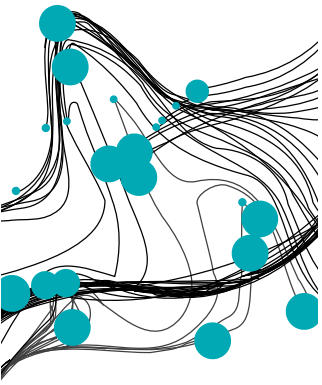
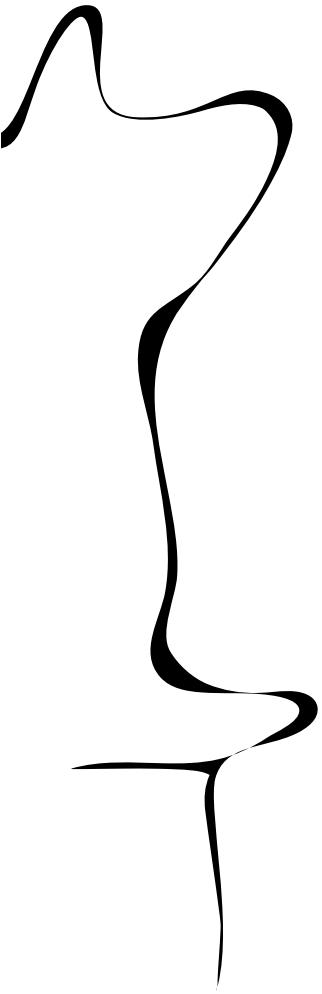


.82481

DMB

DATA MANAGEMENT AND BIOMETRICS



MULTIMODAL 100 DAY POSTOPERATIVE MORTALITY PREDICTION FOR GERIATRIC PATIENTS WITH A HIP FRACTURE: WORKING TOWARDS CLINICAL IMPLEMENTATION

Niels van der Heijden

MASTER'S ASSIGNMENT

Committee:
Prof. Dr. Ir. Maurice van Keulen
Jorn-Jan van de Beld MSc.
Prof. dr. Han Hegeman

August, 2024

2024DMB0007
Data Management and Biometrics
EEMathCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

ABSTRACT

Background
Hip fractures are a highly prevalent type of fracture among elderly patients with a high chance of mortality (13% within a 100 days post surgery), high healthcare costs and a heavy rehabilitation process. The ability to forecast mortality can aid making a more informed treatment plan, such as palliative treatment for highly frail patients.

Objective
This thesis builds upon an existing multimodal model developed by ZGT (Hospitalgroup Twente) to predict 30-day and 100-day post surgery mortality for elderly patients (age >70). The goal is to utilize ZGT's newly available data and extract features from non operatively treated patients (patients who received palliative care) to increase performance. Additionally, the requirements for clinical implementation are directly incorporated into the design process.

Methods
A baseline model is compared with a new iteration that includes the new data and additional features. The baseline model is a recreation of a known multimodal method that combines static patient data with X-ray image modalities. First, comorbidity features, extracted with an NLP algorithm, are added to the baseline. Second, patient similarity features are added, which approximates the similarity between a "to-be treated patient" and a cohort of patients who received palliative care. The performance of the different models is captured with AUCROC. However, it is later proposed to utilize AUCPR, precision and recall, given better clinical applicability.

Results
A Key finding is that changing the target outcome to 100-day mortality significantly improves all performance metrics. The inclusion of comorbidity data shows mixed results; it does not improve the original MM-Y model, but improves performance when image modalities are excluded. Notably, adding comorbidities to a model with static data increases the AUCPR by 2 percentage points. Furthermore, the integration of similarity scores from NOM patients demonstrates a substantial improvement in Maximum precision, with a 20-percentage-point increase, an enhancement of AUCPR by 6 percentage points. The AUCROC remains the same at 0.77.

Conclusion
This thesis has shown that NLP extracted comorbidities can improve predictive performance. Furthermore, it can be concluded that the inclusion of NOM patient similarity scores greatly increases predictive performance in the given context. Additionally, the 100-day mortality target value is proven to be easier to predict and has the same clinical relevance. Finally, it is recommended to utilize AUCPR, precision and recall for future research on postoperative mortality for geriatric patients with a hip fracture.

CLINICAL ABSTRACT

Background
Hip fractures are a highly prevalent type of fracture among elderly patients with a high chance of mortality (13% within a 100 days post surgery), high costs and a impactful treatment process. The ability to forecast mortality can aid making a more informed treatment plan, such as palliative treatment for highly frail patients. In the context of ZGT, this is especially useful for traumasurgeons. Therefore, the included types of hipfractures are intracapsular and extracapsular fractures, with the exclusion of pathological and periprosthetic fractures.

Objective
Previous research within ZGT (Hospital Group Twente) has attempted to predict 30-day post surgery mortality for geriatric patients with a hip fracture (age >70). To achieve this, the previous method combined static patient information, Thorax and Hip radiology images. In this thesis, this is extended by using ZGT's newly available EHR's (Electronic Health Records) and extracting information from NOM patients (patients who received palliative care) to increase performance. Additionally, the requirements for clinical implementation are directly incorporated into the design process.

Methods
A baseline model is compared with a new iteration that includes the new data. The baseline model is a recreation of a known method, developed by ZGT. First, comorbidity features extracted from ER briefs are added to the baseline. Second, patient similarity features are added, who approximate the similarity between a "to-be treated patient" and a cohort of patients that received palliative care. To evaluate the experiments, it is determined how confident each model is in predicting mortality cases, which is highly important for clinical implementation.

Results
Key findings include the observation that changing the target outcome to 100-day mortality significantly improves all performance metrics. In contrast, adding the ASA-score to this method does not increase performance. The inclusion of comorbidities improves a simplified version of the original method, where Thorax and Hip X-ray images are excluded (3 percentage point AUCPR improvement). Furthermore, the integration of similarity scores from NOM patients demonstrates a substantial improvement in Maximum precision, with a 20-percentage-point increase, an enhancement of AUCPR by 6 percentage points.

Conclusion
This thesis has showcased that comorbidities extracted from ER briefs can improve predictive performance. Furthermore, it can be concluded that the inclusion of NOM patient similarity scores greatly increases predictive performance in the given context. Additionally, the 100-day mortality target value is proven to be easier to predict and has the same clinical relevance. Finally, in the given context the ASA-score, Thorax and Hip X-ray images do not improve predictive performance.

CONTENTS

Acronyms	Description
AHFS	Almelo Hip Fracture Score
ASA	American Society of Anesthesiologists
ATC	Anatomical Therapeutic Chemical
AUC	Area Under Curve
CCI	Charlson Comorbidity Index
CNN	Convolutional Neural Network
CvGT	Centrum van Geriatrische Traumatologie / Center for Geriatric Traumatology
EHR	Electronic Health Record
ER	Emergency Room
KNN	K-nearest Neighbors
MM-Y	Multimodal architecture proposed in (Yenidogan, 2021)
NHFS	Nothingham Hip Fracture Score
NLP	Natural Language Processing
NOM	Non Operative Management (Palliative treatment)
PCA	Principal Component Analysis
RF	Random Forest Algorithm
UT	University of Twente
ZGT	Ziekenhuisgroep Twente / Hospital Group Twente

1 INTRODUCTION

One of the major challenges for health care systems is the increasing age of the human population. In many countries, life expectancy has increased and birthrate has declined, resulting in a higher percentage of the population being 60 years or older. In Europe and the United States of America this percentage has reached up to 21%, with Japan reaching percentages up to 31% [1]. This increasing population of elderly has a significant impact on the healthcare systems of such countries. Specifically, for the Hospital Group Twente (ZGT), the hospital in question, the average age of an elderly (geriatric) patient with a hip fracture is approximately 82 years. A geriatric patient is someone 65 years or older, who is being treated for a certain condition. Due to the increased risk and prevalence of medical complications for the elderly, the increase in this part of the population puts additional stress on related healthcare (geriatrics). Additionally, the Dutch government has recognized that the current demand for healthcare professionals is growing faster than the supply [2], adding additional stress to the system.

Hip fractures are one of the most common problems among older adults. Currently, around 14% of total fractures among older adults are hip fractures, with 300.000 occurrences yearly in the USA alone. This 14% of total fractures represents a disproportionate 72% of fracture-related healthcare costs [3]. The United States reported a 6 billion dollar annual bill, the United Kingdom a 2 Billion Pound Bill and the Swedish government estimated a 147 Million euro Bill [4, 5]. Despite prevention efforts, the number of hip fractures is expected to increase as the population ages [6, 7]. Next to the costs and physical trauma, the thirty-day and one-year post-hip-surgery mortality rate lay between 5%-13.3% and 25%-31% respectively [8, 3, 9, 5, 10, 11, 12, 13]. This is in line with the approximate thirty-day mortality rate of 8%, reported by ZGT. The number of hip fracture cases is estimated to increase to 4.5 - 6 Million worldwide annually, therefore increasing costs and absolute mortality [8, 3, 10, 13]. With increasing shortages of healthcare personnel, alternative approaches should be explored.

1.1 Problem Background: General

To keep up with the rising demand an improvement in the efficiency of treatment for geriatric patients with a hip fracture is required. Next to medical research, alternative approaches can be taken to achieve this goal. In recent years, mortality prediction has gained a lot of attention among researchers [11, 9, 12, 3, 4, 13]. Mortality risk predictions can help select a more appropriate and hence effective treatment plans. For example, Non Operative Management (NOM), also know as palliative care can sometimes be preferred over surgery. NOM means that a patient who undergoes palliative treatment is not expected to be cured. Instead, the goal is to improve upon the current situation and maintain a good quality of life, without attempting to achieve a good functional outcome by means of surgery. With this approach, unnecessary harm is reduced while reducing costs for the healthcare system. This form of treatment, with respect to quality of life, has also gained interest in the research community. Recent studies have show that a NOM approach can increase quality of life [14]. Therefore, it is highly valuable to estimate, prior to surgery, whether a patient is likely to develop complications or decease.

To make such predictions, healthcare providers have been exploring Data-driven solutions in an attempt to solve the aforementioned problem. Although the usefulness of Machine Learning (ML) and Artificial Intelligence (AI) algorithms has been proven in medical case studies [5], it remains a challenging task in many regards. Based on related literature and conversation with clinical experts within ZGT the following challenges/requirements are defined for such ML and AI solutions.

Performance

The performance of such algorithms is ought to be comparable or better than clinical experts. Furthermore, such algorithms aid in highly sensitive decision-making processes. This implies that the confidence in classifying a patient as ”likely to decease within 100 days” must be very high. Therefore confidence in a mortality prediction is more important than confidence in a non-mortality prediction. State-of-the-art models have not reached the point where the confidence in a mortality prediction is sufficiently high, thus warranting the need for further research.

Explainability
In cases where an algorithm is confident in its prediction, it is highly important that the decision is justified by empirical proof. Not all algorithms are currently able to deliver such proof. Especially so called black box algorithms are difficult to implement in clinical settings, given that it is nearly impossible to explain how the outcome was established. This, in turn, makes it difficult to be transparent with the patient and others involved.

Practical Usage
The algorithm must be practical when deployed, which means that it must meet the following requirements. Missing values are prevalent in clinical data, hence an algorithm should be able to deal with these. An algorithm must be time-efficient and produce an output in a timely manner. Furthermore, to increase the impact of an algorithm, it should be deployed in multiple healthcare institutions. However, not every institution has the same data availability. Therefore, it would be preferred if the algorithm can handle correlated alternatives to unavailable data.

1.1.1 Research Gaps

A high heterogeneity in research on postoperative mortality for geriatric patients with hip fracture, may be observed. Many different Machine Learning methods have been proposed to increase the predictive performance of mortality. Furthermore, there exist large differences between research population sizes, available features, and time of model inference (pre-, peri-, or post-surgery). However, two gaps have been identified in the literature. First, a model that included palliative treated patients has not been proposed. Normally, only patients that receive treatment in the form of surgery are included. Second, the evaluation of the proposed models is almost exclusively done with the use of the AUCROC. Due to the nature of this dataset, other metrics can be more informative and hence are included in this thesis. With these metrics, new insights are given into the performance of the model.

1.2 Problem Background: ZGT

Naturally, the aforementioned general problems hold for most hospitals, including ZGT. However, some additional complexities should be added to the problem definition. Within ZGT, older adults with hip fractures are treated by traumasurgeons at the Center of Geriatric Traumatology (CvGT), which has been proven to reduce the 1-year mortality rate among ZGT patients [15]. Therefore, the solution (medical decision support algorithm) must be implemented within this department. First, the patients treatment within this department should only be taken into account, imposing selection criteria to the problem. Second, the placement of the algorithm must be taken into account. Depending on when inference is required, different data will be available. Section 2.3 will give a further explanation on the CvGT treatment process and the exact moment of implementation.

Previous Research

In previous years, ZGT internally approached the mortality prediction problem of geriatric patients with hip fracture. This resulted in different approaches, each having its own strengths and weaknesses.

First, a variation of the Nothing Hip Fracture Score (NHFS) [16] was developed, called the Almelo Hip Fracture Score (AHFS) [11]. These results were very promising, resulting in a highly transparent and well-performing model. However, the issue faced was the low confidence in high risk classifications (68% maximum confidence). According to field experts, this would ideally be above 90%.
Second, a new multimodal method was developed. combining hip and thorax X-ray images and static

patient data (Age, Vital signs, Nutrition scores, etc.) [17]. Similarly to AHFS, promising results were achieved. However, due to the limited availability of comorbidity data, questions about the full potential remained. Due to the proven relevance of such features, another project was carried out, attempting to automatically extract such information from patient reports. By utilizing NLP methods, such comorbidities were extracted from text, hence making this data available for future machine learning problems. Whether this increases the performance of post-surgery mortality prediction (for geriatric patients with a hip fracture) has not yet been confirmed.

Improved Data Quality

An increase in ZGT's data quality has taken place over the years, due to more experience. Less data missing and a wider variety of datapoints have likely increased the power of the data. Whether this increase in power improves the performance of post surgery mortality prediction (for geriatric patients with a hip fracture), has not been confirmed yet.

Defining Mortality Prediction

Furthermore, the definition of mortality prediction depends on the context of the problem. ZGT aims to deploy the model as a decision-making support system. This decision is based on the likelihood of mortality. But how is the likelihood of mortality defined? In related work, often 30 days or 1 year mortality is chosen as target value. Because 1-year post-surgery mortality is too long to consider NOM, this target value is excluded from this research. However, in consultation with medical professionals, it became clear that 100-day mortality is similarly useful for the decision-making of the treatment plan. Therefore, in this study two target variables are considered, 30-day and 100-day mortality.

