and overplotting is battled by using opacity on the points, highlighted as Figure 5.15**C**. The legend of the scatterplot allows to filter for specific classes by clicking on the respective item. This allows for *identification* of outliers, patients of interest and connections between variables, as well as comparing different clusters with the help of *encodings*. Additionally a table is provided (Figure 5.15**D**) which lets users inspect the whole raw dataset.

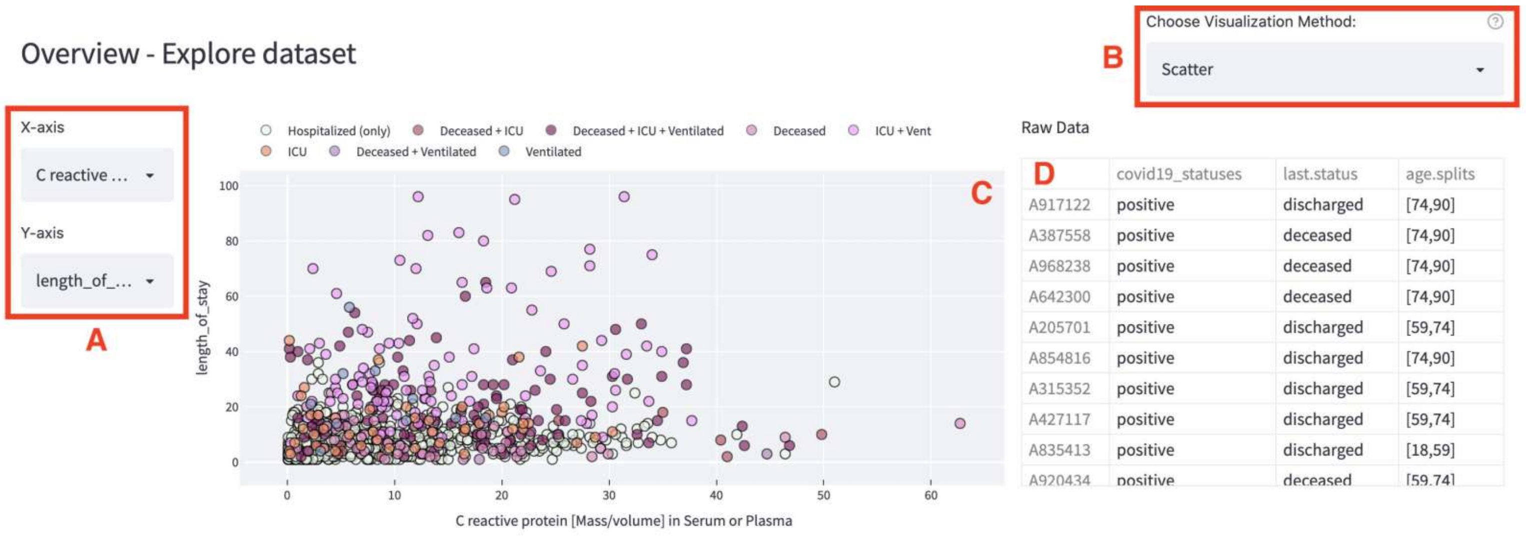
Another dropdown, highlighted in Figure 5.15**B**, enables changes the visualization style to a heatmap, shown in Figure 5.16. Here a typical correlation heatmap is shown for all variables in the dataset. Axis are removed to save space but hovering over the visualization provides tooltips so that interesting areas of the heatmap can be investigated. Zooming is supported by brushing and selecting rectangular areas in the plot. Additionally, another heatmap provides correlation values between features and hospitalization outcomes depicted as Figure 5.15**B**.

Analytical 2: Discover Imputation Strategies

Discovering imputation strategies is implemented in the second section of the analytical experts page, shown in Figure 5.17. A user is able to choose from various imputation strategies described and compared already in Section 3.3.2 and 4.1.1, while also specifying

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.


Overview - Explore dataset

X-axis
C reactive ...

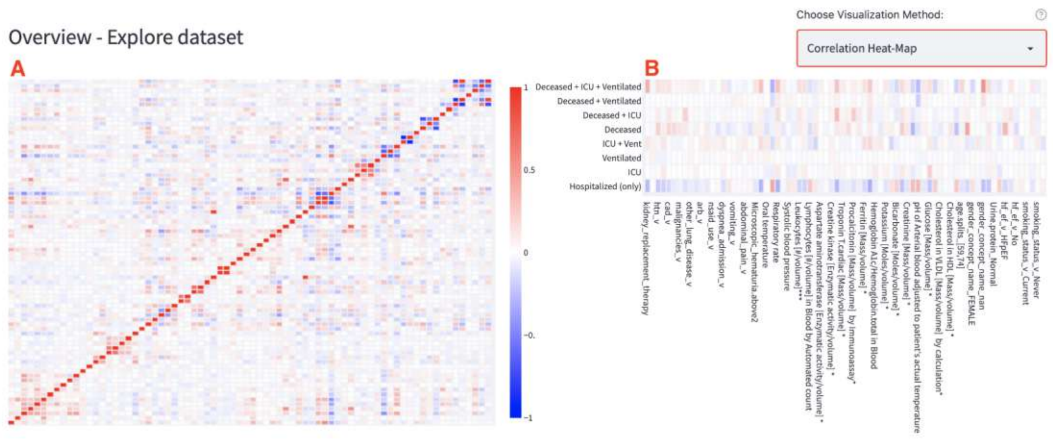
Y-axis
length_of_...

Choose Visualization Method:
Scatter

Raw Data

	covid19_statuses	last.status	age.splits
A917122	positive	discharged	[74,90]
A387558	positive	deceased	[74,90]
A968238	positive	deceased	[74,90]
A642300	positive	deceased	[74,90]
A205701	positive	discharged	[59,74]
A854816	positive	discharged	[74,90]
A315352	positive	discharged	[59,74]
A427117	positive	discharged	[59,74]
A835413	positive	discharged	[18,59]
A920434	positive	deceased	[59,74]

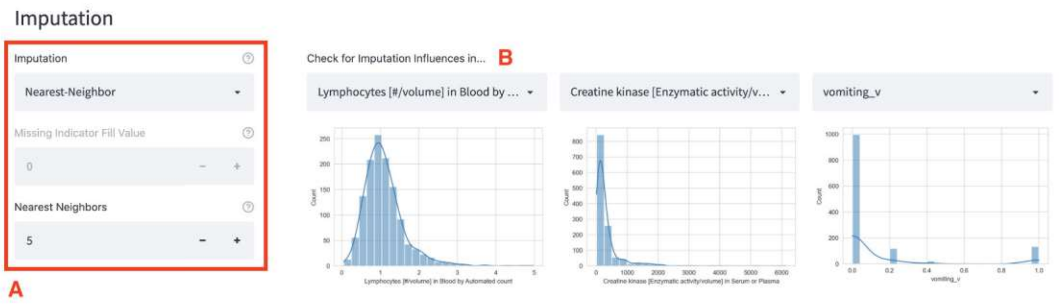
Figure 5.15: Overview for analytical experts. Users can choose between scatterplot shown here (**C**), and heat map overview (**B**) and select features of interest (**A**). Additionally a table with the raw dataset is presented (**D**)


Overview - Explore dataset
A
B
1
0.5
0
−0.
−1
Deceased + ICU + Ventilated
Deceased + Ventilated
Deceased + ICU
Deceased
ICU + Vent
Ventilated
ICU
Hospitalized (only)
kidney_replacement_therapy
htn_v
cad_v
malignancies_v
other_lung_disease_v
arb_v
nsaid_use_v
dyspnea_admission_v
vomiting_v
abdominal_pain_v
Microscopic_hematuria.above2
Oral temperature
Respiratory rate
Systolic blood pressure
Leukocytes [#/volume]***
Lymphocytes [#/volume] in Blood by Automated count
Aspartate aminotransferase [Enzymatic activity/volume]*
Creatine kinase [Enzymatic activity/volume]*
Troponin T.cardiac [Mass/volume]*
Procalcitonin [Mass/volume] by Immunoassay*
Ferritin [Mass/volume]*
Hemoglobin A1c/Hemoglobin.total in Blood
Potassium [Moles/volume]*
Bicarbonate [Moles/volume]*
Creatinine [Mass/volume]*
pH of Arterial blood adjusted to patient's actual temperature
Glucose [Mass/volume]*
Cholesterol in VLDL [Mass/volume] by calculation*
Cholesterol in HDL [Mass/volume]*
age.splits_[59,74]
gender_concept_name_FEMALE
gender_concept_name_nan
Urine.protein_Normal
hf_ef_v_HFpEF
hf_ef_v_No
smoking_status_v_Current
smoking_status_v_Never

Choose Visualization Method:

Correlation Heat-Map

Figure 5.16: Analytical overview using correlation matrix heatmaps. **A**: correlation matrix for all features. **B**: correlation values of all features for each hospitalization outcome.


Imputation
A
B
Check for Imputation Influences in...
Count
Lymphocytes [#/volume] in Blood by Automated count
Creatine kinase [Enzymatic activity/volume] in Serum or Plasma
vomiting_v

Imputation: Nearest-Neighbor

Missing Indicator Fill Value: 0

Nearest Neighbors: 5

Lymphocytes [#/volume] in Blood by ...

Creatine kinase [Enzymatic activity/v...

vomiting_v

Figure 5.17: Comparing imputation options. A: Choosing imputation method and parameters. B: Displaying three changeable features with their distribution.

hyperparameters. This is highlighted as Figure 5.17**A**. Per default three variables of different distributions are shown encoded as histograms, which can be modified as well (Figure 5.17**B**). With this, users directly experience the impact of the imputation method. Changing the imputation method produces a freshly imputed dataset (*Task 2.1*) and takes less than 0.5s, besides multiple-imputation which takes about 10s.

Analytical 3: Discover Patient Clustering Strategies

Figure 5.18 shows the clustering section of the final implementation. Users again are able to *select* different clustering strategies, Ward and k-means, and *change* hyperparameters. The data used here is the previously imputed data from strategies chosen before. Representation is shown in a 2-D scatterplot where the user can choose for 2-D PCA data as well as t-SNE dimensionality reduced transformation. On the right side, two bar plots show unsupervised clustering metrics, Silhouette and CH scores. Settings are shown

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

when hovering over the respective bar blots. With this solution, users are able to *select* and *change* between clustering strategies, compare their outcomes by using scatterplots and barplots for metrics, as well as a stacked barplot Figure 4.29 potentially showing differences in features, solving tasks from requirements *Analytics 3.1* and *Analytics 3.2*. Clustering takes about $2-3$ seconds.

Analytical 4: Automatic Lung Segmentation

For this task we chose to use the same design as we did for Medical 4. In addition to being able to upload a CXR, we also provide the option to change the pre-processing strategy. Predictions are again presented using a barplot. Uploading and predicting takes about 7 seconds. The image needs to be pre-processed, and submitted through the network, before radiomics features are extracted from the new image and predicted using the pre-trained model, so a lot is going on here after uploading a picture.