1.3 Problem Statement

The previous research within ZGT has resulted in promising results. Nonetheless, the general problem statement remains unchanged. A clinically implementable model with sufficient predictive performance has yet to be developed. However, newly available datapoints have sparked a new sub-problem. The effect of the new datapoints is not yet know. Such developments may contribute to an ML model's confidence in predicting mortality. In turn this contributes towards creating a clinically implementable model for the prediction of postoperative mortality in geriatric patients with hip fracture.

In conclusion, the main problem revolves around the model performance. Although it is an important first step, it should be kept in mind that this is only the first step towards clinical implementation.

1.4 Research Question

The previous problem statement can be subdivided into two different problems and subsequent research questions.

Question 1

Main:
Does the current state of the ZGT Dataset and NLP extracted comorbidity data improve the performance of the multimodal model developed in [13].

Subquestion:

- *To what extent do the new features improve the performance of the multimodal architecture [13] and how does this performance compare to the AHFS [11]?*

The first research questions attempts to increase the performance of ZGT's internal post-surgery mortality prediction model. Next to this, this paper tests a more general hypothesis that can aid towards

performance increase. Therefore, the second research question is based on an iterative approach, where gaps in the literature were explored and exploited.

Question 2

Main:
Related literature has pointed out numerous future research directions that require additional data. However, gathering additional data is a costly process. Given the readily available data, can a gap in the literature be defined and exploited to improve post-surgery mortality prediction in the given context?

Subquestion 1:

- *What gaps in the literature may be identified, and which of those may be exploited without changing the current dataset?*

Subquestion 2:

- *Non-treated patients are often excluded in related work. To what extent does extracting features, such as a similarity score, from patients who received palliative treatment, improve the predictive performance?*

1.5 Research Population

The research population consists of patients who were treated within ZGT from 2015 to the beginning of 2024. Each record is related to a patient that has been treated according to the CvGT and with the following characteristics. Patients must be 70 years or older (geriatric), must have suffered a hip fracture of any kind, and must be treated by a geriatric traumasurgeon within ZGT (treatment does not exclusively imply surgery). This resulted in a population of N = 2.082
Additionally, 89 patients who received NOM were included. This sub-cohort served as a reference group for feature extraction. Furthermore, patients with periprosthetic fractures (fractures around a previously placed prothesis) and pathological fractures (fractures caused by a disease, not by impact) are excluded. A more detailed description of the used data will be given in Section 7.

2 BACKGROUND

In this chapter, the background required to understand the given problem is described. This entails preliminaries such as the understanding of processes active within ZGT, a detailed explanation of utilized methods, and terminology.

2.1 Context

The following research is conducted in collaboration with the Hospital Group Twente (ZGT) and the University of Twente (UT). ZGT is a cluster of hospitals, clinics and specialists, that provides healthcare for the Dutch population. Located in Overijssel, The Netherlands, ZGT has a service area with approximately 390.000 inhabitants. Next to providing all standard healthcare, the hospital group specializes in geriatrics, oncology and complex diabetes and obesity. Throughout this investigation, specifically the Department of Geriatric Traumatology (CvGT) was involved.

Internally, ZGT has been expanding its Data Science Lab over the past decade, which uses big data and machine learning techniques, to improve healthcare services. This research will be carried out within the data science team, which has ZGT healthcare experts and researchers of the University of Twente at its disposal. The connected institutes allow for a wider collection of data and knowledge of expertise, which would otherwise not be available.
The research in this article will be carried out for academic purposes, while simultaneously attempting to improve ZGT's provided healthcare.

2.2 Hip Fractures

Hip fractures are one of the most frequent fractures presenting to the emergency department and orthopedic trauma teams. It implies a fracture of the proximal femur between the femoral head and 5 cm distal to the lesser trochanter [18]. There are two main types of hip fracture caused by trauma, the intracapsular fractures and extracapsular fractures. First, intracapsular neck of femur fractures occur within the capsule of the hip joint. The blood supply to the femoral head travels in a retrograde direction via the capsule. As such, any fracture within the capsule could be likely to damage this blood supply. Second, extra capsular neck of femur fractures are fractures of the neck of the femur which occur outside the capsule of the hip joint. As such, the risks of avascular necrosis of the femoral head are no longer a concern. Most hip fractures among geriatric patients are the result of trauma. These are treated within CvGT and fall within the scope of this thesis.
A small percentage however $\tilde{5}$% [18] is not caused by a trauma, such fractures are called pathological fractures. Pathological fractures are caused by a disease process, often due to malignancy and the use of bisphosphonates. Such cases are not treated within the CvGT and therefore do not fall within the scope of this thesis. Additionally, an exception is made for patients with periprosthetic fracture. A periprosthetic fracture, is a fracture around a previously placed prosthesis. Within ZGT, these cases are handled by a orthopedist, which makes it fall outside of the scope of this thesis.

2.3 CvGT Treatment Process

Many important points in time exist between the moment the hip fracture occurs and the moment of surgery. To this end, it is especially relevant to know what processes occur from entering the Emergency Room (ER) up to the operating room. Note that the CvGT treatment process holds for the majority of the

patients, however exceptions can occur.

The process starts in the ER. Here, numerous types of information are gathered. These include the patient's experience during the incident, the patient's history, the medication verification, the physical investigation, results from lab tests, images taken by the department of radiology, and the patient's vitals. With respect to Data Science, this information can mainly be seen as structured data. Ultimately, this information, combined with additional manually written notes (such as Q&A's and reported underlying diseases), results in an ER report. Such a report may be seen as unstructured textual data.
The next step in the process is the hospitalization of the patients. Here, it is determined whether the patient will undergo surgery. Here, additional information is collected, such as the mental state of the patient and possibly more comorbidities. Once it is confirmed that the patient will undergo surgery, the patient will be subject to screening. Here, extra checks prior to surgery are conducted. An important piece of information is gathered, the ASA score. The American Society of Anesthesiologists (ASA) score describes the general health status of a patient [19]. This variable is often used in mortality prediction. After the screening, the patient will undergo surgery, after which the rehabilitation procedure will begin. The moment of algorithmic inference will take place between the hospitalization and the preoperative screening, see Figure 2.1. This is where all information is gathered and synthesized into a well considered treatment plan. The proposed algorithm would serve as one of those sources of information, leading to a more substantiated outcome. Hence, here the algorithm will produce its prediction to aid in the decision to perform surgery.

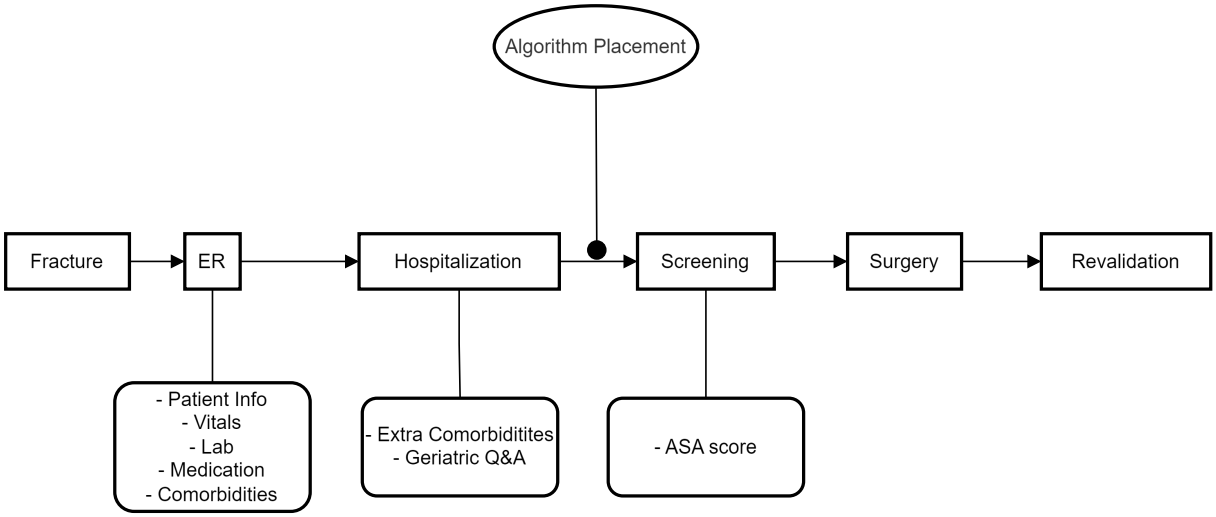


Figure 2.1: The simplified CvGT treatment process, with the point where the algorithm will be utilized.

3 RESEARCH APPROACH

A combination of multiple research frameworks will be used, given the complexity and uncertainty of the process. The process and thesis will be split into separate iterations. Because it is not known how many iterations will occur, the framework was designed with this in mind.

3.1 CRISP-DM

CRISP-DM is the overarching framework, that will lay the foundation of the research approach. It is a widely used framework in Data mining projects [20]. It ensures a structured way of tackling Machine Learning problems and presenting a usable end-product. It does so, by dividing the Data Mining challenge into six different sub-phases. The order of the phases is not strict; however, the most common dependencies and sequences have been illustrated in Figure 3.1. The output of each phase determines the next phase and the specific task that should be carried out. For instance, if the output of the data understanding phase is not satisfactory, the next step may be to increase business understanding in more detail. If the level of data understanding has been sufficiently satisfied, the next phase may be data preparation instead. Furthermore, the cyclic nature of CRISP-DM is illustrated with circular outer arrows. This indicates that a Data Mining project is cyclic and will always raise new questions. Thus, the framework offers a structured way of tackling smaller sub-problems that relate to one large data mining task.

The framework is described in Figure 3.1 and is made up of the components listed below. This will also give a clear overview of how the first part of this thesis is structured.

1. **Business/Problem Understanding**
   Here, the objectives and requirements of the project are stated. The issues facing the company or organization in question are fully understood, to form a clear direction of the project. In case of this report, this requires a clear understanding of ZGT, the Department of geriatric traumatology and the state of the art research, which is described in Section **??**.

2. **Data Understanding**
   In this phase, basic steps in the data understanding process are carried out, such as data collection, familiarization with the data, and quality checks. This is a critical step in any data science project and hence is also applicable in this thesis.

3. **Data Preparation**
   In this phase the raw data, provided by ZGT, are transformed so that a machine learning model may be trained and evaluated. Therefore, the subsequent phases depend on it. Furthermore, data preparation also depends on the preceding data and business understanding phases. This results in a codependency of the two stages.

4. **Modeling**
   During this phase, the selected modeling techniques are applied to the prepared data. Often, multiple architectures will be tested and tuned subsequently. This depends on the goal of the project. Hence a combination of background and Data understanding serve as a foundation of the decisions made in this phase.

5. **Evaluation**
   Before deploying the model, the efficiency and correctness of the chosen models must be considered. This stage serves as a final opportunity to ensure the performance of the developed solution

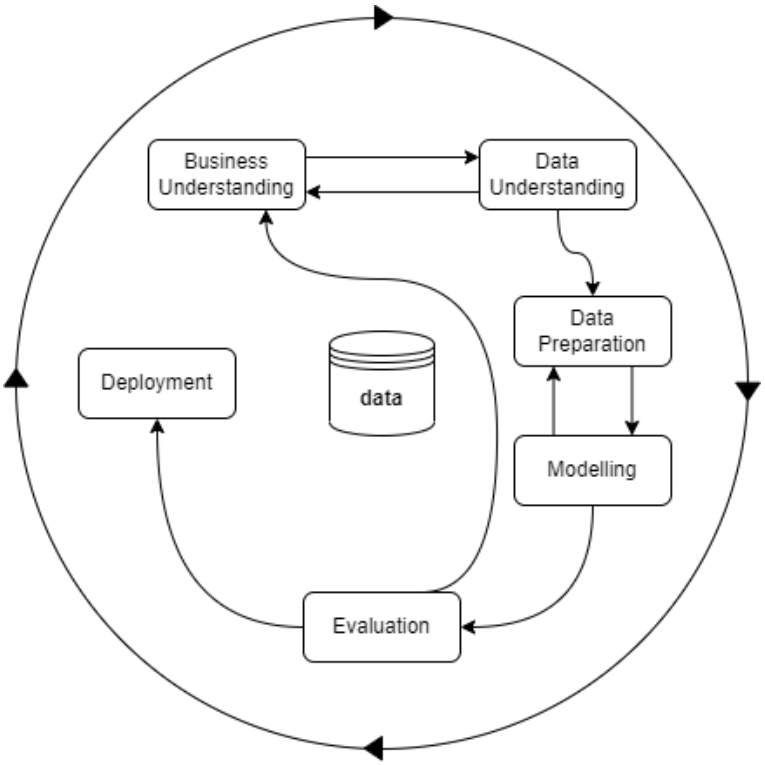


Figure 3.1: The CRISP-DM framework visualized as a cyclic graph.

is desirable. This does not only imply the comparison of performance metrics, it also requires the developer to test whether the model has learnt shortcuts.

6. **Deployment**
   This stage is generally aimed towards deployment. Although real-life deployment is kept in mind, actually deploying the built models is not possible within the scope of this thesis.

3.2 SCRUM

Although CRISP-DM is a suitable framework when a certain goal is kept in mind (see main research question 1), it is challenging to answer main research question 2. Because there is no clear end-point of main research question 2. Given that the problem at hand does not have one distinct solution and hence end-point, aimlessly trying to improve the model can result in a vicious circle. Therefore, I propose to add the following to the CRISP-DM framework, lending ideas from the agile-SCRUM methodology.
The SCRUM method was originally designed to coordinate teamwork. It proposes to use sprints, which are short periods of time in which a certain goal should be reached. These sprints are carefully planned beforehand, which results in a sprint-backlog. This sprint backlog contains all the tasks that are required to reach that sprint's goal. These sprint-backlogs are created by utilizing the product-backlog. Which is a large collection of requirements and idea's that are defined to reach a desirable end-product. Although the teamwork aspect is not crucially important for this research, the iterative nature and planning of sprints are factors that can improve the overall efficiency of the research. Because the end-goal is based on predefined requirements and wishes, the goal can iteratively be defined until a satisfactory result is acquired. So how can this agile teamwork methodology be useful for a data science research? The main idea is to take the concepts of product-backlog, sprint-backlog, and sprints. These are incorporated into the CRISP-DM framework in the following way.

To create a more structured overview of the possible directions of the research, the data gained in the Business understanding and Data understanding are combined with those of the related work. A list, defining the required data preparations is created. It should be used to create a structured pipeline that allows for smooth experimentation in later phases. In the CRISP-DM structure, modeling and data preparation are iterative. This implies that after implementing a certain model, one could go back and

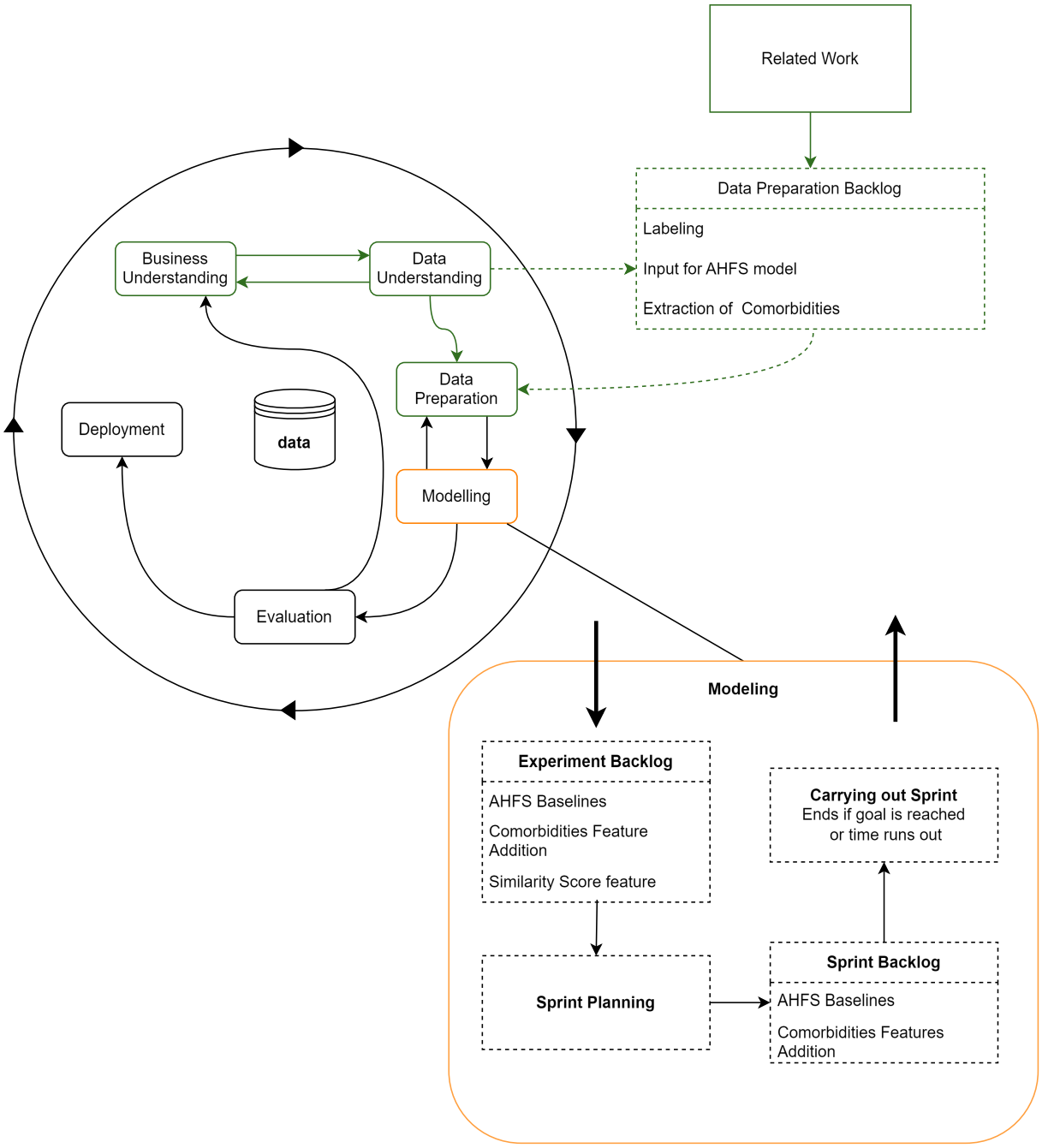


Figure 3.2: This Figure is a graphical representation of the additional concepts, added to the CRISP-DM framework. These concepts are taken from the Agile SCRUM methodology, which is meant to make the process more iterative and suitable for short term projects. The green boxes represent the steps that are affected by or have an effect on the Data Preparation Backlog. This Backlog serves as a guide for the Data Processing Pipeline. The Orange Boxes are those affected by the Experiment Backlog. Which means that they consist of iterative steps that are carried out every 2 weeks. The experiment backlog serves as a list of possible experiments, which are ordered based on their importance.

prepare the data for a new iteration. To make this back-and-forth process as smooth as possible, a data processing pipeline should be created with iterations in mind. This means that it should be robust to different experiments and models, or at least easy to adapt.

After developing the data processing pipeline, the modeling phase can start. In data science, there are plenty of possibilities to explore possible improvements of related work. Therefore, it is crucial to efficiently manage the given time and optional experiments. Therefore, another list of possibilities should be

created beforehand. This list is called the Experiment Backlog (Product Backlog in SCRUM terminology). From the Experiment Backlog, the experiments or additions to those can be selected per sprint. This ensures a finite lifecycle and hence avoids getting stuck in a single idea. With this experiment backlog in mind, the data preparation phase is essentially creating a pipeline to accommodate the different possible experiments. This ensures a structured approach to data preparation, such that new ideas do not require a full revision of the data preparation. Finally, the Experiment Backlog is used to create sprints of 2 weeks, each having their unique Sprint Backlog. With the data prepared, the experiments in the Experiment Backlog can be tackled in order of importance. Figure 3.2 offers a visual representation of this approach.

4 RELATED WORK

This paragraph will give an overview of the related work, related to the topic of postoperative mortality rates for geriatric patients with a hip fracture. The process of analyzing and synthesizing related work is broken down into the following steps, described in the flowchart 4.1.

4.1 Literature Review

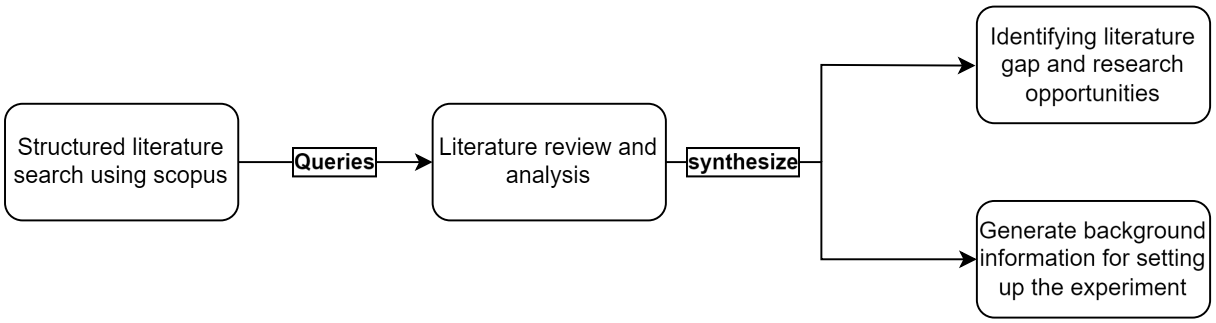


Figure 4.1: Flowchart of the literature review on related work, describing the process from left to right. The queries are logical statements that the Scopus database interprets and uses to generate output. Synthesizing entails the process of converting a stack of raw literature into useful insights.

First, a structured literature review is conducted using Scopus, which is a comprehensive abstract and citation database that covers a wide range of scientific disciplines, including journals, conference proceedings, and patents. It offers extensive coverage of academic literature, facilitating research discovery and citation analysis for scholars, scientists, and institutions around the world. Second, the literature review must be summarized and described in a comprehensive manner. Finally, this information must by synthesized, ultimately resulting in the two main goals; Finding a scientific gap in the literature which can be tackled, and extracting useful information, baselines and data for the experiments.
The first part of the literature review entails the querying of the relevant papers. The query methodology and the definition of ”relevant” will be explained in the following paragraph.

4.1.1 Query

The queries that have been used to extract related papers from the Scopus database are the following:

1. TITLE-ABS-KEY (hip AND fracture AND surgery AND mortality AND geriatric)
2. TITLE-ABS-KEY (hip AND fracture AND mortality AND rate)
3. TITLE-ABS-KEY (hip AND fracture AND mortality AND ((artificial AND intelligence) OR (machine AND learning)))

The reason for excluding anything related to Machine Learning and AI in the first two queries, is to explore methods that include different techniques or fields of study. The latter two queries look at recent work that is especially related to machine learning and AI applications. The results have been ordered into most cited. Afterwards, a recursive search has been applied by looking at papers who had cited the selected query results. Each paper was checked on relevance to the problem definition, described in section 1.3.

4.1.2 Relevance Criteria

Because research in AI and Healthcare tends to develop relatively quickly, relevance is based on the age of the research, as well as the proximity to the definition of the problem. Hence, it was decided to only include papers related to AI and Machine Learning, from the past 10 years (earliest publication date is 2014). For medical papers, the latest version is taken. For instance, the most recent version of a specific research is taken (e.g. the latest research on mortality rates of postoperative mortality for hip fracture in the USA). Furthermore, the paper must at least be related to all of the following topics: Explaining / Predicting postoperative mortality, hip fractures, and geriatrics.

4.1.3 Query Results

Topic	Subtopic	Citation Date
30-day mortality prediction (AI & ML)	Meta Analysis	(Lex, 2023) [3] (Bui, 2023) [10]
	Model Creation/ Prediction improvement	(Li, 2021) [12] (Yenidogan, 2021) [13] (Cao, 2021) [21] (Cary, 2021) [22] (Nijmeijer, 2016) [11] (Debaun, 2021) [23]
	Feature Selection/ Mortality Indicators	(Lin, 2024) [24]
1-year mortality prediction (AI & ML)	Meta Analysis/ Retrospective	(Kitcharanant, 2022) [8] (Forssten, 2021) [4]
	Model Creation/ Prediction improvement	(Xing, 2022) [9] (Cowling, 2021) [25]
Hip fracture mortality (Non-AI related)	Meta Analysis/ Population Studies	(Lin, 2016) [26] (Panula, 2011) [7] (Haentjens, 2010) [27] (Schnell, 2010) [28] (Brauer, 2009) [6]
Non-Mortality & AI for Hip Fractures*	Novel AI techniques	(Murphy, 2022) [5] (VandeBeld, 2022) [29]

Table 4.1: Literature review related to hip fractures, mortality rates and Machine learning/AI. *This section does not concern postoperative mortality, but is still considered relevant due to the case studies in which they are applied.

The queries and relevance features described above have led to the following set of articles (Table 4.1). 21 papers have been found to be currently relevant and closely related to the definition of the problem. This set of papers, concerning postoperative mortality rate of geriatric patients with hip fractures, is subdivided into the following subtopics. 30-day mortality prediction with AI, 1-year mortality prediction with AI, non-AI related research on mortality and non-mortality related research of AI on hip-fractures. The following table provides an overview of these sub-sections. It is noteworthy to mention the overlap between 1-year and 30-day mortality studies. 30-day mortality studies often simultaneously include a forecast for 1-year mortality, however the studies classified as 1-year mortality focus on solely this prediction task.

Within these papers, numerous deep- and machine learning models were presented. The state-of-the-art performance in predicting mortality at 30 days and 1 year is shown in Tables 4.3 and 4.2

State of the Art performance: 1-year post-operative mortality prediction

Model	Report Performance	Sample Size	Most Important Features	# Features	Data Used
Random Survival Forest (Li, 2021)	AUC = 0.75	1.330	Complication Length of Stay Age ASA Creatinine Location Hypopreoteinemia Blood Transfusion (B) Anemi BUN Abnormal	45	Pre- & Post-Operative Patient Age >50
Random Forest (Kitcharanant, 2022)	**AUC = 0.99** **ACC = 0.95**	492	Age Sex BMI CCI Heart Disease Lung Disease Dementia	15	Pre- & Per-Operative Patient Age >50
Random Forest (Xing, 2022)	AUC = 0.81	591	Age COPD Time to surgery Albumin Hemoglobin History of malignancy Perioperative Blood Transfusion	7	Pre- & Per-Operative Patient Age >60
Logistic Regression & XGBoost (Cowling, 2021)	AUC = 0.8	169.646	Age Sex Socioeconomic status ICD-10 Codes	23	Patient Age >60
Logistic Regression (Forssten, 2021)	AUC = 0.74 Specificity = 0.62 Sensitivity = 0.75	124.707	Age Sex ASA score CCI Dementia Congestive Heart Failure Hypertension Surgery Using Pings Chronic Kidney Disease	20	Pre- & Per-Operative Geriatric

Table 4.2: An overview of all created Machine Learning and Artificial Intelligence Models, used to predict 1-year post-operative mortality for patients with hip fractures. Abbreviations: AUC = area under curve, ACC = accuracy, CCI = Charlson's comorbidity index, ASA = American Society of Anesthesliologists

4.2 Literature Review: ZGT

As previously described, ZGT has carried out a multitude of AI related projects in recent years. Some of them have already worked on (sub)problems defined in Section 1.3. This section briefly outlines the two most relevant projects for this research.

Prior to this project, a variation of the Nothing Hip Fracture Score (NHFS) [16] was developed, called the Almelo Hip Fracture Score (AHFS) [11]. The performance of this model was very competitive and simultaneously achieved a high transparency and interpretability. Despite the good results, the confidence in mortality predictions stagnated around 68%, making it too sensitive to errors for real life application.
Second, research of (Berk, 2021) [13] was carried out using the data available at ZGT. This article developed a multimodal approach for predicting postoperative mortality of geriatric patients with hip fractures. Static patient data was combined with features of chest and hip X-ray images (extracted using a CNN). The features of the static patient data were interpreted using a Random Forest model. The results of this paper were positive given the relatively small sample size. A limitation however, was the lack of data on patient comorbidities.
Due to the proven relevance of such features, another project was carried out, attempting to automatically extract such information from patient reports. By utilizing NLP methods, such comorbidities were extracted from text, hence making this data available for future machine learning research.