Analytical 5: Discover Radiomics Features

In the condensed view shown right side of Figure 5.20 radiomics features correlation values with target hospitalization outcomes are shown, encoded in a 2-D heatmap. Here, the hovering is also highlighted. Since the radiomics features are not very expressive by their name alone, we added a textual description with a link to the official documentation that describe each feature on the left side. The correlation values can be sorted by target which can be chosen using the drop-down menu. The full view display all radiomics features which adds another layer of explainability.

Analytical 6: Prediction

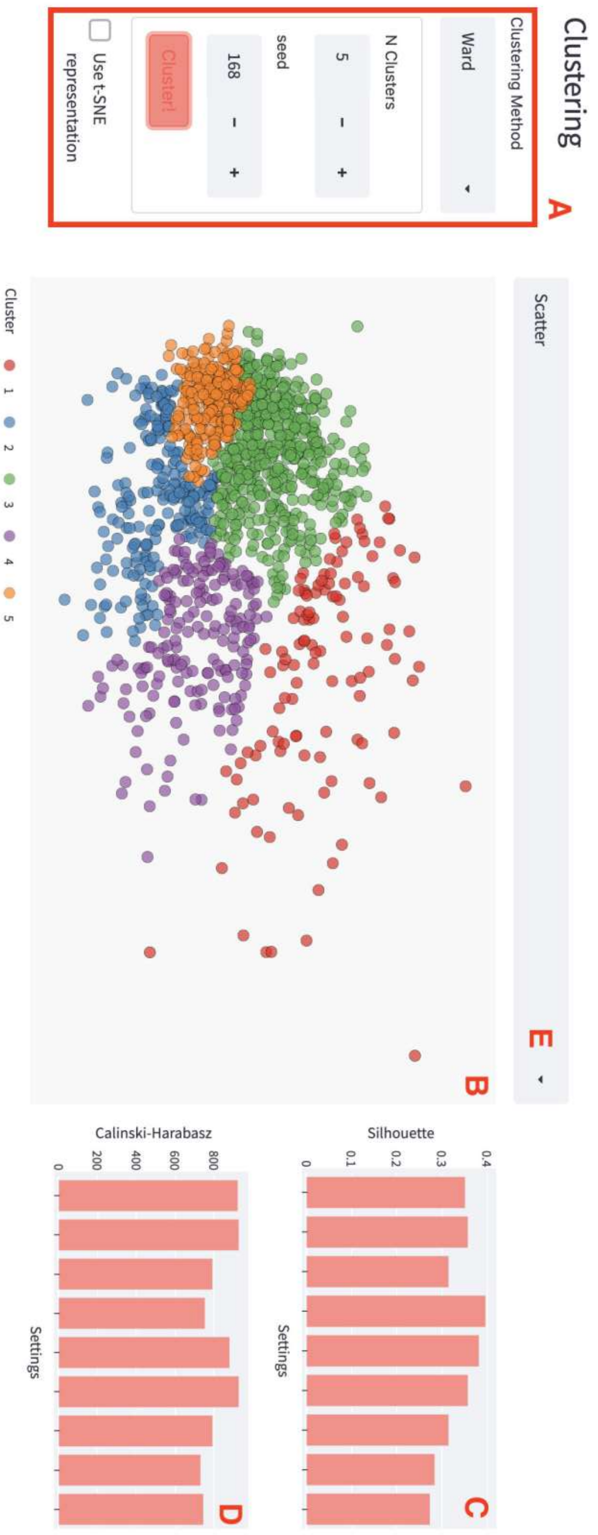
For the prediction task, our final design is displayed in Figure 5.21. Here the dataset as well as a classifier of the five proposed classifiers in this thesis can be chosen. Classifier can be optimized by changing hyperparameters using sliders and text inputs as highlighted. The previously chosen imputation strategy in Task 2 is chosen here as well to impute data after performing a train-validation split. No cross-validation is provided here as some of the classifier trainings took several hours, especially for the radiomics features. Models can be trained and compared using the lineplot Figure 5.21**B**, which shows common metrics on the validatoin set. In addition to that, we provide the confusion matrix for the validation data, allowing users to *compare* different models.

5.2.3 General Pop

General 1: Overview

For the general population we created an infographic. To keep it simple, we only included information that we believed was common sense and where information was available in the given dataset, including certain preconditions, demographic information and lifestyle

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub


Clustering

Clustering Method: Ward

N Clusters: 5

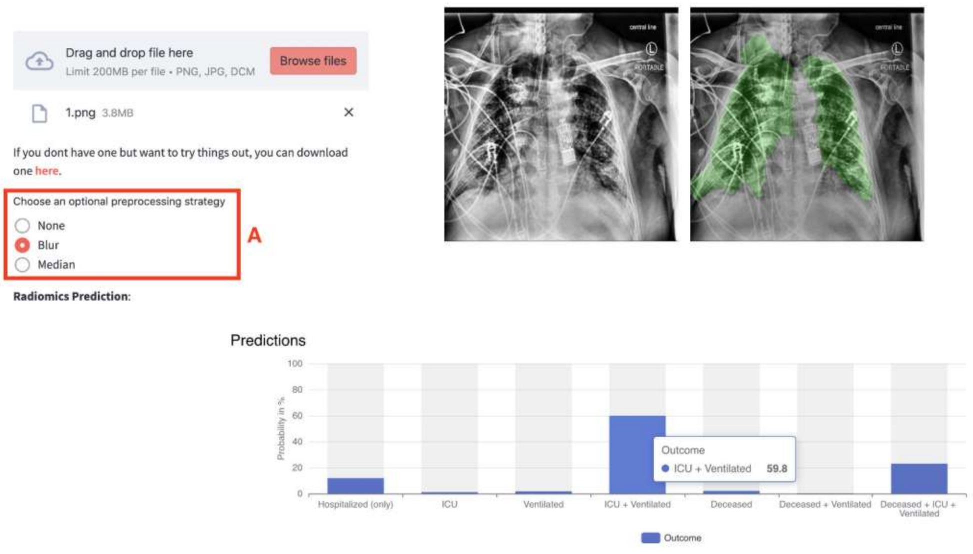
seed: 168

Cluster!

Use t-SNE representation

Figure 5.18: Clustering and comparison of results. **A**: Clustering strategy and hyperparameters, scatterplot dimensionality reduction. **B**: Scatterplot with 2D, dimensionality reduced data and color encoded clusters. **C**: Silhouette scores for unsupervised clustering, consecutive clustering runs are appended here. Used settings are shown when hovering over the bars as a tooltop. **D**: Calinski-Harabasz scores, similar to **C**. **E**: Change from scatterplot to stacked barplot representation, as shown in Figure 4.29 to display cluster statistics.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub


Drag and drop file here
Limit 200MB per file • PNG, JPG, DCM
Browse files
1.png 3.8MB

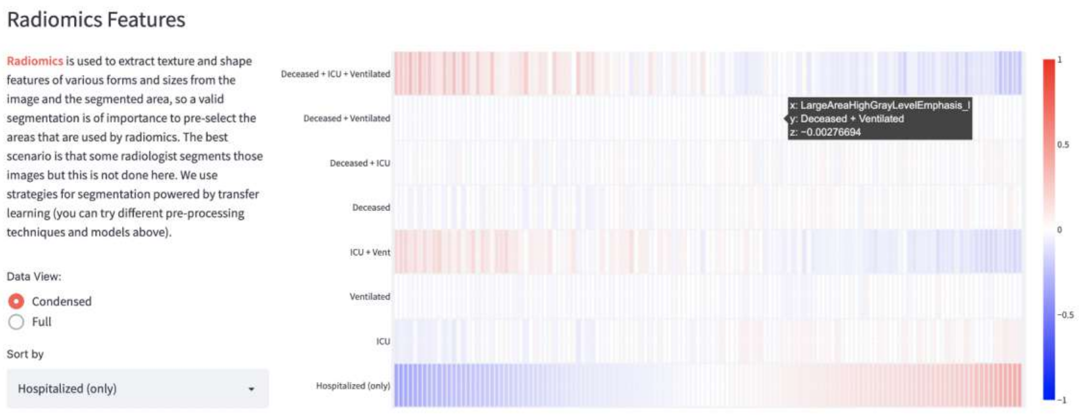
If you dont have one but want to try things out, you can download one here.

Choose an optional preprocessing strategy
None [✓ Blur] Median

A

Radiomics Prediction:

Figure 5.19: CXR image upload and outcome prediction with optional preprocessing selection **A**. This selection is not removed in the the medical user groups page.


Radiomics Features

Radiomics is used to extract texture and shape features of various forms and sizes from the image and the segmented area, so a valid segmentation is of importance to pre-select the areas that are used by radiomics. The best scenario is that some radiologist segments those images but this is not done here. We use strategies for segmentation powered by transfer learning (you can try different pre-processing techniques and models above).

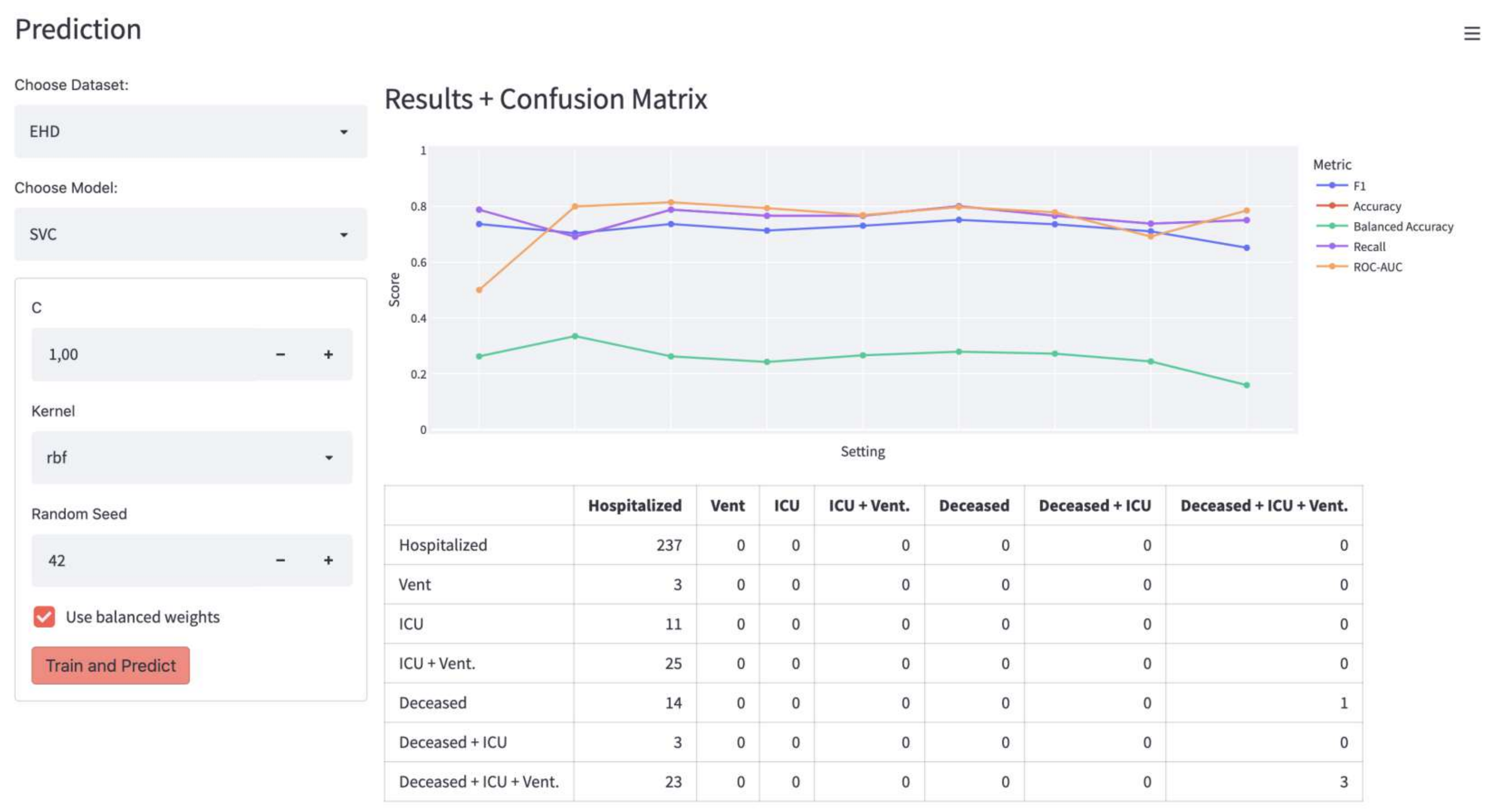
Data View:
[✓ Condensed] Full

Sort by
Hospitalized (only)

x: LargeAreaHighGrayLevelEmphasis_l
y: Deceased + Ventilated
z: −0.00276694

Figure 5.20: Comparing radiomics features. Condensed view showing correlation between targets and features, full view showing correlation matrix for all features.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub



	Hospitalized	Vent	ICU	ICU + Vent.	Deceased	Deceased + ICU	Deceased + ICU + Vent.
Hospitalized	237	0	0	0	0	0	0
Vent	3	0	0	0	0	0	0
ICU	11	0	0	0	0	0	0
ICU + Vent.	25	0	0	0	0	0	0
Deceased	14	0	0	0	0	0	1
Deceased + ICU	3	0	0	0	0	0	0
Deceased + ICU + Vent.	23	0	0	0	0	0	3

Figure 5.21: Prediction model training. **A**: User inputs, including dataset, model and hyperparameters. **B**: Metrics displayed in a lineplot for comparison. **C**: Confusion matrix of predictions on a validation split.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

choices such as smoking habits. When abstracting data for the general population, we tried to encode the data in special ways that fulfil our need to communicate differences.

The graphic highlighted as Figure 5.23**C** shows this abstracted information for each cluster. Preconditions includes diabetes, heart, lung and kidney disease as well as malignancies as depicted in the legend Figure 5.23**B**, or Figure 4.37. For each of the preconditions, an expressive icon was chosen that encodes it. If the precondition was present in the cluster more often then in others, it was colorized, if not it was gray-scaled. More detailed information on the differences is given Section 5.1.1. As for demographics, we used the given age groups in the dataset and encoded patients age distributions accordingly by corresponding gender symbols. Where one gender was significantly more present, the size has been adjusted. Additionally we created the possibility to get a textual representation of the infographic as highlighted as Figure 5.23**A**. The explanations are shown in detail in Figure 5.22.

Finally, we chose a scatterplot representation highlighted in Figure 5.23**D**: Instead of sampling we computed outcome probabilities for each type of outcome for each respective cluster. This implicitly meant that we were unable to visualize dedicated patients outcomes and had to summarize and round values in order to encode in a meaningful way. Therefore we included non-optional textual description below the actual plots to improve understandability. For instance, for Group 3 the description is:

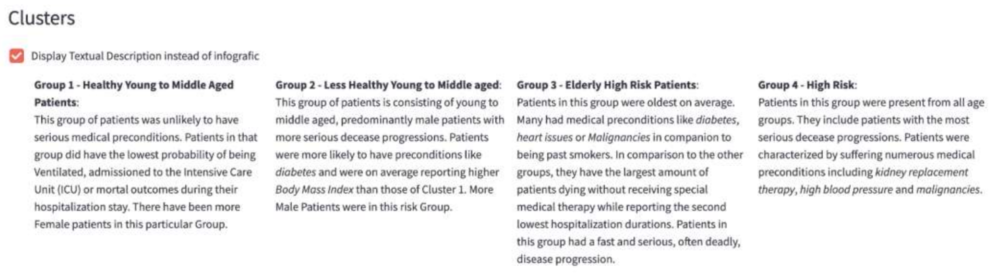
> Out of a 100 Patients in this group, 18 patients were admissioned to the ICU, 12 patients had to be Ventilated, and 25 died.

This translates to the following in the scatterplots: out of a 100 patients, each patients could either be admissioned to the ICU, ventilated or had a mortal outcome. For patients sharing multiple outcomes, we either had to mix outcomes in the encoding (for example one scatter with multiple colors) or multiple points per patient. We chose the latter, so for each group this translates to 300 points. Another possibility here would have been to sample patients from the groups (with replacement) so that real patients could have been displayed. This could have the negative effect of skewing the distribution once again though.

The clusters were reordered in ascending order by their actual risk. In particular this meant that cluster 2 and 3 have been swapped. Furthermore we used groups as the notation of the clusters, since it is not given that layman users are aware of what a cluster actually is.

With this, we *present* groups of patients that should indicate which groups are most similar to themselves or are of interest. We furthermore engage users to try out different inputs for the next task, as we encourage interaction by abstract *encodings* which may lead to a user inputting not only themselves but beloved ones and sharing the outputs and findings with others.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub


Clusters

✓ Display Textual Description instead of infografic

Group 1 - Healthy Young to Middle Aged Patients:
This group of patients was unlikely to have serious medical preconditions. Patients in that group did have the lowest probability of being Ventilated, admissioned to the Intensive Care Unit (ICU) or mortal outcomes during their hospitalization stay. There have been more Female patients in this particular Group.

Group 2 - Less Healthy Young to Middle aged:
This group of patients is consisting of young to middle aged, predominantly male patients with more serious decease progressions. Patients were more likely to have preconditions like *diabetes* and were on average reporting higher *Body Mass Index* than those of Cluster 1. More Male Patients were in this risk Group.

Group 3 - Elderly High Risk Patients:
Patients in this group were oldest on average. Many had medical preconditions like *diabetes*, *heart issues* or *Malignancies* in companion to being past smokers. In comparison to the other groups, they have the largest amount of patients dying without receiving special medical therapy while reporting the second lowest hospitalization durations. Patients in this group had a fast and serious, often deadly, disease progression.

Group 4 - High Risk:
Patients in this group were present from all age groups. They include patients with the most serious decease progressions. Patients were characterized by suffering numerous medical preconditions including *kidney replacement therapy*, *high blood pressure* and *malignancies*.

Figure 5.22: Textual information optionally provided instead of the infographic, toggable by the checkbox.

General 2: Hospitalization Outcome Prediction

Given the input for this task stems from the previous task, *General 1*, the VA application is also meant to intertwine usage in between both areas of the page. The design is shown in Figure 5.24. User inputs are gathered via conventional web-tools like check-boxes and drop down menus. Where multiple inputs are available that are abbreviated in ways to save space, help pop-ups are implemented as illustrated in Figure 4.36. The group chosen in the preceding task is now used as the prediction model is trained on the whole datasets variables. Here, we only ask for a subset of inputs as it would not be feasible to ask for detailed medical variables that need prior laboratory examination. Those values are instead merged to the inputs and used as inputs for the prediction model in the background. Group-wise means are used here. Additionally there is the possibility to set this to *Automatic* which will instead use a k-Nearest Neighbour approach using the inputs provided to estimate the remaining data from the dataset. Prediction results are presented in form of a doughnut plot, as most users should be familiar with pie plots, although concerns exist about the use of pie charts when displaying percentages (see Section 4.7.1).

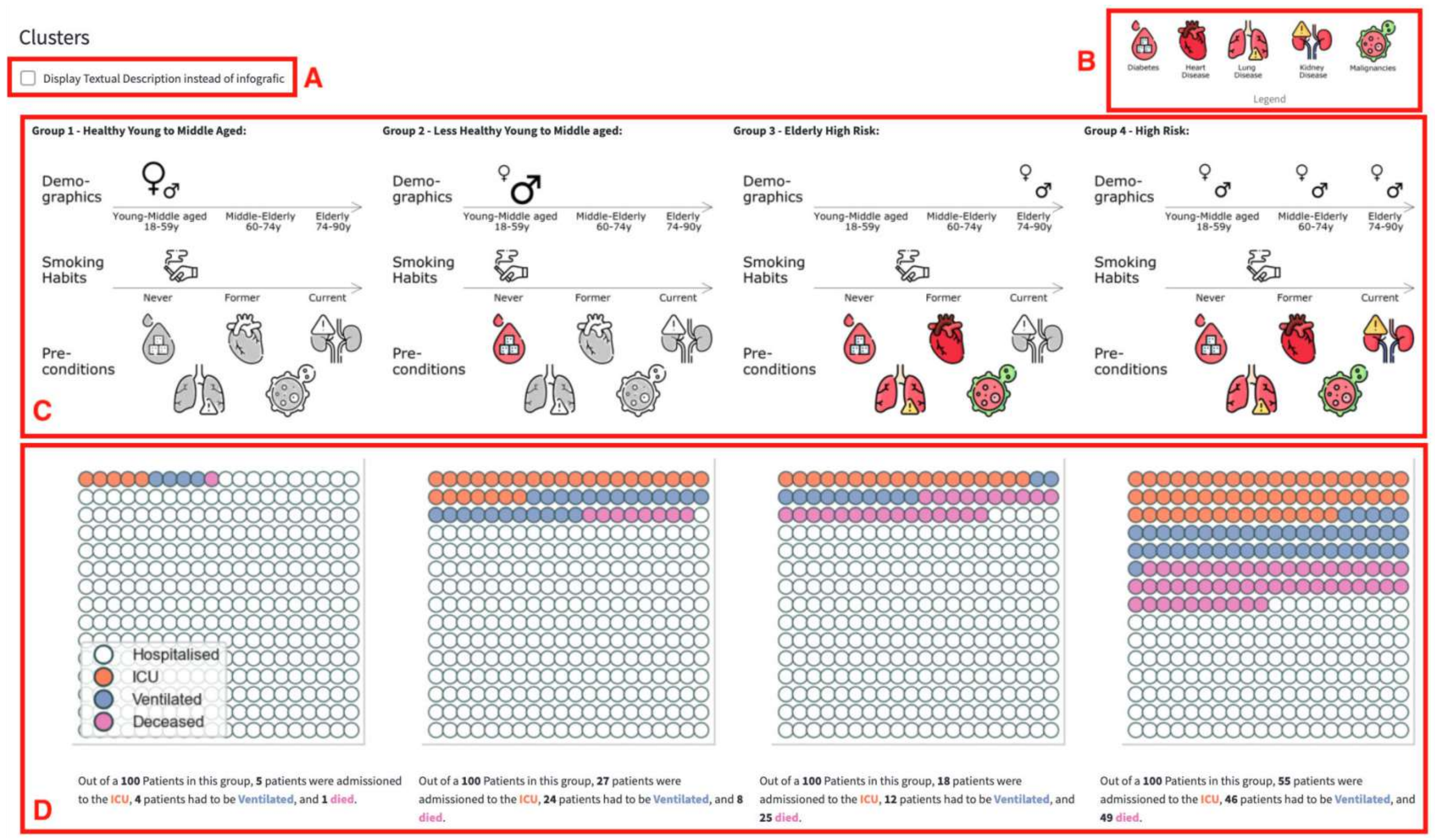
In addition to that we also display the five most similar patients estimated from the user inputs per nearest neighbour search. This should improve prediction explainability by providing insights into *consistency* as well as *stability* of a model and is proposed by research in the explainable machine learning area [Molnar, 2022, Section 3.5 and 3.6].

5.2.4 Runtime Performance

As for performance, we measured runtime directly in the code with debug outputs before and after function calls or dashboard interaction. Interactive plots built with *Plotly* update in almost real-time without active re-rendering, as used in Figure 5.13, Figure 5.15**C**, Figure 5.16**A**). Building the carousel plot with images (Figure 5.14) runs for 0.1s to 0.8s depending on the number of images available for the selected patient. Uploading an image and processing it for automatic segmentation takes up to 3s, while radiomics feature extraction is fast with 0.3s and prediction with less than 0.01s.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



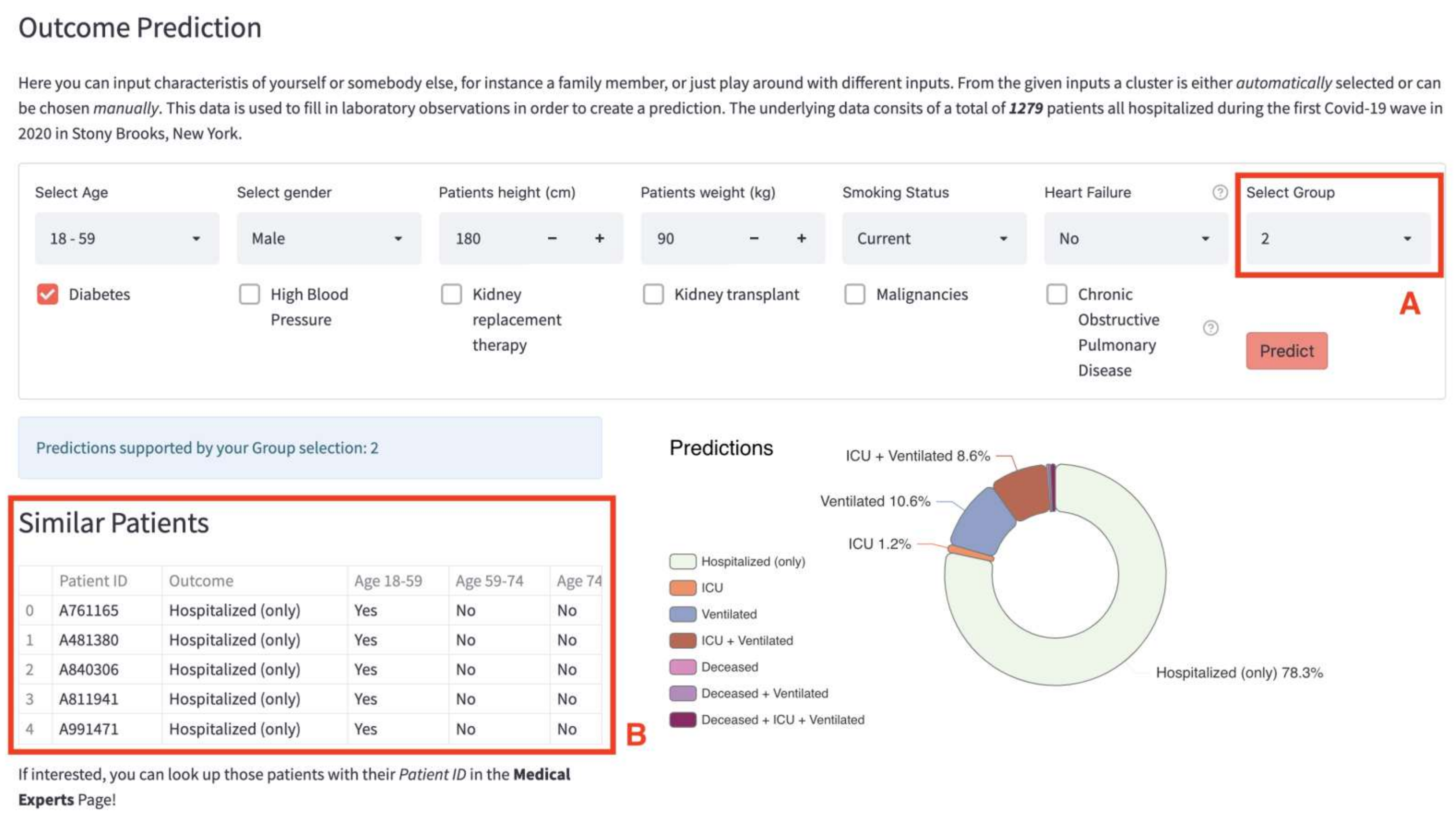
Out of a **100** Patients in this group, **5** patients were admissioned to the ICU, **4** patients had to be Ventilated, and **1** died.

Out of a **100** Patients in this group, **27** patients were admissioned to the ICU, **24** patients had to be Ventilated, and **8** died.

Out of a **100** Patients in this group, **18** patients were admissioned to the ICU, **12** patients had to be Ventilated, and **25** died.

Out of a **100** Patients in this group, **55** patients were admissioned to the ICU, **46** patients had to be Ventilated, and **49** died.

Figure 5.23: General population overview. Main focus on abstract communication of patient groups and risk factors for hospitalized patients. **A**: Tickbox to switch between infographic and textual description. **B**: Legend for the infographics preconditions. **C**: Actual infographic showing the four clusters of patients derived from clustering as discussed in Section 5.1.1. **D**: Scatterplot representation of a sample of 100 patients for each group of patients, indicating the outcome by color encoding.



Outcome Prediction

Here you can input characteristis of yourself or somebody else, for instance a family member, or just play around with different inputs. From the given inputs a cluster is either *automatically* selected or can be chosen *manually*. This data is used to fill in laboratory observations in order to create a prediction. The underlying data consits of a total of ***1279*** patients all hospitalized during the first Covid-19 wave in 2020 in Stony Brooks, New York.

Select Age	Select gender	Patients height (cm)	Patients weight (kg)	Smoking Status	Heart Failure	Select Group
18 - 59	Male	180	90	Current	No	2

[✓] Diabetes ☐ High Blood Pressure ☐ Kidney replacement therapy ☐ Kidney transplant ☐ Malignancies ☐ Chronic Obstructive Pulmonary Disease

Predictions supported by your Group selection: 2

Similar Patients

	Patient ID	Outcome	Age 18-59	Age 59-74	Age 74
0	A761165	Hospitalized (only)	Yes	No	No
1	A481380	Hospitalized (only)	Yes	No	No
2	A840306	Hospitalized (only)	Yes	No	No
3	A811941	Hospitalized (only)	Yes	No	No
4	A991471	Hospitalized (only)	Yes	No	No

If interested, you can look up those patients with their *Patient ID* in the **Medical Experts** Page!

Figure 5.24: General population outcome prediction. Using inputs from a user to predict the outcome, also making use of the previously presented patient groups. **A**: chose group from Figure 5.23. **B**: Additionally, the five most similar patients are gathered from the dataset using k-Nearest Neighbour search.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Imputation performance is depending on the method. Except for multiple imputation, which may take up to $12s$, methods run for less than $0.8s$, and only need to run once at a preprocessing step. Interactive clustering is happening in near-real time with less than $0.5s$. Classifier training and prediction, as shown in Figure 5.21, takes less than $1s$ in most cases, depending on the hyper parameters used. Cross validation is not supported here, because especially for radiomics features it took minutes to finish, which would limit the responsiveness of the dashboard as it is not trivial to build a multi-thread environment in streamlit.

Most of the methods results and trained models as well as user uploads are cached manually in the background to improve iterative tasks, such as *Medical 1*, *Analytical 2.2*, *3.1* and *6*. For instance if a new clustering method is tried out for task *Analytical 3.1*, the previously selected imputation is re-used rather than re-trained and reduces runtime to less than $0.03s$.

When comparing to the thresholds provided in Section 3.5.3 defined by Nielsen [1994], we observe a freely navigable application with most wait times below 1s. A percentage-based indication is not provided in the case of the multiple imputation which showed the highest wait times, as it is the only case. All wait times are indicated by streamlit itself, as shown in Section 4.7.2.

5.2.5 Case Studies

Finally we conducted case studies with a small sample of users from the general population user group ($N = 6$, 3 male, mean age 28 ± 2.4). Therefore two assignments were prepared to validate encoding and visualization design on the previously specified user group. Before the assignments, the users had time to get familiar with the application and were able to ask questions.

Case 1 - Till: A fried of mine is Till. He is 28 years old and has diabetes. He is a current smoker and not very sporty. To which risk group would you assign Till to? What would be his outcome prediction if he caught COVID-19 and was hospitalised?

The idea was to assign the patient Till to group 2, as he was male, of young to middle age and had diabetes. This can be seen in Figure 5.23**C**, where group two has a bigger demographics male glyph, and the only precondition that is color coded is diabetes. The additional information of being not very sporty and a current smoker lead four interviewees to choose group 4 over group 2 instead, wavering between the two groups when asked. This also was a legitimate answer given the introduced graphic describing the groups. Inputting the patients data for the prediction was successful in all cases and the prediction was successfully interpreted using the doughnut chart.

Case 2 - Aunt Mary-Ann: Mary-Ann is in her sixties. She loves cake and thus struggles with diabetes. As a former smoker, she also struggles with high blood pressure and coronary artery disease. To which risk group would you assign her to? What would be her outcome prediction if she caught COVID-19 and was hospitalised?

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

For this case, the appropriate answer would have been to assign this test patient to group 4. This is again depicted in Figure 5.23**C**. When only looking at the preconditions here, group 3 and 4 would be legitimate. Given that group 3 only has gender glyphs in the elderly region (74 to 90 years old), while group 4 has them everywhere, group 4 would be the correct assignment. Two interviewees chose group 3 instead, which indicated misinterpreting the depicted age information inside the infographic. Interviewees did not do well when inputting specific pre-conditions like coronary artery disease, whose abbreviation CAD was just an option in a drop-down menu. Although a help icon was provided which explained this, interviewees rather asked for exact guidance in this case.

General Feedback

For feedback, all interviewees positively outlined the possibilities of inputting your own data and playing with the application. Four mentioned the doughnut plot was *intuitive* and *easy to understand*, while two additionally pointed out the chosen colors, which successfully divided outcome categories in their opinion. One mentioned the doughnut chart could have been a little bit bigger. Four also mentioned the infographic in addition to the group-wise outcome scatterplot was *nice*, *intuitive* or *easy to understand*. Two found that it was *confusing*, *hard to understand* or required *detailed attention*. As a help for this, we initially provided additional textual feedback which could have been used in order to get a textual description instead of the graphic. Only one interviewee tried this out without previously hinting and stated that it indeed makes it very clear to understand afterwards.