State of the Art performance: 30-day post-operative mortality prediction

Model	Report Performance	Sample Size	Most Important Features	# Features	Data Used
Lasso Regression (Lin, 2024)	AUC = 0.83	107.660	Acites Disseminated Cancer No wound complications Ventialtor ASA classification: Moribund Totally/Partially dependent Male sex Septic Shock Age Malnourishment	64	Pre-Operative Geriatric
Logistic Regression (Nijmeijer, 2016)	AUC = 0.83	850	Age Sex Admission Serum Hemoglobin Comorbidities Living in an Institution Maligancy Cognitive Frailty Parket Mobility Score (PMS) ASA score	9	Pre-Operative Geriatric
Logistic Regression (Cary, 2021)	AUC = 0.76 ACC = 0.78	17.140	Age Sex Stroke Liver disease Chronic Kidney disease Chronic Hearth failure Lung Disease Depression CCI	15	Pre- & Post-Operative Geriatric
Logistic Regression (Cao, 2021)	AUC = 0.76	**134.915**	Age Sex Hypertension Dementia ASA score RCRI (Revised Cardiac Risk Index)	26	Pre-Operative All Patient Ages
CNN (Cao, 2021)	AUC = 0.76	See Above	See Above	See Above	See Above
ANN (Debaun, 2021)*	**AUC = 0.93**	19.835	Age Sex Failure to wean off ventilator Pre-op ventilator Post-Op Pneumonia History of CHF History of Cancer History of COPD Ascites	47	Pre- & Post-Operative Geriatric
ANN (Cary, 2021)	AUC = 0.76 ACC = 0.74	17.140	Age Sex Stroke Liver disease Chronic Kidney disease Chronic Hearth failure Lung Disease Depression CCI	15	Pre- & Post-Operative Geriatric
Multi-Modal: Random Forest + ResNet on Hip and Chest Images (Van de Beld, 2024) (Yenidogan, 2021)	AUC = 0.78 ± 0.04 Precision = 0.18 ± 0.04 Recall = 0.80± 0.11	1.669	Demographics Daily Living Conditions Nutrition Surgery Information Lab results Medication Comorbidities	**76**	Pre-Operative Geriatric
Random Survival Forest (Li, 2021)*	AUC = 0.83	1.330	Complication Mechanical Ventilation Length of Stay Creatinine levels Age Hypertension Anemia Renal Disease Location Pneumonia	45	Pre- & Post-Operative Patient Age >50

Table 4.3: An overview of all created Machine Learning and Artificial Intelligence Models, used to predict 30-day post-operative mortality for patients with hip fractures. Note that not all studies contain the same characteristics and underlying populations. Abbreviations: ANN = Artificial Neural Network, CNN = Convolutional Neural Network.

4.3 Findings

Based on Table 4.1, 4.2, 4.3 and Section 4.2 the following findings are made. These findings are divided in the following subtopics.

4.3.1 Heterogeneity

It can be concluded that many different factors (such as sample size and population selection) cause a high level of heterogeneity within this research domain. Although this covers a larger variety of research topics, it makes it difficult to directly compare performance results. Therefore, the high performance of some logistic regression and neural network models [23, 8, 12] cannot be taken for granted. A multitude of different aspects must be considered.

4.3.2 Sample Size

First, the variance of the sample size in the selected articles is rather large. Although some articles use data sets of approximately 1000 samples [11, 12, 9, 8] , some exceed the 100.000 mark [25, 4, 21, 24] the heterogeneity in sample sizes can cause differences in performance. A larger sample size is expected to generate more robust and better performing models if all other factors remain static. Three papers however, have achieved a similar AUC of 0.83 [12, 24, 11] with sample sizes ranging from 1.330 to 107.660 patients. This implies differences between the available data, methods, etc.

4.3.3 Datapoints over Time

Second, different types of data are used with respect to time. This means that some papers use pre-, peri- and postoperative data to predict mortality. Naturally, datapoints closer to the point of mortality contribute significantly to the predictive performance. Therefore, such studies can only be interpreted in light of their context. Because this thesis is applied to the CvGT process, the point of inference is made before surgery, which means that only preoperative data can be taken into account. This implies that a fair comparison in predictive power can be made with the following studies [24, 11, 21, 13]. These studies realized an AUCROC of 0.83, 0.83, 0.76, and 0.78, respectively.

4.3.4 Feature Availability

Third, there is a significant difference in the available features and their quality. For example, numerous papers reported a lack of comorbidities, a strong predictor of mortality [21, 13, 4, 25]. This could have potentially caused weak performance compared to similar models, despite the relatively large sample size. More papers have reported that their future work should include a specific feature or the improvement of feature quality [17, 21, 4, 25, 9]. In addition to this, many studies suggested retesting their model on external data, testing the robustness of interhospital deployment [3, 24, 12]. Finally, it was found that many paper under-utilize features(Bui, 2023) [10], due to generalization. Per example, due to generalized comorbidities such as heart failure, without specifying the severity, information is lost.
Adding on the notion of inter hospital testing, a lack in robustness might be observed within the literature. Although interhospital testing can answer whether a model is robust, there is no inherent robustness built into the models. Per example, most hospitals gather similar data, but it is hard to gather exactly the same features across hospitals. To the best of the author's knowledge, little research was conducted on utilizing similar (but not exactly) the same features.

4.3.5 Research Population

Fourth, some differences in research population may be observed, altering the set-up of some experiments. Some articles utilize the whole adult population [21], while the others set a cutoff point at 50, 60 or 70 years old. Because most mortality incidents occur among geriatric patients, model performance is expected to change based on this. Additionally, many studies seem to exclude non-treated patients [9, 23, 8, 11, 15]. Depending on the purpose of the model, useful information could be extracted from non-treated patients. Hence, this may be identified as a gap in the literature.

4.3.6 Metrics

Another insights that was drawn from the literature research, is the similarity between used metrics. Every paper reports mainly on the Area Under Curve of the Receiver Operator Curve (AUCROC). As will be motivated in a later section, this metric is not always suitable for medical research. Here lies an opportunity to compare models based on different metrics, which may reveal new insights into performance differences.

4.4 Related Work: Conclusion

In conclusion, it is not possible to pinpoint a single ”best” performing model. Each model is subject to a given context. For example, the Random Forest of (Kircharanant, 2022) [8] works extremely well for that specific dataset (probably due to overfitting or specific bias present in this hospital). Others utilize the power of Deep Learning with vast amounts of data prior, during, and after surgery. This has resulted in promising post-hoc predictions [23]. When attempting to make predictions in early stages (right after ER), such models might be less practical. Other models score lower on the given metrics, but are easier to interpret, enhancing the practical usability by clinicians [11, 24, 12]. Finally, a promising underutilized technique was developed by (Yenidogan, 2021) [13], which focusses on multimodal learning.

Furthermore, some interesting gaps in the literature may be observed, which do not require and altering of the dataset (or data collection process). First, comorbidities are often excluded from models, although it has been proven to benefit model performance. Patients who did not receive treatment due to the high likelihood of postoperative mortality are excluded in all cases. Second, the AUCROC metric is the most reported, although it is not always the most suitable for highly imbalanced datasets. Third, given the context, there is no model that tests whether features can be substituted in case of unavailability. Per example, could image modalities replace comorbiditiy data and vice versa? These identified gaps may be seen as potential research directions that were considered in this thesis.

5 RESEARCH DIRECTION

In this section the problem context described in Section 1.1, and the described literature review in Section 4.1 is taken into account. With this, the direction of this research is defined. Here, the research is split into two subsequent sections, where the output of Part I will serve as the input of Part II.

5.1 Part I: Adding Comorbidities to a Multimodal Architecture

Given the problem background and the literature review the model architecture described in (Yenidogan, 2021) [13] is the most promising for future research, for a multitude of reasons. From now on this model architecture will be referred to as MM-Y (Multimodal Yenidogan).
First, the model has shown promising results, given the combination of multimodal learning, explainable methods and deep learning. Although missing important features (Comorbidities and socioeconomic status) and being subject to a relatively small samplesize, it achieved a good AUCROC of 0.78 (compared to competing research). Another benefit of this architecture, is that this research also took place within ZGT. Given the improved data quality and NLP extraction methods, this creates an unique opportunity. The limits of the multi-modal model can be tested with a new and higher quality dataset.
Second, the clinical usability within ZGT heavily relies on the conformity with the CvGT treatment plan, where the algorithm will be utilized in pre-operative decisionmaking scenarios. Hence, variables used peri- and post-operative cannot be included in the algorithm, since these data are not given at the time of inference. MM-Y was trained almost exclusively (with the exception of the ASA score), on pre-operative data. Note that ASA Score is excluded due to the unavailability at the moment of model inference in the CvGT treatment process. Refer to Section 2.3 for more details.
Answering whether comorbodities extracted by NLP improve the MM-Y model, completes the first research question. To this end, the re-testing of the MM-Y Architecture will be referred to as Phase I of this research. The resulting model will be utilized as input to Phase II.

5.2 Part II: Working Towards Clinical Implementation

After answering the first research question, plenty of work remains before clinical implementation can be achieved. Therefore, the remainder of this research looks at possibilities to work towards clinical implementation, without changing the experimental setup (same dataset and context). More precisely, this section attempted to iterate over multiple gaps, identified in the literature review (Section 4.3. These mainly entail the use of NOM patients as information source, the use of new metrics and the testing of feature replaceability within a model. However, due to time constraints, the first two options will be elaborated on in this thesis.

5.2.1 NOM Patient Similarity

Ultimately, the goal of mortality prediction is to spare patients a drastic treatment plan, if the outcome almost certainly results in passing away. Therefore, NOM patients (patients who received palliative care) can be seen as another truth value for mortality prediction. A high similarity with a patient who received NOM might help the model forecast mortality. Furthermore, it can also aid to showcase similarities with historic NOM patients. This can strengthen the argument for palliative care. Nonetheless, to the best of the author's knowledge, such patient similarity has not been incorporated into postsurgical mortality prediction in the given context. This underutilized information could improve the predictive performance of the MM-Y model, and is therefore subject to research in Part II.

5.3 Academic Relevance

This research direction holds significant academic relevance in the field of Medical Data Science. The re-validation of the MM-Y architecture has the potential to add extra power to previous findings. This contributes towards the research into multimodal post-surgery mortality prediction and multimodal learning in general. Furthermore, the use of NLP extracted comorbidity features will be evaluated through this approach. Addressing this contributes towards the use of NLP to extract features from raw text, more specifically comorbidity data.
Furthermore, the introduction of a NOM similarity score, to improve classification performance has not been a popular topic of research. This research can contribute towards utilizing NOM patients in post-surgery mortality studies, where palliative care is a realistic option.

5.4 Clinical Relevance

Although this research will not result in a clinically applicable solution, a contribution towards clinical implementation can be made. By keeping clinical relevance in mind while conducting research, different design choices are being made. Per example, as will be explained in Section 6.2.1, performance metrics who take into account clinical relevance are utilized. Furthermore, each research question has been developed, with the aim to work towards clinical implementation. Increasing performance is an excellent starting point, however it is kept in mind that this is only the beginning and that there are plenty of other obstacles to overcome.

6 METHODOLOGY

To act upon the aforementioned research directions (Phase I & II) the following methods are used. It contains the baselines, metrics and experimental setup for both phases. More detail on Phase II will be given in a later section.

6.1 Baselines

A new baseline will be constructed according to the methodology of [13, 11], rather than copying the the exact model weights. This is because the utilized dataset has changed over the past years, hence an exact one-on-one replication is hard to achieve. A misrepresentation would introduce noise into the experiment, resulting in a lack of causal inference. The reproduced baselines will serve as the reference points for the results of the future experiments.

6.1.1 AHFS-b

The primary baseline was derived from the Almelo Hip Fracture Score (AHFS) [11]. AHFS-a was developed in the paper of MM-Y [13], to serve as a baseline for the MM-Y model. Similarly the AHFS-b will be used as a baseline in this thesis, to benchmark the performance of the multimodal model. The reason for the suffixes -a and -b is due to different input data and model architecture, compared to the original AHFS. The altered architecture is a simplified version that consists of a logistic regression model with a stochastic average gradient (SAG) optimizer and l2 regularization. This architecture was copied from AHFS-a [13] with an added regularizer, which slightly improved performance on the validation set. Specifically, the variables used are Age, Female, Delerium, Memory problems, Hemoglobin, Charlson Comorbidity Index (CCI), Maligancy except skin neoplasms, Dementia, Unwanted weightloss, SNAQ score, Katz ADL score, Help with taking shower, help with going to toilet, help with eating, help with transferring bed to chair, fell within last 6 months.
Note that these features differ from those included in the AHFS-a. The feature ”living situation” of the AHFS-a was replaced in the AHFS-b. In the AHFS-b, the features which start with ”help..” are used to indicate living situation. Additionally, the prefracture mobility is not available in the AHFS-b model.

6.1.2 MM-Y

The second baseline that will be used is a close replication of the MM-Y architecture, called the MM-Y Baseline. The multimodal architecture allows for the combination of different data modalities. The first modality being the static data (S), which consists of tabular patient data, extracted from the EHR's. The second modality consists of Medical Hip X-ray images (H). The third consists of the Medical THorax X-ray images (T).
The architecture that combines these 3 modalities is nearly the same as the original proposal by (Yenidogan, 2021).The architecture extracts features from both medical image modalities, by using a ResNet152 for Hip images and Xception for Thorax images. Each image model outputs 8 features, which summarize the input image. These features are combined with the static data and fed into a Random Forest algorithm (RF), which outputs a pseudoprobability, classifying the patient as 1 (diseases after surgery) or 0 (does not disease after surgery), based on the decision threshold. A graphical representation is given in Figure 6.1.