Three interviewees wished there was a more granular differentiation between the age groups, in particular they stated the age group of 18 to 59 years was too large. Two interviewees would have found it helpful to have information about patients fitness or sportiness in the groups. One pointed out it would be helpful to update the data with vaccination data in order to show it to vaccine sceptical people. Most interviewees did not make much use of the similar patients table provided. One stated that it was more *confusing* than helping him understanding the predictions. He proposed to make this information optional by providing a link to another page for instance.

5.2.6 Discussion & Limitations

The solution proposed here was motivated by the ongoing COVID-19 pandemic and is a highly interdisciplinary topic, ranging from machine learning and predictive analytics, to communication of results to different user groups including laymen users, while handling high volume medical data.

The SBU COVID-19 dataset used in this thesis comes with some limitations. The dataset itself is not considered a large dataset by machine learning definition, hence only having 1297 records with patients that have CXR images available as well. But for medical purposes it can be considered quite big. One downside is that no data of non-hospitalized patients is included, or available in the same way (including CXR), which introduces

TU Bibliothek, WIEN Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

some bias towards already worse disease progressions, with only serious illnesses having the probability of being included in the dataset. Additionally, no information about the vaccination status is available in the dataset given it was recorded prior to vaccine availability. Given the supposedly high protection from serious illness through vaccinations [Polack et al., 2020; Anderson et al., 2020], the data could look different when recorded again after successful vaccination strategies haven taken place.

Segmentation of the available images showed that real-life applications and real-life data introduce certain problems for automatic solutions by including artefacts which are not included in clean, openly available datasets prominently used for transfer learning processes (hence Section 2.2). Additionally, the CXR data used in this thesis was reported to be recorded anterior to posterior (AP), when normally data is recorded posterior to anterior (PA), also also the transfer learning datasets MC and SD were recorded PA. Additionally, it is not clear which exact X-ray device model was used to record the SBU data. How different device types can influence transfer learning and segmentation has already been researched by Vidal et al. [2021]. Figure 5.25 shows how inter-device type learning helped reducing segmentation artifacts that are similar to the ones reported in the dataset of this thesis.

While working with the medical image data, we decided to ignore the temporal aspects of the images (and also dropped temporal information from the electronic health data, f.i. days hospitalized). Whilst the information of how long a patient was already hospitalized and when the image was taken would have been available, we decided to build prediction on segmentation of all images. This introduces some bias as to how the features are extracted by *PyRadiomics*: given the already progressed disease is sometimes clearly visible in terms of artefacts such as cables from the ICU admission (hence Section 5.1.3) in the images and could have been mitigated by pre-selecting only clean images. We prove that radiomics features are able to detect those artefacts and predictions are possible based on this data to some degree, but for patients with images already from the ICU this is in danger of being a *confirmation bias*, meaning we confirm what was already known prior. Our results in terms of ROC-AUC outperforms results of previous work done by Varghese et al. [2021]. Their work was based on the same data, but only kept the first two observations (CXR images) per patient, removing this bias. Following this, our application is able to predict the outcome of a patient given *any* state of the hospitalization, whilst the solution of Varghese et al. [2021] would only be legitimate to be used in the beginning of the hospitalization stay.

Lastly, it is key to mention that non of the work that was done in this thesis was done in cooperation with medical experts or clinicians. This is a downside which was discussed already in research done about COVID-19 in various works [Roberts et al., 2021; Heaven, 2021].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

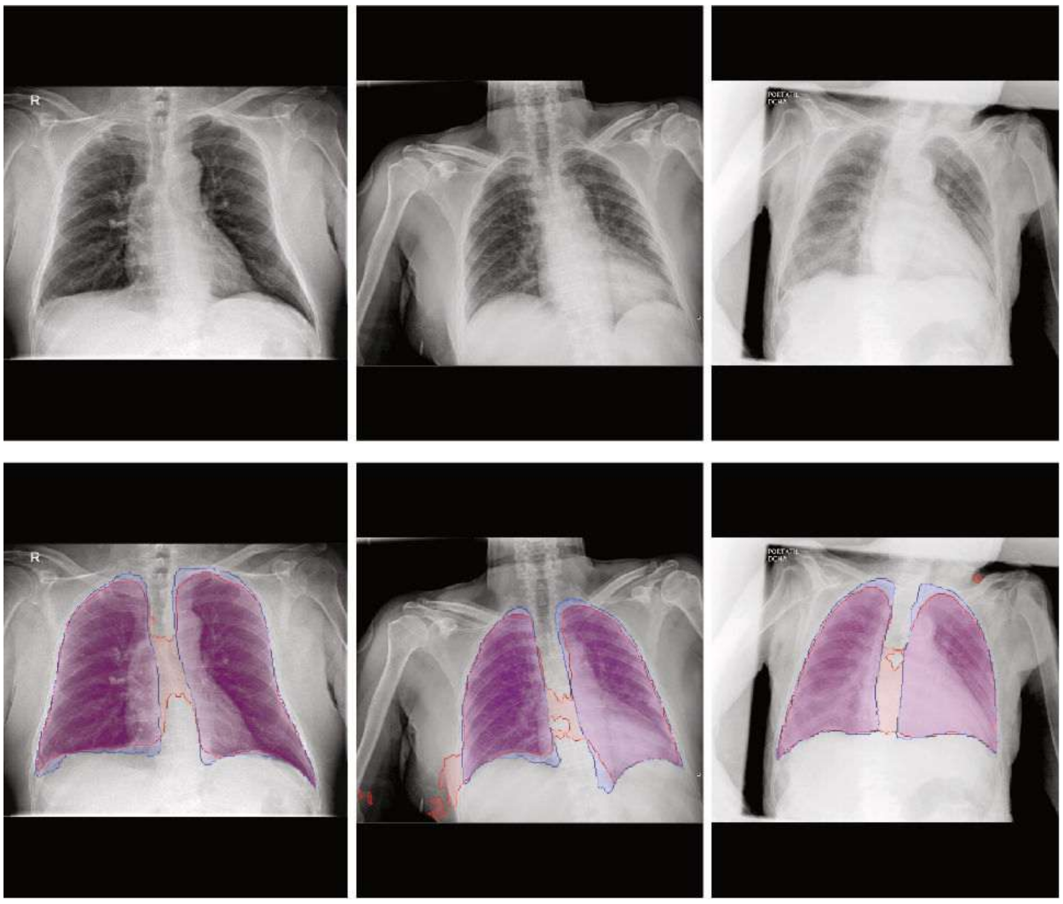


Figure 5.25: Results from Vidal et al. [2021]. Red regions depict automatic segmentation from inter-domain transfer learning, blue regions are results from inter-device-type transfer learning. Inter-device-type learning helped reducing artifacts.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 6

Conclusion and Future Work

Revisiting the research questions defined in Section 1.2, this work contributes to the state of the art in various ways:

Segmentation: We show that automatic lung segmentation using transfer learning is possible in real-life data scenarios with artefacts, such as COVID-19 related CXR scans. We reason with qualitative and quantitative assessment, where we proposed a **novel** way of quantitatively **evaluating segmentation error without GT** using a *reverse transfer learning* approach. Furthermore, we show that **reproducible radiomics features** such as Informational Measure of Correlation (Imc) or Gray-Level Non Uniformity (GLN) extracted from CXR images can be used as biomarkers on CXR images regarding COVID-19 disease prediction, given the features provide methods to detect irregularities and complexities of gray level intensities in a region of interest.

Prediction: When comparing to state-of-the-art research, we propose a granular prediction algorithm which performs in-line with previously proposed solutions predicting COVID-19 (hospitalization) outcomes. We **indicate that merging image features with electronic health data may improve prediction results** by a short amount, from electronic health data to merged features with $0.780 \rightarrow 0.785$ and from sole image features to merged with $0.776 \rightarrow 0.785$ ROC/AUC. We clustered patients into groups that inherited different patients proven by significance analysis, leading to four different risk groups of patients.

Visual Analytics: We created a **novel VA application** to present findings in an effort to **increase risk perception** in the general population. In addition, the application provides **decision making support for medical experts** and **analysts** working with this COVID-19 data. We estimated algorithm runtime by measuring our system between interactions as proposed by Munzner [2009], indicating interactive usage is possible while the runtime of almost all algorithms supports interactive data analysis with wait times of less than 1s. We evaluated design choices by conducting **user interviews** using case

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

studies with users from the general population. Evaluation indicated good usability and understandable presentation of prediction results using a doughnout plot and color encodings. Users found it interesting to input different data and experience how the outcome changes given different inputs. An infographic (Figure 5.23**C**) aimed to present different risk groups had mixed effects, with the majority of users being able to use it as designed. We additionally provided **usage scenarios** for the remaining user groups conducted by ourselves.

The thesis opens up interesting pathways for future work and scientific research. On the fly and explainable segmentation of *real-life* data is a very important topic in the future, as medical workers should trust and rely on applications built to reduce manual work. It is of further importance that applications are deployed in practice in order to measure their value in real-life scenarios and to find shortcomings and improve their results while improving trust of medical workers and experts into this systems. This also includes evaluation of machine learning solutions beyond development in practice, building robust solutions to measure performance when no GT is available. The pandemic ultimately put a lot of stress on medical health care workers [Ruiz-Fernández et al., 2020; Gavin et al., 2020] which may increase the urgent need of developing automated and robust machine learning assisted solutions in the future.

Future work on the SBU data can include other areas such as predicting the hospitalization length of stay for instance. Furthermore, gathering data including vaccination status of patients would yield a promising way of mitigating vaccination hesitancy, when included in applications like the one proposed in this thesis. Further work would be required in a long-term and wide user study to determine the usability and insights (and therefore the impact) that visual analytics solutions can provide to the general population with regard to COVID-19 risk perception and education. As hinted in the limitations section, a natural next step would be to evaluate the application with medical experts to estimate usability.

All in all, this is a first step towards building dedicated VA applications for different user groups, from providing interactive analytical possibilities to raising awareness into publicly available COVID-19 data which may be helpful in battling and controlling the pandemic.

The work conducted in this thesis has been submitted and accepted for presentation at EG VCBM 2022 (the 12th Eurographics Workshop on Visual Computing for Biology and Medicine) under the title "Predicting, Analyzing and Communicating Outcomes of COVID-19 Hospitalizations with Medical Images and Clinical Data" by Stritzel and Raidou, and will be published at the Digital Library of Eurographics[1] *in September 2022.*

[1] https://diglib.eg.org/

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

APPENDIX A

Appendix

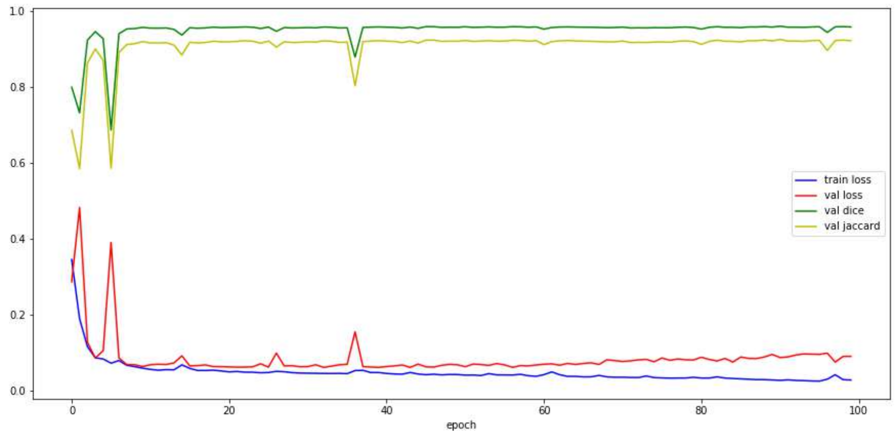


Figure A.1: Model A loss curves and validation metrics on the training process. TernaususNet VGG11, no rotation, Adam optimizer, no early stopping.

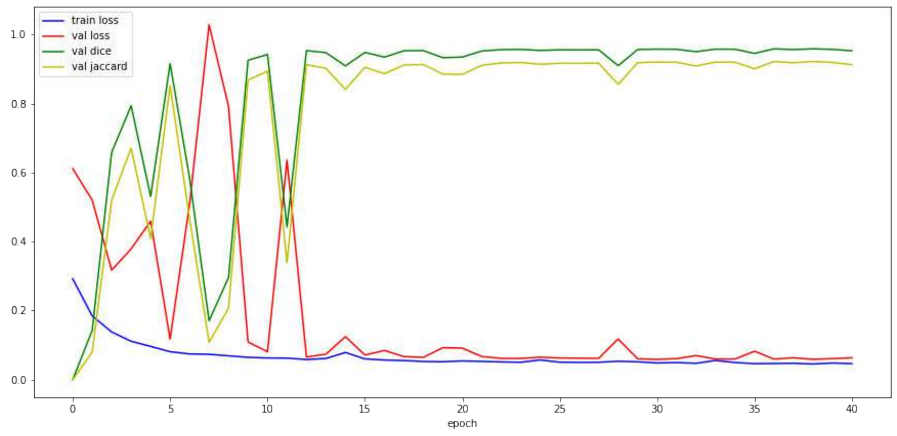


Figure A.2: Model B loss curves and validation metrics on the training process. TernaususNet VGG11, ±30, AdamW optimizer, early stopping after 10% of maximum epochs do not yield performance increase.

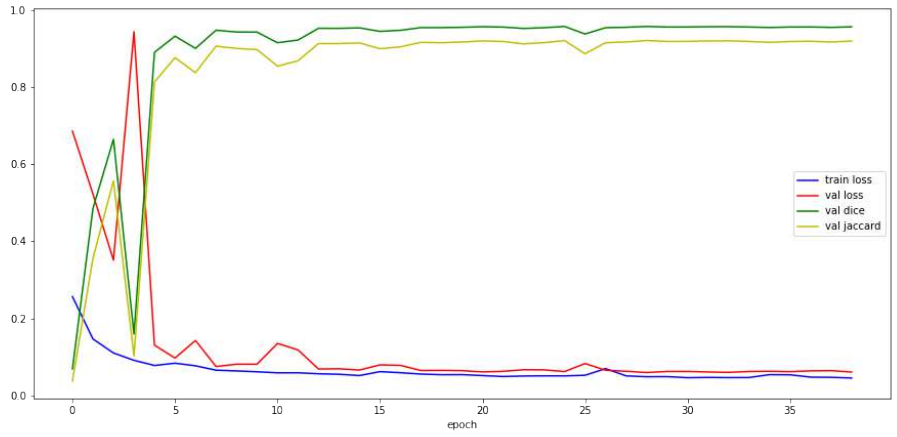


Figure A.3: Model C loss curves and validation metrics on the training process. TernaususNet VGG11, ±30, Adam optimizer, early stopping after 10% of maximum epochs do not yield performance increase.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

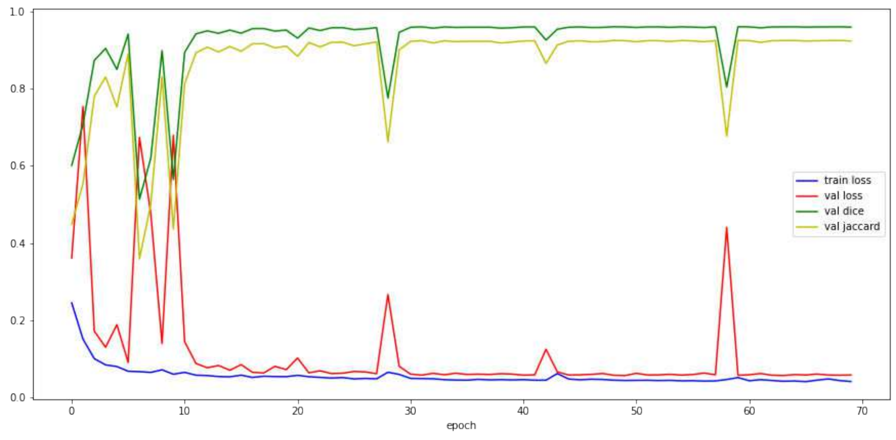


Figure A.4: Model D loss curves and validation metrics on the training process. TernaususNet VGG11, ±25, Adam optimizer, early stopping after 20% of maximum epochs do not yield performance increase.

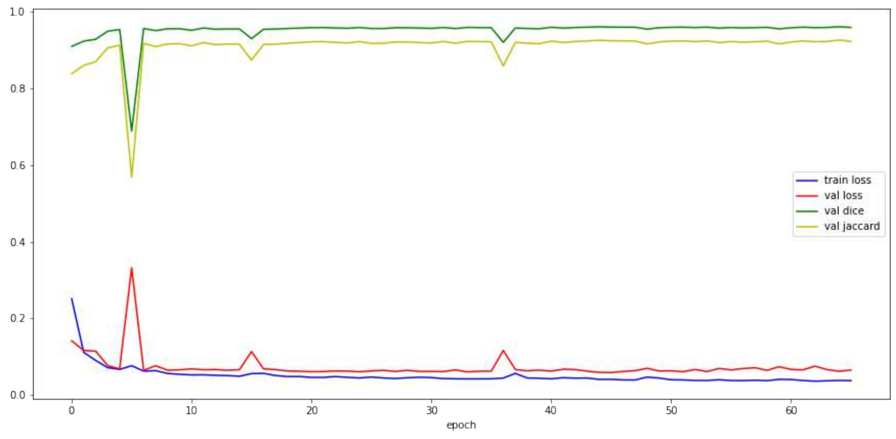


Figure A.5: Model D loss curves and validation metrics on the training process. U-Net blueprint, ±25, Adam optimizer, early stopping after 20% of maximum epochs do not yield performance increase.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

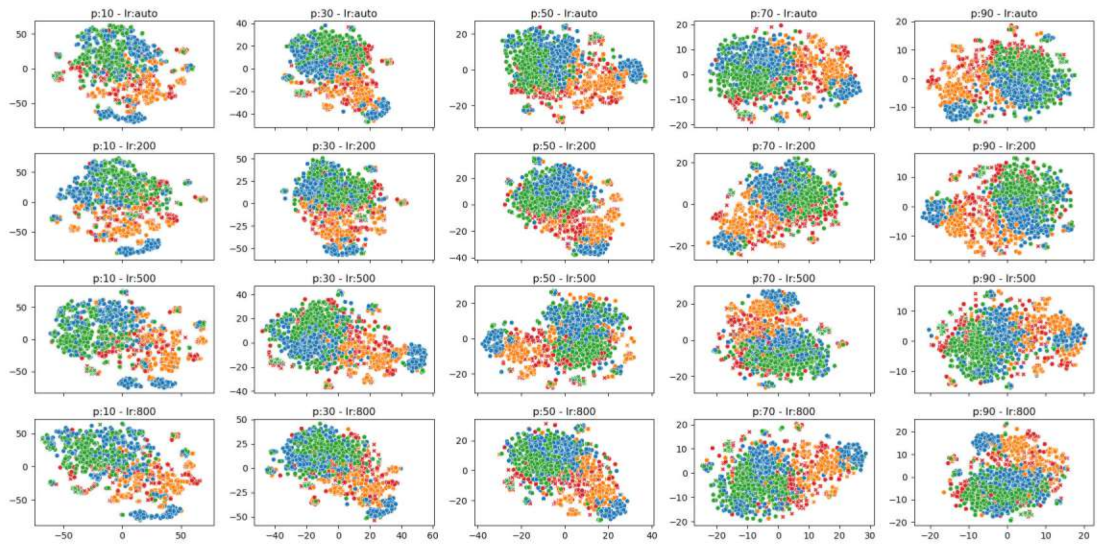


Figure A.6: Trying out different hyperparameter settings *perplexity* and *learning-rate* for t-SNE. Chosen values: $perplexity = 50; lr = 500$.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Figures

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Tables

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Bibliography

(2022). Imagenet benchmark. Online ; last accessed 20.01.2022.

Ahsan, M., Based, M., Haider, J., Kowalski, M., et al. (2021). Covid-19 detection from chest x-ray images using feature fusion and deep learning. *Sensors*, 21(4):1480.

Ake, C. F. (2005). Rounding after multiple imputation with non-binary categorical covariates. In *annual meeting of the SAS Users Group International, Philadelphia, PA*.

Akhloufi, M. A. and Chetoui, M. (2021). Chest XR COVID-19 detection. `https://cxr-covid19.grand-challenge.org/`. Online; accessed September 2021.

Akiyama, S., Hamdeh, S., Micic, D., and Sakuraba, A. (2021). Prevalence and clinical outcomes of covid-19 in patients with autoimmune diseases: a systematic review and meta-analysis. *Annals of the rheumatic diseases*, 80(3):384–391.

Anderson, E. J., Rouphael, N. G., Widge, A. T., Jackson, L. A., Roberts, P. C., Makhene, M., Chappell, J. D., Denison, M. R., Stevens, L. J., Pruijssers, A. J., et al. (2020). Safety and immunogenicity of sars-cov-2 mrna-1273 vaccine in older adults. *New England Journal of Medicine*, 383(25):2427–2438.

Andridge, R. R. and Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64.

Apostolopoulos, I. D., Aznaouridis, S. I., and Tzani, M. A. (2020). Extracting possibly representative covid-19 biomarkers from x-ray images with deep learning approach and image data related to pulmonary diseases. *Journal of Medical and Biological Engineering*, 40(3):462–469.

Arthur, D. and Vassilvitskii, S. (2006). k-means++: The advantages of careful seeding. Technical report, Stanford.

Barandela, R., Sánchez, J. S., Garcıa, V., and Rangel, E. (2003). Strategies for learning in class imbalance problems. *Pattern Recognition*, 36(3):849–851.

Becker, R. A. and Cleveland, W. S. (1987). Brushing scatterplots. *Technometrics*, 29(2):127–142.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Bengio, Y. (2015). Deep learning:theoretical motivations. Online; last accessed 23.03.2022.

Bengio, Y. and Delalleau, O. (2011). On the expressive power of deep architectures. In *International conference on algorithmic learning theory*, pages 18–36. Springer.