Although the main architecture remained the same, minor changes were applied to the last training layers of the model. It was found that both the ResNet152 and Xception model suffered from the dying ReLu problem [30]. This means that the weights in these layers tend to go to 0. With already a small number

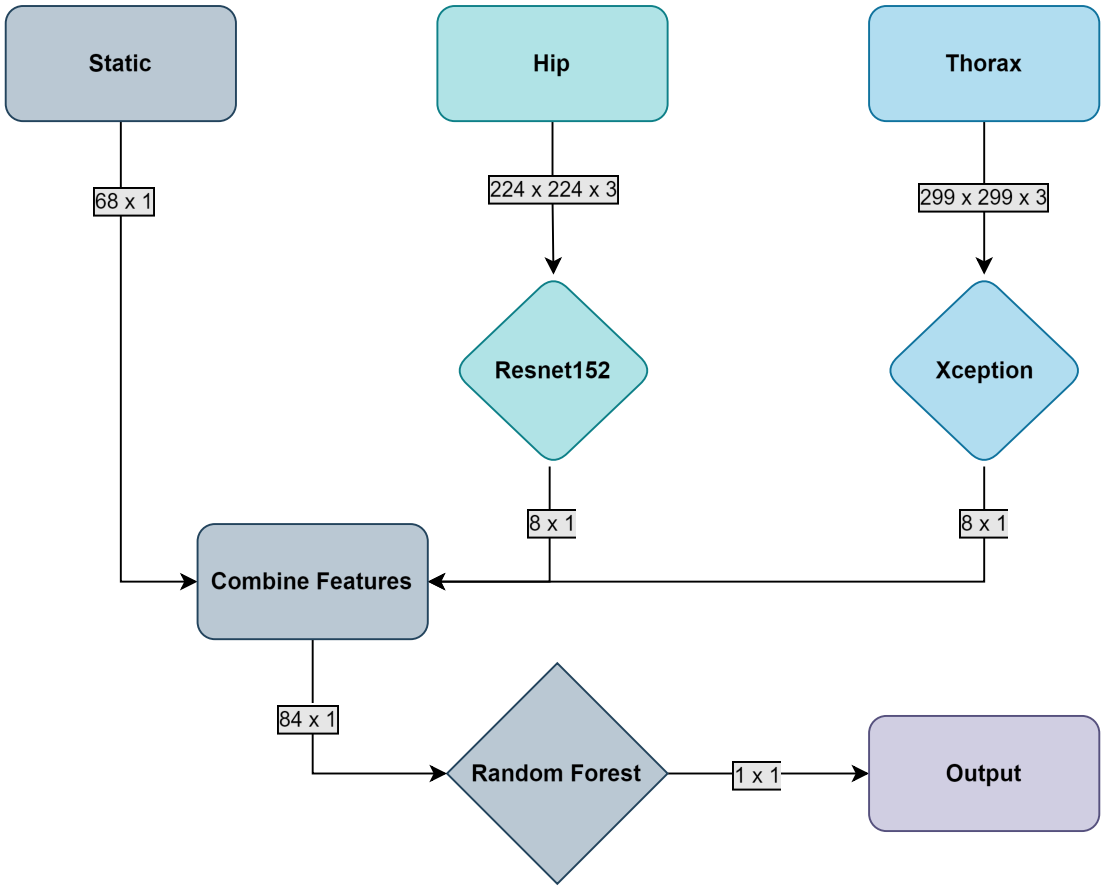


Figure 6.1: High-level overview of the MM-Y architecture, derived from literature [13].

of features (8) to capture the whole image, dead neurons can force all the information into even fewer features. To combat this data loss, a l2 regularizer was proposed by [31]. This decision was copied and applied to the original MM-Y architecture. For more details, refer to Table 6.1 .
Finally, a different dataset was used. Therefore, the just described MM-Y Baseline model will be retrained on this dataset. This allows for a fair comparison with the new models that are introduced in this thesis.

6.1.3 Training, Validating and Testing

The baseline architectures will be retrained on the newly available dataset. This will be done via extraction from the SQL database, preprocessing and combining those data in Python and performing a quality check with HIX, the EHR (Electronic patient files) software. Once this has been achieved, a 60-20-20 train-validation-test split will be made. The 20% newest patients files will make up the test set, resulting in 416 health records. The reason for this rests with the patients who will be hospitalized in the years to come. Their data will be collected in a similar fashion to newest patients. Therefore, testing the model's performance on these data will result in the most realistic and clinically relevant performance indicator.
Although little to no finetuning of the hyperparameters will take place, a stratified cross validation on validation set is still encorporated in the experimental setup. The reason for this rests on the expected data leakage from multimodal training. The Image and Hip modalities will be trained on the same population as the Random Forest which combines the modalities. Therefore, we introduce an extra validation set, making sure no data leakage can take place between the Image modalities and the multimodal training.

6.2 Metrics

To enable clear interpretation of the results, it is critical that the right metrics are utilized. In this section, metrics are explained, and prioritized with subsequent motivation.

Type	Variable	MM-Y Literature	Difference (MM-Y Baseline)
Parameters Image Model			
	Input Size	Hip: (224,224,3) Thorax: (299,299,3)	-
	Weights	imagenet	-
	Pooling	Average	Max
	Non-Trainable Rows	Hip: 0 - 483 Thorax: 0 - 116	-
	Regularizer	-	l2(0.001)
	Activation Function	Layer 1: ReLu Layer 2: ReLu Output: Sigmoid	Layer 1: ReLu Layer 2: Sigmoid Output: Sigmoid
Training Image Model			
	Epochs	100	-
	Callback monitor	Validation AUCROC	-
Parameters Random Forest			
	Estimators	50	-
	Criterion	Gini	-
	Max Depth	5	-
	Min Sample Split	40	-
	Min Sample Leaf	1	-
	Max Leaf Nodes	100	-
	Bootstrap	True	-

Table 6.1: This Table summarizes the features that were used in the best performing architecture in [**?**], and the differences between the MM-Y architecture in this thesis.

6.2.1 Metric Definitions

Performance and performance metrics cannot be seen separately of the context in which they are in. The performance (metric) can be completely misunderstood if the objective it attempts to measure is not clear.
In this thesis the performance is defined as ”to what extent the model behaves as is required”. This behavior is quantified by the following metrics:

Because the main problem is a binary classification problem, the True and False Positives/Negatives are the cornerstones of all used metrics.

- True Positive (TP) = predicted 1 when truth label is 1
- True Negative (TN) = predict 0 when truth label is 0
- False Positive (FP) = predict 1 when truth label is 0
- False Negative (FN) predict 0 when truth label is 1
- Discrimination threshold = the probabilistic threshold in a binary decision problem which determines to which class a prediction will belong.

In the case of this study, 1 means that the patients has deceased within 30 or 100 days after surgery. 0 implies that the patient did not decease shortly after surgery. Given this terminology, the following metrics may be formulated.

- **Recall**, True Positive Rate (TPR) or sensitivity explains how sensitive the model is to predicting the positive class. A recall of 1 implies that every patient with truth value 1 has been predicted as 1 by the model.

$$Recall = \frac{TP}{TP + FN} \tag{6.1}$$

- **FPR**, stands for False Positive Rate, which measures the proportion of patients with truth value 0 have been wrongly classified as 1.

$$FPR = \frac{FP}{TP + FP} \tag{6.2}$$

- **Precision**, tells us something about the success probability of correctly classifying a patient as 1. A Precision of 1 implies that each prediction of 1 had a truth value of 1.

$$Precision = \frac{TP}{TP + FP} \tag{6.3}$$

- **ROCAUC**, stands for Area Under Curve - Receiver Operating Curve. It plots the Recall (or **TPR**) against the **FPR** when the **Decision Threshold** is changed. Interpeting the ROC can give insights into how well the model can distinguish between class 0 and class 1. To summarize this into a single metric, the Area Under the Curve (AUC) can be computed. A value of 0.5 implies random guessing (in case of a perfectly balance binary problem), while 1 implies perfect discrimination. See Figure 6.3 for an illustrative example.

- **AUCPR**, stands for Area Under Curve - Precision Recall Curve. It is similar to the ROC, except that **recall** is plotted against **precision**. The AUCPR is mathematically equivalent to the Weighted Average Precision (Weighted AP), which is defined as follows:

$$WeightedAP = \frac{1}{n} \sum_{i=1}^{n} (r_i - r_{i-1}) p_i \tag{6.4}$$

  Here r represents the recall at threshold i, p represents the precision at threshold i. Note that in this study the threshold i is limited to the interval $[0, 1]$, due to the pseudo-probabilities that the utilized models output. Hence, for each possible threshold i, the precision is corrected for the increase in recall that the change in threshold realizes. The sum is taken of these resulting precisions, after which the Weighted AP remains. See Figure 6.2 for an illustrative example.

- **Max P given Min R** stands for maximum precision given minimum recall. This metric has been introduced by the author in an attempt to capture the highest precision a model can generate, given a minimum required recall. Therefore, this metric returns two values, a Maximum achieved precision and its corresponding recall. It does so by changing the classification decision threshold. It returns the recall and precision where both values are optimized within their constraints. These constrains are defined as follows, refer to Figure 6.2 for a graphical explanation:

  1. minimum recall: a number between 0 and 1 that defines the minimum recall that is accepted by the user of the metric. If set to 0.5, it will not consider any thresholds that result in a recall lower than 0.5.
  2. Sufficient precision: number between 0 and 1. Once the precision reaches this value, it is returned with the highest possible recall. If this value is reached, that threshold generates a sufficiently high precision value.

In related work, the AUCROC is often the metric of choice. Therefore, AUCROC will be used to compare this study's models with other literature. Although its popularity, in this study it is argued that the AUCPR is a more suitable metric. PR curves are often preferred over ROCs in case of imbalanced datasets and when there is a much higher interest in positive labels (label 1 - mortality) [32] [33] [34]. This is due to the nature of the FPR when the data is highly skewed. The FPR is likely to be high for most thresholds, given that there is a larger likelihood for False Positives in a dataset with a small amount of Positive values. These characteristics hold for this study, where a relatively small portion of the patients decease (Around 13%) and where there is a high interest in this class. Therefore, when comparing models developed within this study, the AUCPR will be utilized.

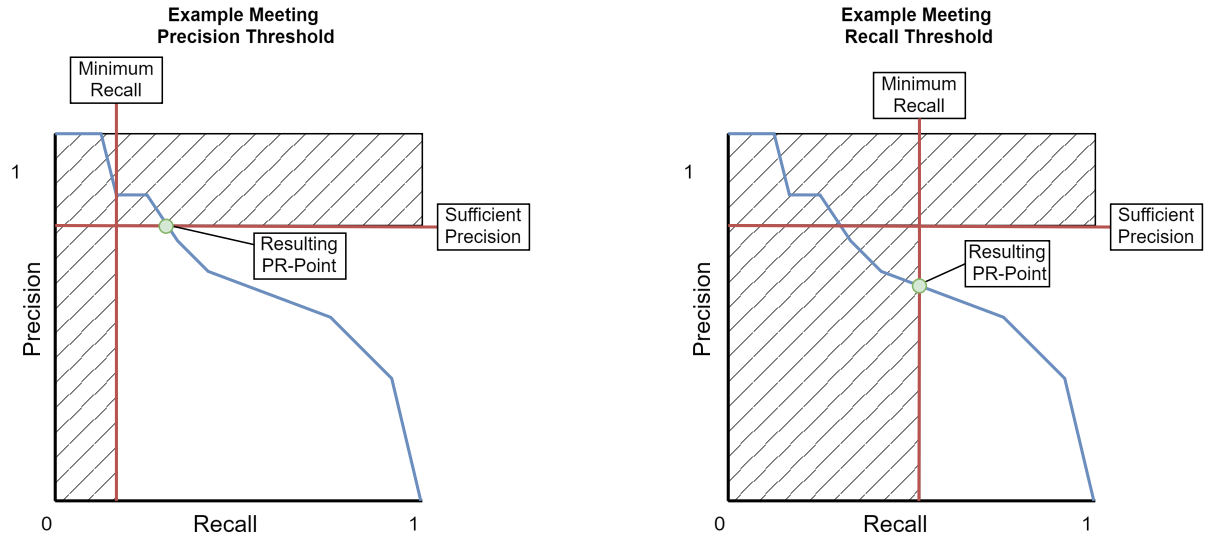


Figure 6.2: **Max P given Min R**: Left shows an example of the returned precision-recall point where the sufficient precision threshold is met, before reaching the minimum recall threshold. Right shows an example of the selected point when the minimum recall constraint is met, before the precision is sufficiently high. The hashed areas of the plot are not reachable for metric computation.

6.2.2 The Importance of Precision

When utilizing the PR-curve (and its AUC) as a comparison between models, it is important to define if there is a difference in importance between precision and recall. Two similar AUCPR values can have different underlying behaviors. One can have only high precision with very low recall and the other can have the opposite, see Figure 6.3. In this specific case, precision is highly important. When a model classifies a patient as highly likely to decease, the probability of being right (precision) should also be very high. This implies that a high precision (0.8-1) is a minimum requirement and models should be compared based on their recall at high precision. A higher recall at e.g. 0.9 precision means that a larger portion of patients who are likely to decease are correctly predicted with high confidence.

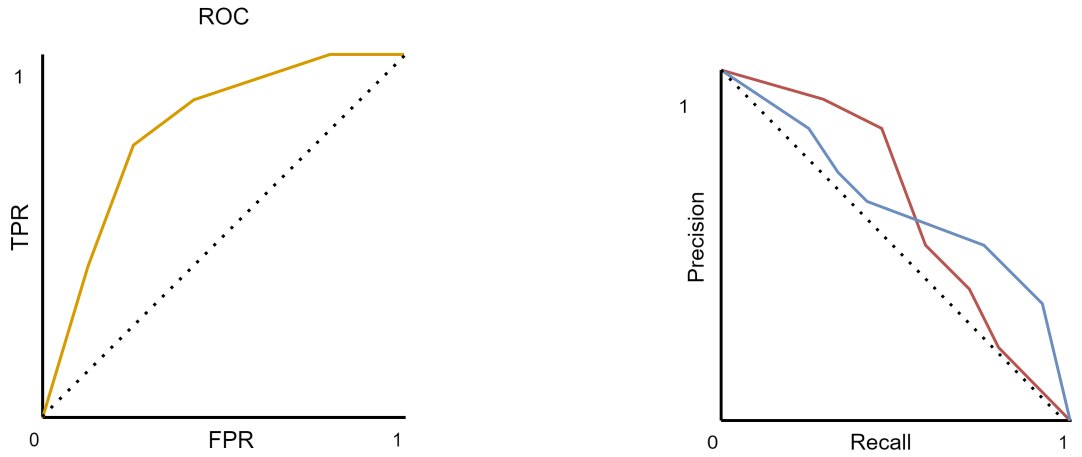


Figure 6.3: Left:An example of a standard ROC curve, where the dotted line implies random discrimination between two binary outcome variables. Right: An example of a PR-curve with similar AUCs, but with different underlying behaviour.