Bernold, G., Matkovic, K., Gröller, M. E., and Raidou, R. (2019). preha: Establishing precision rehabilitation with visual analytics. In *Eurographics Workshop on Visual Computing for Biology and Medicine (2019)*, pages 79–89.

Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., and Blaschko, M. B. (2019). Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 92–100. Springer.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Brehmer, M. and Munzner, T. (2013). A multi-level typology of abstract visualization tasks. *IEEE Trans. Visualization and Computer Graphics (TVCG) (Proc. InfoVis)*, 19(12):2376–2385.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brown, J. (2009). Choosing the right number of components or factors in pca and efa. *JALT Testing & Evaluation SIG Newsletter*, 13(2).

Cai, M., Bowe, B., Xie, Y., and Al-Aly, Z. (2021). Temporal trends of covid-19 mortality and hospitalisation rates: an observational cohort study from the us department of veterans affairs. *BMJ open*, 11(8):e047369.

Caliński, T. and JA, H. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3:1–27.

Card, M. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

Carrieri, V., Madio, L., and Principe, F. (2019). Vaccine hesitancy and (fake) news: Quasi-experimental evidence from italy. *Health economics*, 28(11):1377–1382.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1985). *Graphical methods for data analysis*. Wadsworth.

Chan, T. and Vese, L. (1999). An active contour model without edges. In Nielsen, M., Johansen, P., Olsen, O. F., and Weickert, J., editors, *Scale-Space Theories in Computer Vision*, pages 141–151, Berlin, Heidelberg. Springer Berlin Heidelberg.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Chen, B., Zhang, R., Gan, Y., Yang, L., and Li, W. (2017). Development and clinical application of radiomics in lung cancer. *Radiation Oncology*, 12(1):1–8.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American statistical Association*, 68(342):361–368.

Choo, J. and Liu, S. (2018). Visual analytics for explainable deep learning. *IEEE computer graphics and applications*, 38(4):84–92.

Conti, S., Ferrara, P., Fornari, C., Harari, S., Madotto, F., Silenzi, A., Zucchi, A., Manzoli, L., and Mantovani, L. G. (2020). Estimates of the initial impact of the covid-19 epidemic on overall mortality: evidence from italy. *ERJ Open Research*, 6(2).

Cook, K. A. and Thomas, J. J. (2005). Illuminating the path: The research and development agenda for visual analytics.

Correia, P. L. and Pereira, F. (2002). Stand-alone objective segmentation quality evaluation. *EURASIP Journal on Advances in Signal Processing*, 2002(4):1–12.

Dai, Z., Zeng, D., Cui, D., Wang, D., Feng, Y., Shi, Y., Zhao, L., Xu, J., Guo, W., Yang, Y., et al. (2020). Prediction of covid-19 patients at high risk of progression to severe disease. *Frontiers in Public Health*, 8.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Deiana, C., Geraci, A., Mazzarella, G., and Sabatini, F. (2022). Perceived risk and vaccine hesitancy: Quasi-experimental evidence from italy. *Health Economics*, 31(6):1266–1275.

Eavis, T. and Japkowicz, N. (2000). A recognition-based alternative to discrimination-based multi-layer perceptrons. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 280–292. Springer.

Fernandez, N. F., Gundersen, G. W., Rahman, A., Grimes, M. L., Rikova, K., Hornbeck, P., and Ma'ayan, A. (2017). Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Scientific data*, 4(1):1–12.

Floricel, C., Nipu, N., Biggs, M., Wentzel, A., Canahuate, G., Dijk, L. V., Mohamed, A., Fuller, C., and Marai, G. E. (2021). Thalis: Human-machine analysis of longitudinal symptoms in cancer therapy. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Fontanet, A., Tondeur, L., Madec, Y., Grant, R., Besombes, C., Jolly, N., Pellerin, S. F., Ungeheuer, M.-N., Cailleau, I., Kuhmel, L., Temmam, S., Huon, C., Chen, K.-Y., Crescenzo, B., Munier, S., Demeret, C., Grzelak, L., Staropoli, I., Bruel, T., Gallian, P., Cauchemez, S., van der Werf, S., Schwartz, O., Eloit, M., and Hoen, B. (2020). Cluster of covid-19 in northern france: A retrospective closed cohort study. *medRxiv*.

Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(4):453–476.

French, J., Deshpande, S., Evans, W., and Obregon, R. (2020). Key guidelines in developing a pre-emptive covid-19 vaccination uptake promotion strategy. *International Journal of Environmental Research and Public Health*, 17(16).

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Furmanová, K., Grossmann, N., Muren, L. P., Casares-Magaz, O., Moiseenko, V., Einck, J. P., Gröller, M. E., and Raidou, R. G. (2020). Vapor: Visual analytics for the exploration of pelvic organ variability in radiotherapy. *Computers & Graphics*, 91:25–38.

Furmanová, K., Muren, L. P., Casares-Magaz, O., Moiseenko, V., Einck, J. P., Pilskog, S., and Raidou, R. G. (2021). Previs: Predictive visual analytics of anatomical variability for radiotherapy decision support. *Computers & Graphics*, 97:126–138.

Gavin, B., Hayden, J., Adamis, D., and McNicholas, F. (2020). Caring for the psychological well-being of healthcare workers in the covid-19 pandemic crisis.

Gehrau, V., Fujarski, S., Lorenz, H., Schieb, C., and Blöbaum, B. (2021). The impact of health information exposure and source credibility on covid-19 vaccination intention in germany. *International Journal of Environmental Research and Public Health*, 18(9).

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Grus, J. (2019). *Data science from scratch: first principles with python*. O'Reilly Media.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Han, Y., Chen, C., Tewfik, A., Ding, Y., and Peng, Y. (2021). Pneumonia detection on chest x-ray using radiomic features and contrastive learning. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 247–251.

Harrison, L., Reinecke, K., and Chang, R. (2015). Infographic aesthetics: Designing for the first impression. In *Proceedings of the 33rd Annual ACM conference on human factors in computing systems*, pages 1187–1190.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Heaven, W. D. (2021). Hundreds of ai tools have been built to catch covid. none of them helped. Online; last accessed 03.05.2022.

Heinrich, J. and Weiskopf, D. (2013). State of the art of parallel coordinates. *Eurographics (State of the Art Reports)*, pages 95–116.

Hinton, G. E. and Roweis, S. (2002). Stochastic neighbor embedding. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Holzinger, A. (2018). From machine learning to explainable ai. In *2018 world symposium on digital intelligence for systems and machines (DISA)*, pages 55–66. IEEE.

Hua, J. and Shaw, R. (2020). Corona virus (covid-19)"infodemic" and emerging issues through a data lens: The case of china. *International journal of environmental research and public health*, 17(7):2309.

Huang, Q. and Dom, B. (1995). Quantitative methods of evaluating image segmentation. In *Proceedings., International Conference on Image Processing*, volume 3, pages 53–56 vol.3.

Iglovikov, V. and Shvets, A. (2018). Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *ArXiv e-prints*.

Inselberg, A. (1985). The plane with parallel coordinates. *The visual computer*, 1(2):69–91.

Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., and Möller, T. (2013). A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Islam, J. and Zhang, Y. (2018). Towards robust lung segmentation in chest radiographs with deep learning. *arXiv preprint arXiv:1811.12638*.

Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X. J., Lu, P.-X., and Thoma, G. (2014). Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., and Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, 50(2):105–115.

Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331.

Katal, A., Wazid, M., and Goudar, R. H. (2013). Big data: issues, challenges, tools and good practices. In *2013 Sixth international conference on contemporary computing (IC3)*, pages 404–409. IEEE.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, 8(1):1–8.

Keim, D. A., Mansmann, F., Stoffel, A., and Ziegler, H. (2008). Visual analytics.

Kermali, M., Khalsa, R. K., Pillai, K., Ismail, Z., and Harky, A. (2020). The role of biomarkers in diagnosis of covid-19–a systematic review. *Life sciences*, 254:117788.

Khan, I. U. and Aslam, N. (2020). A deep-learning-based framework for automated diagnosis of covid-19 using x-ray images. *Information*, 11(9):419.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Kohlberger, T., Singh, V. K., Alvino, C. V., Bahlmann, C., and Grady, L. J. (2012). Evaluating segmentation error without ground truth. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 15 Pt 1:528–36.

Kosara, R. and Skau, D. (2016). Judgment error in pie chart variations. page 91–95.

Lambin, P., Leijenaar, R. T., Deist, T. M., Peerlings, J., De Jong, E. E., Van Timmeren, J., Sanduleanu, S., Larue, R. T., Even, A. J., Jochems, A., et al. (2017). Radiomics: the bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*, 14(12):749–762.

Legido-Quigley, H., Mateos-García, J. T., Campos, V. R., Gea-Sánchez, M., Muntaner, C., and McKee, M. (2020). The resilience of the spanish health system against the covid-19 pandemic. *The lancet public health*, 5(5):e251–e252.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Lever, J. (2016). Classification evaluation: It is important to understand both what a classification metric expresses and what it hides. *Nature methods*, 13(8):603–605.

Li, F.-F., Krishna, R., and Xu, D. (2022). Cs231n convolutional neural networks for visual recognition. Online; last accessed 19.01.2022.

Li, X., Ge, P., Zhu, J., Li, H., Graham, J., Singer, A., Richman, P. S., and Duong, T. Q. (2020a). Deep learning prediction of likelihood of icu admission and mortality in covid-19 patients using clinical variables. *PeerJ*, 8:e10337.

Li, Y., Horowitz, M. A., Liu, J., Chew, A., Lan, H., Liu, Q., Sha, D., and Yang, C. (2020b). Individual-level fatality prediction of covid-19 patients using ai methods. *Frontiers in Public Health*, 8:566.

Little, R. J. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*, volume 793. John Wiley & Sons.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Magg, C., Toussaint, L., Muren, L. P., Indelicato, D. J., and Raidou, R. G. (2021). Visual Assessment of Growth Prediction in Brain Structures after Pediatric Radiotherapy. In Oeltze-Jafra, S., Smit, N. N., Sommer, B., Nieselt, K., and Schultz, T., editors, *Eurographics Workshop on Visual Computing for Biology and Medicine*. The Eurographics Association.

Mahase, E. (2020). Covid-19: death rate is 0.66% and increases with age, study estimates. *BMJ: British Medical Journal (Online)*, 369.

Malik, P., Patel, U., Mehta, D., Patel, N., Kelkar, R., Akrmah, M., Gabrilove, J. L., and Sacks, H. (2021). Biomarkers and outcomes of covid-19 hospitalisations: systematic review and meta-analysis. *BMJ evidence-based medicine*, 26(3):107–108.

Mayerhoefer, M. E., Materka, A., Langs, G., Häggström, I., Szczypiński, P., Gibbs, P., and Cook, G. (2020). Introduction to radiomics. *Journal of Nuclear Medicine*, 61(4):488–495.

Metsalu, T. and Vilo, J. (2015). Clustvis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic acids research*, 43(W1):W566–W570.

Miksch, S. and Aigner, W. (2014). A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data. *Computers & Graphics*, 38:286–290.

Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition.

Munzner, T. (2009). A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6):921–928.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Munzner, T. (2014). *Visualization analysis and design*. CRC press.