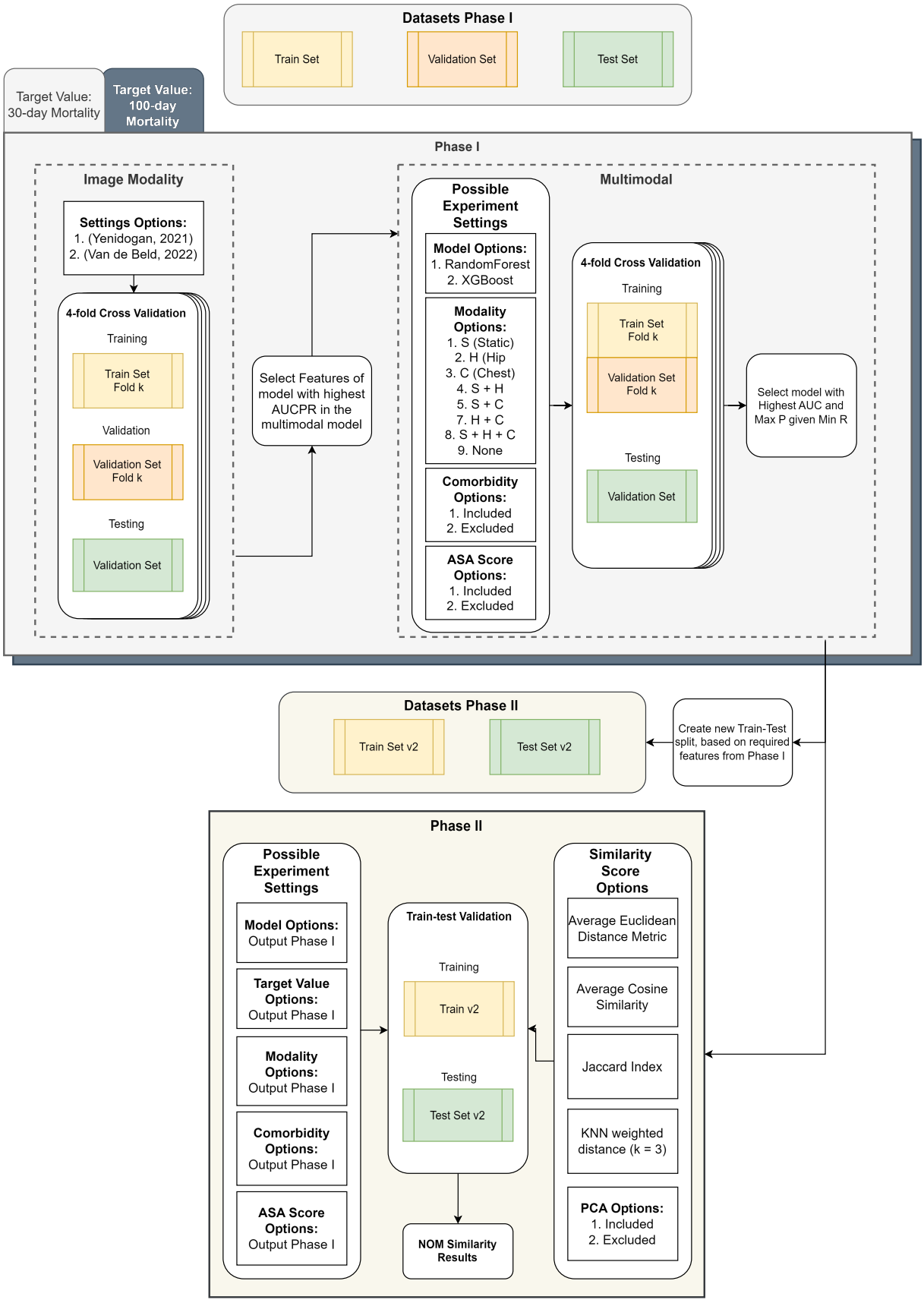


Figure 6.4: This Flowchart displays the experimental setup. Note that the grey boxes display Phase I of the research, which focuses on the addition of comorbidity data to MM-Y. Phase II is displayed in beige, which addresses the addition of Similarity scores to the data. Each ”Options section” implies the different settings, for each type of setting, a combination is made with other settings. Per Example, each Model option is combined which each of the modality options and each Comorbidity option etc.

6.3 Experimental Setup

The experimental setup is visualized in Figure 6.4. Both phases and the number of iterations are showcased. Furthermore, the experiments, train-validation-test sets, and stratified cross validation settings are summarized in Table 6.2. Note that no hyperparameter tuning is incorporated in the experiment. This is considered outside the scope of this research. Furthermore, it has been shown that for this specific architecture, hyperparameter tuning has little to no effect [31]. The only planned tuning of parameters will take place in the training of image and hip modalities. In the original architecture of MM-Y, the extracted features from the image models tend to go to 0. Therefore, (Van de Beld, 2022) [31] has proposed to add l2 kernel regularizes to the model. These kernel regularizers will refrain the weights and hence the feature output to go to 0. The best performing architecture will be chosen based on AUCPR, computed with the Multimodal options in the following experiment.
After modality training, the features, extracted from the hip and thorax radiographs, will be combined with the static patient data. Every option of modality inclusion and exclusion is tested, except for excluding static data as a whole (this was proven to perform significantly worse in (Yenidogan, 2022)). Furthermore, this process is executed twice, altering the target variables between 30 and 100 day mortality. Finally, the addition of ASA score was tested. Although the ASA cannot be utilized in the current clinical setting, it might shed light on whether the ASA score should be generated earlier in the CvGT treatment process. If ASA scores is deemed highly important, inclusion in the model should be considered.
Finally, the best-performing model based on AUCPR and Max P given Min R will be selected. This model becomes subject to testing in Phase II. In Phase II, non utilized modalities are removed from the dataset. The process of missing value removal, missing imputation, data scaling, and train-test splitting is revisited (resulting in train v2 and test v2). The resulting train and test set will be subject to a Phase II baseline, which is created with the chosen settings from Phase I and the train-test v2. Subsequently, each similarity feature is added to the train test v2, with or without dimensionality reduction (PCA). The effects of the addition are measured on the test set, compared to the Phase II baseline. Note that stratified cross-validation will not be carried out for this last experiment, because the test-set is not subject to change. For the motivation, refer to Section 6.1.3.

Experiment	**Training**	**Validation**	**Testing**	**4-fold Stratified Cross-Validation**
Image Modality Feature Extraction	Trainset v1	Validation set v1	Validation set v1	Yes
Multimodal (Static + Image Modality)	Trainset v1	Validation set v1	Test set v1	Yes
NOM similarity scores	Trainset v2	-	Test set v2	No

Table 6.2: Carried out subcategories of experiments, with accompanying train, test an validation sets. Whether 4-fold stratified cross-validation was carried out, has also been indicated.

Part I: Adding Comorbidities to a Multimodal Architecture

7 DATASET

Now that the direction of the research has been defined, the data exploration phase is a critical next step. It serves as the foundation on which to build the subsequent analysis, modeling, and experiments. The goal of this data exploration section is to provide a clear and thorough understanding of the dataset, ensuring that subsequent analysis are based on reliable and well-understood data.
The main data source consists of ZGT's Electronig Health Records. Electronic health records (EHR) or electronic medical records (EMRs [35]), are digital versions of patient paper charts. They contain comprehensive health information, including medical history, diagnoses, medications, treatment plans, radiology images, and laboratory test results. Note that in 2021 the EHR management systems were updated. This resulted in differences over time, which are to be taken into account.
The EHR's are stored in a relational database, which stores the different groups of variables in tables. Each table stores a specific group of variables, such as vital signs or medications. Table 7.1 describes the different tables/variable groups utilized in this study. The merging between the tables is described in a later section.

7.1 Selection Criteria

First, to replicate the results of (Yenidogan, 2021) [13], the included features are adopted as closely as possible. Hence, the feature groups that are taken from [13] are: General Patient information, Vitals, Lab results, Medication, questionnaires, X-ray images, and Emergency Rooms briefs. Appendix Section A displays the variables utilized with their subsequent groups.

Second, in (Yenidogan, 2021) it was established that the data quality before 2015 was insufficient for multimodal training. This argument was based on the high percentage of missing data and the difficulty of accurately imputing large amounts of missing data. Given this reasoning and the pursuit of high data quality, data before 2015 was not included.

Third, only patients who received their screening and treatment in CvGT were considered. This implies that patients who were transferred to a different department or hospital prior to treatment are not included. In addition, patients with periprosthetic or pathalogical fractures are excluded, given that treatment is not carried out within CvGT.

Fourth, only geriatric patients (>70 years) with a hip fracture are included in the research population. Therefore, only patients that match this criteria have been extracted from the database.

Table/Variable Group
Emergency Room (ER)
Emergency Room Briefs (Comorbidities)
Lab Results
Vital Signs
Medication
Survey
X-ray Images (Thorax & Hip)

Table 7.1: This tables describes the Individual tables that were utilized in this study.

7.2 Data Processing

The data processing phase occurred in two phases. First, each table was explored and pre-processed individually (outlier removal, duplicate handling etc.). Second, the processed tables are combined and processed all together (missing imputation, scaling, etc.).

7.2.1 Emergency Room

This table serves as the foundation for the final dataset. Each patient's treatment process starts in the Emergency Room. Here the main selection criteria are applied (such as age restrictions and complication type). Furthermore, information on surgery was added to this initial dataframe, which was utilized to ensure treatment took place within CvGT. This step ensures correct treatment, elimination last moment transfers to different departments (e.g. orthopedics). Simultaneously, patients who received palliative treatment were excluded. The following processing steps occurred:

1. Duplicate rows were removed
2. Emergency room visits that were later cancelled were removed
3. Target variables were created based on the days between ER arrival and mortality
4. Patients who did not receive surgery by a traumasurgeon were excluded

7.2.2 Emergency Room Briefs

The Emergency Room briefs are textual summaries of a patient's ER visit. Here, the patient's story, medication usage, treatment history, and treatment plan are described. Although this is valuable information, textual data is difficult to interpret for most machine learning techniques. An internal study within ZGT has developed a Natural Language Processing technique that extracts comorbidity data from ER briefs. This information is used in this study.

Furthermore, the information in the ER briefs was utilized to determine whether a fracture is periprosthetic or not. Logical statements were used to determine whether the term "periprosthetic" was present in the conclusion of the brief. A manual check was carried out to determined whether the removal of these patients was accurate.

In total the following actions were performed.

1. Duplicate ER brief were removed. In case of slight differences the latest version was taken.
2. Periprosthetic fractures were removed using logical statements.
3. Logical statement mistakes were manually corrected.
4. Text was anonymized using the de-identify package[1]. The de-identify package was developed to anonymize dutch medical records.
5. Classical NLP processing steps were applied such as tokenization, stemming and stopword removal.
6. A one-versus-rest Random Forest, fitted during previous research within ZGT, was utilized to predict the comorbidities given the ER briefs.

The result of these processing steps is the removal of patients who were treated for a periprosthetic fracture, and the extraction of comorbidities using NLP. Compared to (Yenidogan, 2021) this extraction is a new addition to the dataset.

[1] https://github.com/nedap/deidentify

7.2.3 X-ray Images

The multimodality of MM-Y originates from the input of image data. These images were created by the department of radiology and stored in a DICOM (Digital Imaging and Communications in Medicine) file format. The DICOM file format is a standard for storing and transmitting medical imaging information, combining image data and patient information into a single file. For each patient that was selected after the initial criteria, the DICOM file was extracted. Furthermore, due to manual errors in the image labeling, checks were carried out to ensure good data quality.

For each file, the following processing was performed.

1. The DICOM files were split into textual information and .png files
2. The .png files were greyscaled
3. For both Thorax and Hip, only Anterior Posterior (AP) or Posterior Anterior (PA) X-ray images were selected
4. For the Hip, only pelvis X-ray images were selected
5. Due to data-entry errors, some Thorax images were registered under the Hip label. These were manually corrected.

7.2.4 Survey

These variables cover the answers to a standardized list of questions, regarding the background and current status of the patient. These questions are posed to all geriatric patients, treated in the CvGT. In addition, a high percentage of variables need to be transformed into numeric data, because the survey table consists mainly of binary and nominal data, often expressed in text.
An additional difficulty that was found while handling the data is the transition between version of the EHR systems in 2021. During the transition, questions were removed, added, or altered. The variable changes are reported in Table 7.2. Note that some variables remained the same but were given a different variable name. Such cases were combined into a single value. These combinations are displayed in a similar way in Table 7.2.

Removed Features	**Added Features**	**Combined Features**
Pre-Fracture Mobility	Problems Sleeping	Delerium Score + Prone to Delerium*
Living Situation	Help with Mobilization	Fall Risk Score + Prone to Falling
Blood Thinners		Help with Selfcare past day + Help with Selfcare pas 24h
Usage of Incontinence Products		
Prone to Malnutrition		

Table 7.2: This table described the removed, added and combined features, extracted from the patient surveys. * Implies that not only the variable name has changed, also the underlying nature of the variable.

The combined features marked with a * are subject to information loss. Due to a change in the underling nature, the variables cannot be simply concattinated. The delerium score is an ordinal variable ranging from 0 to 3, indicating the severity of the delerium. After 2021 however, the delerium score was replaced for a binary variable, indicating whether someone is prone to a delerium or not. Therefore, the ordinal variable is reduced to a binary variable, mapping all values greater than zero to 1. A delerium score of 0 remains 0 in the binary indication.

To summarize, the following processing steps were performed in the given order.

1. Converted ASA-score to an ordinal variable (example: ”III-severe systematic disease” $\rightarrow$ 3)
2. Filled missing ASA-scores with preliminary projections of the ASA-score
3. Combined the variables described in Table 7.2
4. Converted ”Yes” and ”No” questions to binary

- Yes $\rightarrow$ 1
- No $\rightarrow$ 0

5. Converted ”weightloss” to ordinal value
    - ”6kg or more” $\rightarrow$ 2
    - ”3-6kg” $\rightarrow$ 1
    - ”<3 kg” $\rightarrow$ 0
6. dropped columns with more than 30% of missing values
7. dropped rows (patients) that missed four standard questions in the CvGT survey (Implies high likelihood of not completing the full treatment processes within CvGT).

7.2.5 Vitals

The vital signs used in this study are collected in the Emergency Room. Although vital signs are collected over longer periods of time, it was decided to take into account only the first measurement at arrival. The vital signs selected in this study are noninvasive blood pressure and heart rate.

The following processes were applied to ensure high data quality.

1. The two variables ”Pulse” and ”Hearthrate” are combined due to a namechange after the EHR system change
2. Non-invasive Bloodpressure above 250 was seen as an outlier and removed.

7.2.6 Lab

Similar to Vital signs, lab values are gathered during the emergenecy room visit. Laboratory values describe the presence of certain substances in the patient's blood. Similarly to Vital signs, lab values can be collected multiple times throughout the treatment process. Because the first measurement is most important for deciding the treatment process, the first laboratory measurements are collected. The identification of outliers was done in discussion with a clinical professional. It was determined that no outlier were present in the data and that all lab results were within a realistic range. Furthermore, not all values were stored as numerical values and had to be converted.

The overview of all pre-processing steps is given below.