Márquez-Neila, P., Baumela, L., and Alvarez, L. (2014). A morphological approach to curvature-based evolution of curves and surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):2–17.

Nielsen, J. (1994). *Usability engineering*. Morgan Kaufmann.

Norval, M., Wang, Z., and Sun, Y. (2019). Pulmonary tuberculosis detection using deep learning convolutional neural networks. In *Proceedings of the 3rd International Conference on Video and Image Processing*, ICVIP 2019, page 47–51, New York, NY, USA. Association for Computing Machinery.

Nusrat, S., Harbig, T., and Gehlenborg, N. (2019). Tasks, techniques, and tools for genomic data visualization. In *Computer Graphics Forum*, volume 38, pages 781–805. Wiley Online Library.

Ovcharenko, I. (2019). Lung-segmentation. Online; last accessed 23.04.2022.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pearson, K. (1895). X. contributions to the mathematical theory of evolution.—ii. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London.(A.)*, (186):343–414.

Pham, D. L., Xu, C., and Prince, J. L. (2000). Current methods in medical image segmentation. *Annual review of biomedical engineering*, 2(1):315–337.

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Marc, G. P., Moreira, E. D., Zerbini, C., et al. (2020). Safety and efficacy of the bnt162b2 mrna covid-19 vaccine. *New England journal of medicine*.

Ponti, G., Maccaferri, M., Ruini, C., Tomasi, A., and Ozben, T. (2020). Biomarkers associated with covid-19 disease progression. *Critical reviews in clinical laboratory sciences*, 57(6):389–399.

Prince, S. J. (2012). *Computer vision: models, learning, and inference*. Cambridge University Press.

Provost, F., Fawcett, T., and Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms 1998. In *Proceedings of the 15th international conference on machine learning ICML-98 Morgan Kaufmann. San Mateo, CA*.

TU WIEN Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Pu, J., Leader, J., Bandos, A., Shi, J., Du, P., Yu, J., Yang, B., Ke, S., Guo, Y., Field, J. B., et al. (2020). Any unique image biomarkers associated with covid-19? *European radiology*, 30(11):6221–6227.

Puri, N., Coomes, E. A., Haghbayan, H., and Gunaratne, K. (2020). Social media and vaccine hesitancy: new updates for the era of covid-19 and globalized infectious diseases. *Human vaccines & immunotherapeutics*, 16(11):2586–2593.

Radiomics (2022). Pyradiomics documentation. Online; last accessed 23.03.2022.

Reitermanova, Z. et al. (2010). Data splitting. In *WDS*, volume 10, pages 31–36.

Reno, C., Maietti, E., Fantini, M. P., Savoia, E., Manzoli, L., Montalti, M., and Gori, D. (2021). Enhancing covid-19 vaccines acceptance: results from a survey on vaccine hesitancy in northern italy. *Vaccines*, 9(4):378.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217.

Rokach, L. and Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Ruiz-Fernández, M. D., Ramos-Pichardo, J. D., Ibáñez-Masero, O., Cabrera-Troya, J., Carmona-Rega, M. I., and Ortega-Galán, Á. M. (2020). Compassion fatigue, burnout, compassion satisfaction and perceived stress in healthcare professionals during the covid-19 health crisis in spain. *Journal of clinical nursing*, 29(21-22):4321–4330.

Saary, M. J. (2008). Radar plots: a useful way for presenting multivariate health care data. *Journal of clinical epidemiology*, 61(4):311–317.

Sacha, D., Sedlmair, M., Zhang, L., Lee, J. A., Peltonen, J., Weiskopf, D., North, S. C., and Keim, D. A. (2017). What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing*, 268:164–175. Advances in artificial neural networks, machine learning and computational intelligence.

Saltz, J., Saltz, M., Prasanna, P., Moffitt, R., Hajagos, J., Bremer, E., Balsamo, J., and Kurc, T. (2021). Stony brook university covid-19 positive cases (covid-19-ny-sbu). The Cancer Imaging Archive. Online; Version 2021/08/11; accessed December 2021.

TU WIEN Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Sarikaya, A. and Gleicher, M. (2017). Scatterplots: Tasks, data, and designs. *IEEE transactions on visualization and computer graphics*, 24(1):402–412.

Schmid, P., Rauber, D., Betsch, C., Lidolt, G., and Denker, M.-L. (2017). Barriers of influenza vaccination intention and behavior–a systematic review of influenza vaccine hesitancy, 2005–2016. *PloS one*, 12(1):e0170550.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X. (2017). Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Sedlmair, M., Meyer, M., and Munzner, T. (2012). Design study methodology: Reflections from the trenches and the stacks. *IEEE transactions on visualization and computer graphics*, 18(12):2431–2440.

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.-i., Matsui, M., Fujita, H., Kodera, Y., and Doi, K. (2000). Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74.

Shlens, J. (2014). A tutorial on principal component analysis.

Shneiderman, B. (2003). The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*, pages 364–371. Elsevier.

Siirtola, H. (2019). The cost of pie charts. In *2019 23rd International Conference Information Visualisation (IV)*, pages 151–156.

Skau, D. and Kosara, R. (2016). Arcs, angles, or areas: Individual data encodings in pie and donut charts. 35(3):121–130.

Smiciklas, M. (2012). *The power of infographics: Using pictures to communicate and connect with your audiences.* Que Publishing.

Souza, J. C., Bandeira Diniz, J. O., Ferreira, J. L., França da Silva, G. L., Corrêa Silva, A., and de Paiva, A. C. (2019). An automatic method for lung segmentation and reconstruction in chest x-ray using deep neural networks. *Computer Methods and Programs in Biomedicine*, 177:285–296.

Spence, I. and Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology*, 5(1):61–77.

Steinbach, M., Ertöz, L., and Kumar, V. (2004). The challenges of clustering high dimensional data. In *New directions in statistical physics*, pages 273–309. Springer.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Strimbu, K. and Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463.

Tamal, M., Alshammari, M., Alabdullah, M., Hourani, R., Alola, H. A., and Hegazi, T. M. (2021). An integrated framework with machine learning and radiomics for accurate and rapid early diagnosis of covid-19 from chest x-ray. *Expert systems with applications*, 180:115152.

Trutschl, M., Grinstein, G., and Cvek, U. (2003). Intelligently resolving point occlusion. In *IEEE Symposium on Information Visualization 2003 (IEEE Cat. No. 03TH8714)*, pages 131–136. IEEE.

Tufte, E. R. (1985). The visual display of quantitative information. *The Journal for Healthcare Quality (JHQ)*, 7(3):15.

Udeshani, K., Meegama, R., and Fernando, T. (2011). Statistical feature-based neural network approach for the detection of lung cancer in chest x-ray images. *International Journal of Image Processing (IJIP)*, 5(4):425–434.

Ullman, S. (2000). *High-level vision: Object recognition and visual cognition*. MIT press.

User:Schutz (2007). Pie vs bar chart. Online; last accessed 28.06.2022.

Valindria, V. V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E. O., Rockall, A. G., Rueckert, D., and Glocker, B. (2017). Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth. *IEEE Transactions on Medical Imaging*, 36(8):1597–1606.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Van der Heijden, G. J., Donders, A. R. T., Stijnen, T., and Moons, K. G. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *Journal of clinical epidemiology*, 59(10):1102–1109.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., and Yu, T. (2014). scikit-image: image processing in python. *PeerJ*, 2:e453.

van Griethuysen, J. (2020a). Github issue - [feat extraction] wonder how features are caculated for mask with multi rois. Online; last accessed 27.03.2022.

van Griethuysen, J. (2020b). Github issue - how to extract features from multi rois in one scan slice? Online; last accessed 27.03.2022.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G., Fillion-Robin, J.-C., Pieper, S., and Aerts, H. J. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104–e107.

Varghese, B. A., Shin, H., Desai, B., Gholamrezanezhad, A., Lei, X., Perkins, M., Oberai, A., Nanda, N., Cen, S., and Duddalwar, V. (2021). Predicting clinical outcomes in covid-19 using radiomics on chest radiographs. *The British Journal of Radiology*, 94(1126):20210221.

Veneti, L., Bøås, H., Kristoffersen, A. B., Stålcrantz, J., Bragstad, K., Hungnes, O., Storm, M. L., Aasand, N., Rø, G., Starrfelt, J., et al. (2022). Reduced risk of hospitalisation among reported covid-19 cases infected with the sars-cov-2 omicron ba. 1 variant compared with the delta variant, norway, december 2021 to january 2022. *Eurosurveillance*, 27(4):2200077.

Vidal, P. L., de Moura, J., Novo, J., and Ortega, M. (2021). Multi-stage transfer learning for lung segmentation using portable x-ray devices for patients with covid-19. *Expert Systems with Applications*, 173:114677.

Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to use t-sne effectively. *Distill*, 1(10):e2.

Yenduri, S. and Iyengar, S. S. (2007). Performance evaluation of imputation methods for incomplete datasets. *International Journal of Software Engineering and Knowledge Engineering*, 17(01):127–152.

Yim, O. and Ramdeen, K. T. (2015). Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *The quantitative methods for psychology*, 11(1):8–21.

Zebin, T. and Rezvy, S. (2021). Covid-19 detection and disease progression visualization: Deep learning on chest x-rays for classification and coarse localization. *Applied Intelligence*, 51(2):1010–1021.

Zhao, Z., Chen, A., Hou, W., Graham, J. M., Li, H., Richman, P. S., Thode, H. C., Singer, A. J., and Duong, T. Q. (2020). Prediction model and risk scores of icu admission and mortality in covid-19. *PloS one*, 15(7):e0236618.

Zhu, J., Shen, B., Abbasi, A., Hoshmand-Kochi, M., Li, H., and Duong, T. Q. (2020). Deep transfer learning artificial intelligence accurately stages covid-19 lung disease severity on portable chest radiographs. *PloS one*, 15(7):e0236621.

Zwanenburg, A., Leger, S., Vallières, M., and Löck, S. (2016). Image biomarker standardisation initiative. *arXiv preprint arXiv:1612.07003*.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J. W. L., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R. J., Boellaard, R., Bogowicz, M., Boldrini, L., Buvat, I., Cook, G. J. R., Davatzikos, C., Depeursinge, A., Desseroit, M.-C., Dinapoli, N., Dinh, C. V., Echegaray, S., Naqa, I. E., Fedorov, A. Y., Gatta, R., Gillies, R. J., Goh, V., Götz, M., Guckenberger, M., Ha, S. M., Hatt, M., Isensee, F., Lambin, P., Leger, S., Leijenaar, R. T., Lenkowicz, J., Lippert, F., Losnegård, A., Maier-Hein, K. H., Morin, O., Müller, H., Napel, S., Nioche, C., Orlhac, F., Pati, S., Pfaehler, E. A., Rahmim, A., Rao, A. U., Scherer, J., Siddique, M. M., Sijtsema, N. M., Fernandez, J. S., Spezi, E., Steenbakkers, R. J., Tanadini-Lang, S., Thorwarth, D., Troost, E. G., Upadhaya, T., Valentini, V., van Dijk, L. V., van Griethuysen, J., van Velden, F. H., Whybra, P., Richter, C., and Löck, S. (2020). The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.