1. The variables used in (Yenidogan, 2021) were selected
2. Exact duplicate rows were removed
3. Take the first measurement of each patient's ER visit
4. Combine the two different column names that measure Glucose levels (due to EHR system change)
5. The GFR (glomerular filtration rate) measurement, indicates the function of the kidneys. In recent years, a new formula was adopted for estimated the GFR, called CKD-EPI. An older version, called MDRD, is often also available. Both columns are merged with the preference given to the CKD-EPI formula, in case both values are available.
6. Values are transformed to numerical values
    - Angular brackets were removed (> & <)
    - Bloodgroup O was mapped to a binary variable, indicated with 1 if present
    - Positive Irregular Antibodies were mapped to a binary value, indicated with 1 if present

7.2.7 Medications

A similar selection as in the MM-Y architecture was extracted from the database. Here, the ”active” medication was selected at the end of the emergency room visit. This is the moment when medication verification has taken place and the latest status of patient medication is known. Any medication data after the emergency room visit is excluded, since it may be related to the treatment process and hint towards the outcome. E.g. medication to treat a certain post-surgery complication. The medication data are represented by binary variables, each indicating the presence of a certain subgroup of drugs. These subgroups are based on ATC (Anatomical Therapeutic Chemical) Classification. The second order subgroup is utilized to describe the type of medication a patient receives.

An overview of the processing steps is given below.

1. The required ATC codes are selected, based on (Yenidogan, 2021)
2. Dropped rows where the medication was not validated
3. Transform the medication data to dummy variables. Each row indicates the present medication of an unique ER room visit per unique patient.

7.3 Data Merging

After individually assessing and processing the tables, the data must be combined into a single dataframe. Figure 7.1 is a visual representation of the merging. This flowchart shows the sources, the type of merges and the order in which they occurred. First, emergency room, survey, and surgery information were combined. This table forms the baseline of patients that adhere to the restrictions of the research population. Subsequently, the other features (e.g. Vitals and X-ray Images), were added by means of a *Left-join*.

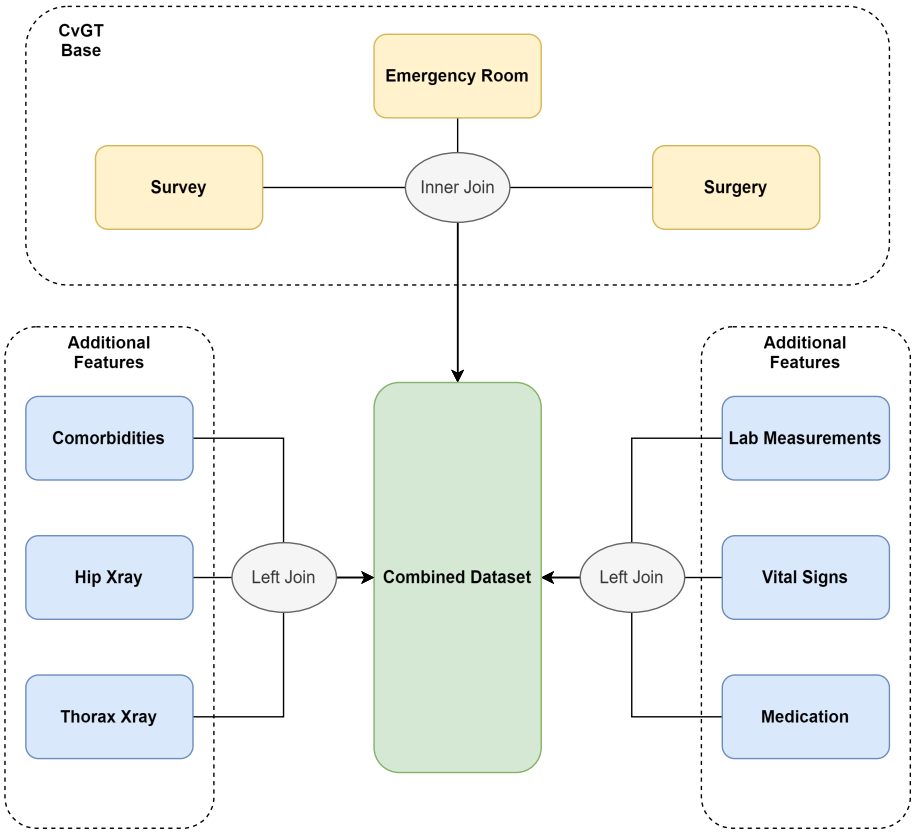


Figure 7.1: Flowchart of different sources of data and their general processing. The CvGT Base is used as the main table, to which the additional features are added with a left-join (if available).

7.4 Combined Dataset

The merging of the individual tables resulted in a single dataframe with 2082 rows and 68 columns. Every row represents a patient's treatment process (duplicate patients may occur if a patient is treated multiple times). Given this dataframe, the overall representation of the research population may be presented. The mean age of the population is 83 ± 6.8. Furthermore, a larger portion of the population is female, although mortality being higher for men (see Figure 7.3 and 7.4. As expected, Figure 7.2) shows that the mortality percentage increases after 80 years and stagnates around 90 years. Furthermore, a high imbalance can be observed in the data set, with the average mortality rate being 13%. Such findings are in line with related work. Additionally, there are large differences between genders within the dataset, hence why this is included in the model. Finally, mortality at 30 and 100 days have large differences in prevalence (See Figure 7.4). The 30 day mortality rate is approximately 7% in this data set, while the 100 day mortality increases to 13%. This is high difference can have a large effects on the predictive power, per target value.

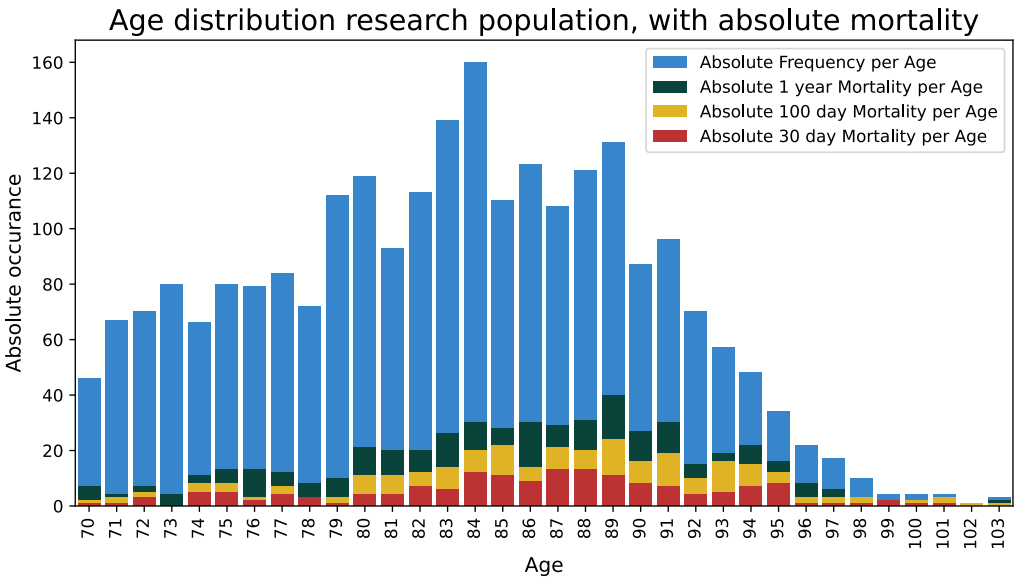


Figure 7.2: Distribution of Age and Mortality in the research population. The numbers are absolute and are computed using the filters previously described.

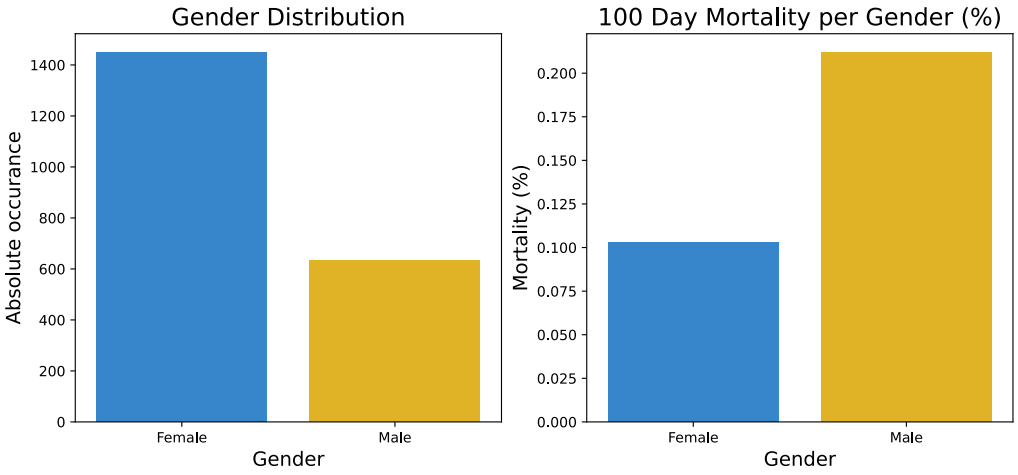


Figure 7.3: The distribution of gender and the respective mortality rate is visualized. It may be observed that a larger amount of female patients are hospitalized, although men have a higher mortality rate per hospitalization.

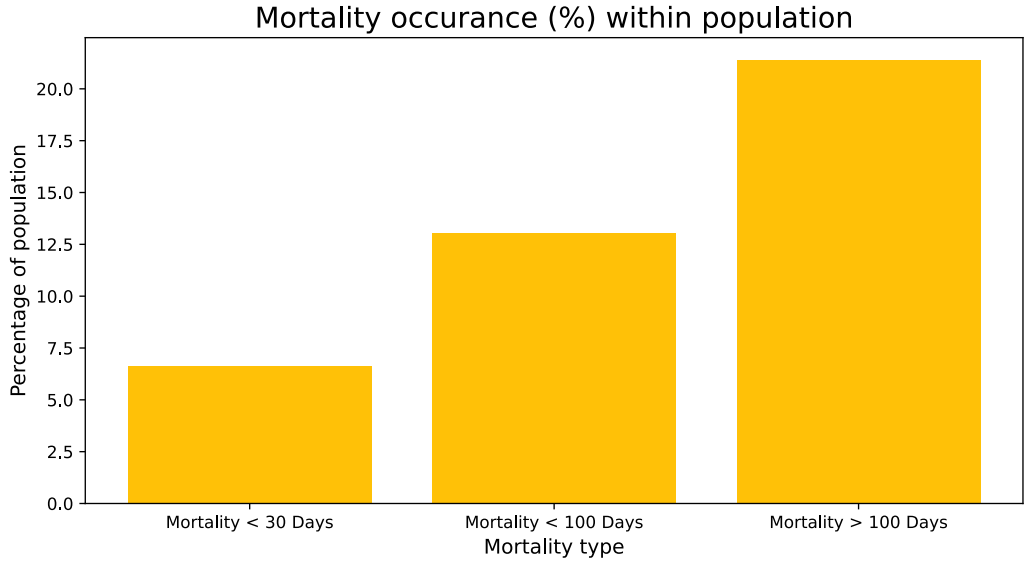


Figure 7.4: Shows the percentage of patient mortality per classification (based on date). Often mortality after 100 days is not directly related to the hip surgery or fracture.

7.4.1 Missings

Although data processing took place on an individual table level, the (left-join) merging creates new missing values. These missing values are visualized in Figure 7.5. It may be observed that variables often miss in patterns. Often for one patient, multiple fields are missing. This problem is dealt with in two different ways. First, when too many missing values are present in a row, it is removed from the research population. The reason for this is the balance between missings removal and imputation. Data quality is of high importance in this study, however it is unrealistic to always expect perfectly filled datasets. Therefore, rows are dropped if more than 2 missings occur in a single variable group (See Figure 7.1, and otherwise imputated. Second, the remaining missing values are imputated using an iterative imputation technique K-Nearest Neighbor (KNN). However, this process takes place after then train-test splitting of the data, and will be explained after the splitting of the data.

7.5 Train-Validation-Test Splitting

The data was split in a 60-20-20 train-validation-test split. The train and validation set are utilized to avoid overfitting and data leakage from the image modality to the static modality. In total, 4 stratified folds of the train-validation split were created. The test set is used to draw a final conclusions after testing different hyperparameters settings in the image modality. The data is split in two different ways.

- First, the test set was created. Since the data contains discrepancies over the years, the test-set was selected based on time, where the 20% newest occurrences are taken. This means that the test-set only contains patients, that were treated in 2022 or later. This ensures that the test-set is representative for the patients that will be hospitalized in the near future.
- Second, the remainder of the dataset (non test-set) is split in train-validation folds, using 4-fold Stratified Cross-validation. This ensures an even distribution of the target value Y amongst all splits.

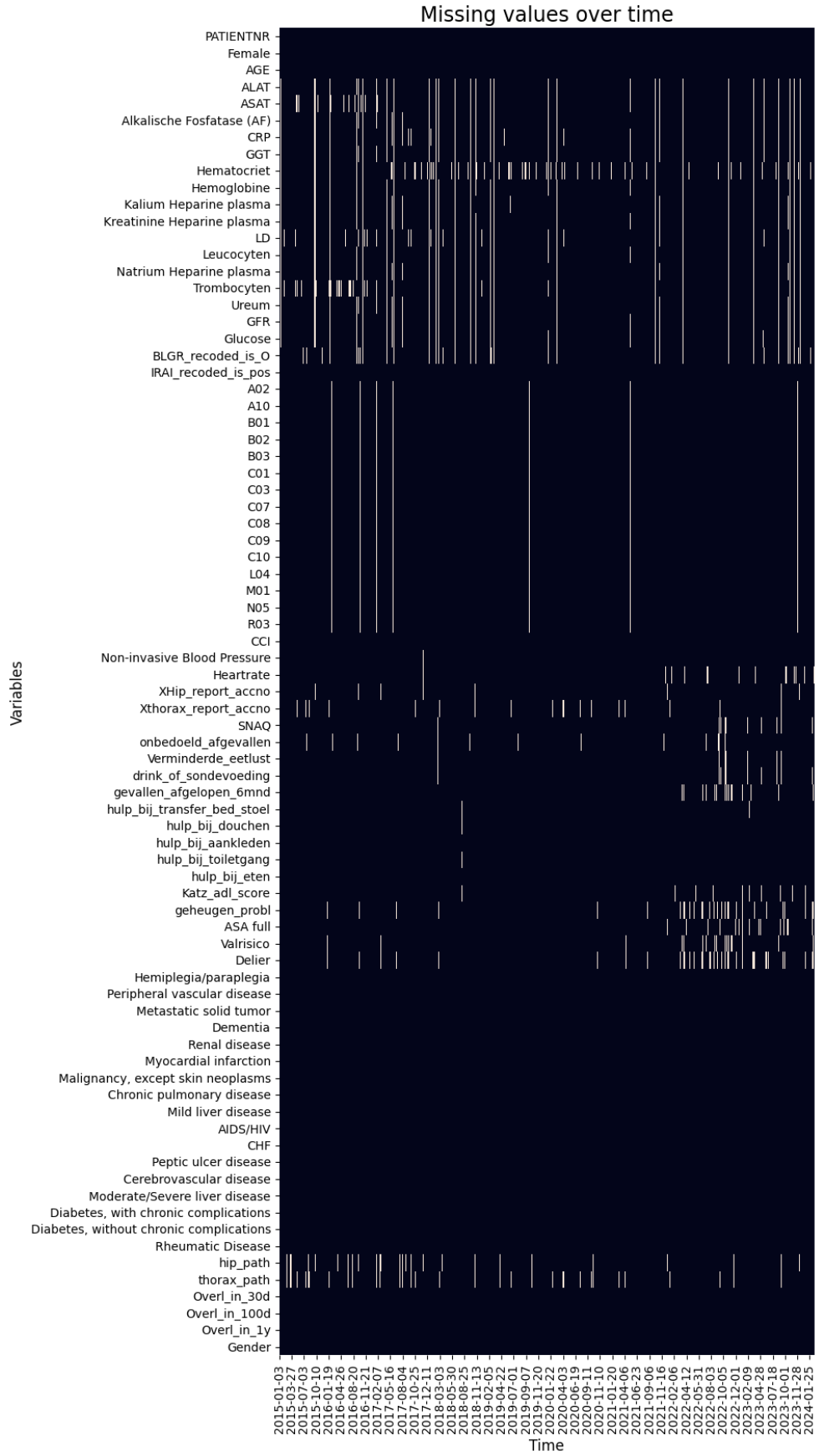


Figure 7.5: This Figure showcases the missing value of the combined dataset, prior to dropping or imputation. On the X-axis, all variables present in the final dataset are showcased. On the Y-axis, the time is visualized in ascending order.

7.5.1 Missing Imputation and Scaling

After splitting the data, the remaining missing values are imputed. In (Yenidogan, 2021), the data imputation strategy was thoroughly tested. Different algorithms were compared, with the Iterative Imputer KneighborsRegressor giving the best results (k = 10 with max 20 iterations). For each different train-validation-test fold (of the 4-fold stratified cross-validation), the imputer weights were fitted on the train set. These weights were then used to impute the missing values of the train, validation and test set. Similarly, a MinMax scaler was fitted and applied to each fold's train set. The resulting scale parameters were used to scale the values in the validation and test set. The reason for fitting the imputation and scaling method for each fold, is to prevent data leakage. Although the effect is expected to be minimal, in real life situations, the validation and test data would not be available to fit the imputation and scaler methods.

8 RESULTS PHASE I

This section presents the results derived from the conducted experiments in chronological order. First, the most relevant results from related work are given in Table 8.1. Second, the literature baselines will be held against this thesis' baselines. Third, the results after the addition of comorbidities to the baseline will be presented. Fourth, the modality and comorbidity inclusions will be compared. Fifth, metrics given the addition of the ASA-score are presented.
Note that the original AHFS and MM-Y baselines included the ASA-score as independent variable. However, the following results exclude the ASA-score in their model, except for the special section dedicated to this.

Model	Test - AUCROC	Test - Precision	Test - Recall
AHFS literature (Nijmijer, 2016)	0.82	-	-
AHFS-a literature (Yenidogan, 2021)	0.72	-	-
MM-Y literature (Yenidogan, 2011)	0.79	0.25	0.3

Table 8.1: This table shows the reported metrics from the original literature of the baselines.

8.1 Results: Image Modality

A comparison was made between the original MM-Y architecture and the proposed version in (Van de Beld, 2022). By validating on the validation split and taking the average, the architecture proposed by (Van de Beld, 2022) resulted in the best performance. While ROCAUC yielded relatively similar results, the AUCPR of the architecture with regularizers slightly outperformed the original architecture. Hence, this image-modality architecture was adopted for feature extraction, which is utilized in the following experiments.

8.2 Results: Reproducing Baselines

Here, the baselines are reproduced with the new dataset, where the ASA-score is excluded, and the two different target variables are used. These baselines will serve as a means of comparison for future experiments.

Model	Target Value	Test - AUCROC	Test - AUCPR	Test - Max Precision	Test - Recall Given Max Precision
AHFS-b Baseline	30-day mortality	0.67	0.15	0.23	**0.22**
AHFS-b Baseline	100-day mortality	0.72	0.27	0.33	0.16
MM-Y Baseline	30-day mortality	0.70±0.01	0.17±0.01	0.31±0.07	0.14±0.04
MM-Y Baseline	100-day mortality	**0.75±0.01**	**0.34±0.01**	**0.54±0.11**	0.19±0.07

Table 8.2: This table shows the results from the replicated MM-Y and AHFS-b. Both target variable 30-day mortality and 100-day mortality are shown. The variables included are based on the literature. For MM-Y, the included modalities are the Hip, Thorax and the Static data (excluding ASA-score).

Table 8.2 describes the baseline results that have been achieved in this study. In Table 8.2 the direct

comparisons with literature baselines, shown in Table 8.1, may be made. In addition the baselines for 100 day mortality prediction are given.

8.3 Results: MM-Y Comorbiditiy Addition

it was decided to continue the future experiments with the target value of 100-day mortality, due to the baseline AUCPR performance increasing greatly. The baselines is altered by adding, NLP extracted comorbidities as features to the MM-Y architecture. The results of predicting 100-day mortality with the MM-Y architecture and added comorbidities, are given in Table 8.3.

Model	Added Variables	Target Value	Test - AUCROC	Test - AUCPR	Test - Max Precision	Test - Recall Given Max Precision
MM-Y Baseline	-	100-day mortality	0.75±0.01	0.34±0.01	0.54±0.11	0.19±0.07
MM-Y Baseline	Comorbidities	100-day mortality	0.76±0.001	0.35±0.01	0.49±0.06	0.20±0.06

Table 8.3: This table shows the effect of the addition of comorbidities, on the best performing model architecture in (Yenidogan, 2021)

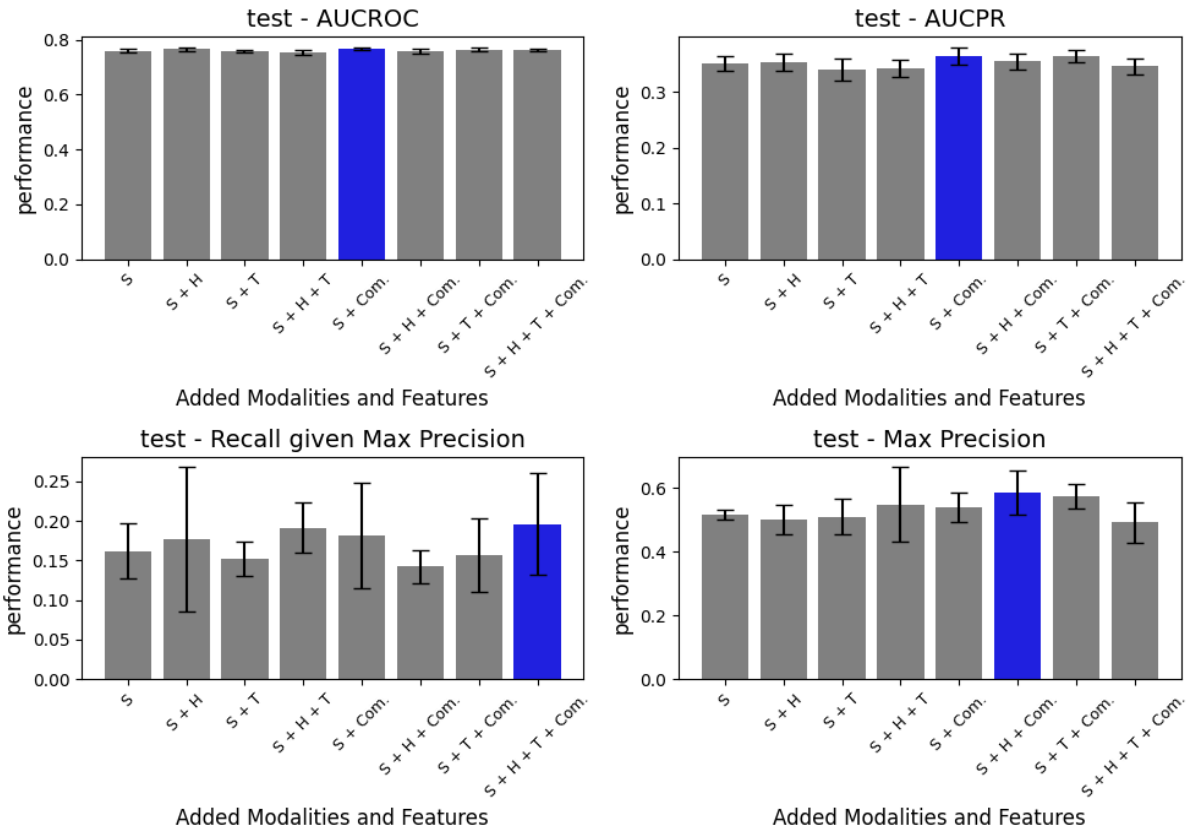


Figure 8.1: This grid shows for each performance metric, the comparison between the inclusion of different modalities. The comorbidity features are also added and shown in the comparison. H = Hip, T = Thorax, Com. = Comorbidity, S = Static data.

8.4 Comparing Modalities

To further measure the effect of comorbidities, they were added to each combination of Static, Hip and Thorax modalities. An iteration without comorbidities was also computed, to serve as a means of comparison. Based on these results, the model for Phase II will be selected based on best AUCPR performance.

Note that the previous results are taken into account by applying the following settings. The target value is 100-day mortality, the ASA-score is excluded because it is not yet known at time of model inference, the comorbidity data are included, and static patient data are always included. Figure 8.1 shows the described findings. It may be observed that different trade-offs occur between the different inclusion of modalities, especially between maximum precision and recall given max. precision. The AUCROC however, seems to be relatively stable between the modalities, while the AUCPR (Average precision) seems to be slightly higher with comorbidities included.

8.4.1 Adding ASA-score

From the previous results, it seems that no model was able to reach a high enough precision (>0.9) for clinical implementation. Therefore, the AUCPR (Average Precision) was chosen to determine the best performing model. Here the model that utilized both comorbidities and static data appears to perform the best. In this section, the effect of the ASA-score on this chosen model is presented.

Model	Modality Label	Target Value	Test - AUCROC	Test - AUCPR	Test - Max Precision	Test - Recall Given Max Precision
MM-Y	S + Com.	100-day mortality	**0.77±.01**	**0.36±0.02**	**0.54±0.05**	0.18±0.07
MM-Y	S + Com. + ASA Score	100-day mortality	0.76±0.01	0.34±0.02	0.47±0.07	**0.18±0.05**

Table 8.4: This table shows the effect of adding ASA-score, to the MM-Y model with Static and Comorbidity data included.

It may be observed that the addition of the ASA-score does not increase the performance for all metrics. Note that can only be observed for this specific model architecture and in combination with additional comorbidities. Therefore, the following settings are selected for the following phase:

- X-ray Modalities are excluded
- ASA-score is excluded
- Comorbidities are included
- Target variable is 100 day mortality

Part II: NOM Similarity

9 PATIENT SIMILARITY SCORES

Based on the findings in Section 4.3, the Patient Similarity Score was found to be underutilized in the given context. Nonetheless, there recent research has been finding promising results when utilizing patient similarity scores [36, 37, 38, 39]. Before investigating the effects of Patient Similarity for this use case, the topic is generally introduced.

The general idea of computing patient similarity is to provide personalized predictions based on the similarity to an index patient (reference patient). Here an index patients may be seen as a reference group, E.g. all patients who suffer from disease A. Given the patient in question and the index patients, a similarity score indicates how much the two are alike. There are multiple ways to utilize such a similarity score to solve a given problem. First, similarity scores can be used to identify a precision cohort from the total population. Per example, the 100 most similar patients (based on patient similarity score) are chosen to be the precision cohort. This precision cohort may then be used to train a specialized model for that specific subgroup. Second, the similarity score can be directly used for treatment selection. A patient may be highly similar to a group of patients that all received a certain treatment [37]. Finally, similarity scores can be used directly to classify patients into groups corresponding to a certain diagnosis (clustering algorithms).
Sharafoddini et al. have performed a meta study into the topic. Based on this meta study, the previously mentioned use-cases are commonly applied to the following problems:

1. Disease Diagnosis (9)
2. Risk assessment of future diagnosis (6)
3. Drug plan development (2)
4. Survival (2)
    - Proved that the addition of a patient similarity score improved the prediction of hospital mortality at 30 days of ICU patients [40].
    - Introduced a Knn method to forecast mortality of patients with renal failure [41]

More specifically, the studies focused mainly on patients with cardiovascular disease, cancer, diabetes, liver disease and renal failure. Only one study looked specifically at mortality [37]. No studies applied their developed solution to any type of fractures.

In addition to the different applications of Patient Similarity Scores, a multitude of algorithms and metrics can be deployed to achieve the desired result. The most used types of algorithms and metrics are the following. Neighborhood-Based Algorithms (E.g. K-nearest neighbors), Distance-Based Metrics (E.g Euclidean, Manhattan and Mahalanobis distance.), Correlation-Based Metrics (E.g. Multiple correlation coefficient), Cosine-Similarity Metrics, Cluster-Based Algorithms (E.g K-means), and Set-Based Metrics (E.g. Jaccard Index).
This short introduction shines light on the large amount of differences between the studies that utilize patient similarity scores. Although the high heterogeneity, there are many gaps to be found with respect to applications. As mentioned previously, patient similarity scores are rarely used for mortality prediction and, to the best of the authors knowledge, never in combination with fractures. The next section shines light on the specific problem that is faced in this thesis, and how patient similarity score can add to the solution.

9.1 Patient Similarity Scores for Hip Fractures

In this thesis a similarity score is a value between 0 and 1, where 0 implies no similarity and 1 an exact match. The similarity score is computed between the 'Patient in question' and the index group "Non Operatively Managed Patients" (Patients who receive palliative care). By doing so, information is captured from patients who did not receive surgery purposefully (essentially what this study aims to predict), which would otherwise be lost. Hence, the most important terms are:

- **Patient in Question** = A patient whose likelihood of mortality will be inferred by the model. In this case, each patient in the dataset described in Chapter 7 is a patient in question.
- **Index Group (reference group)** = Group of patients who have received NOM treatment and diseased within 100 days after fracture.

In summary, for each patient seen in Phase I, the average similarity to the group of patients with NOM is calculated. This score is then used as a feature in the Random Forest algorithm, as described in Phase I. A graphical representation is given in Figure 9.1.

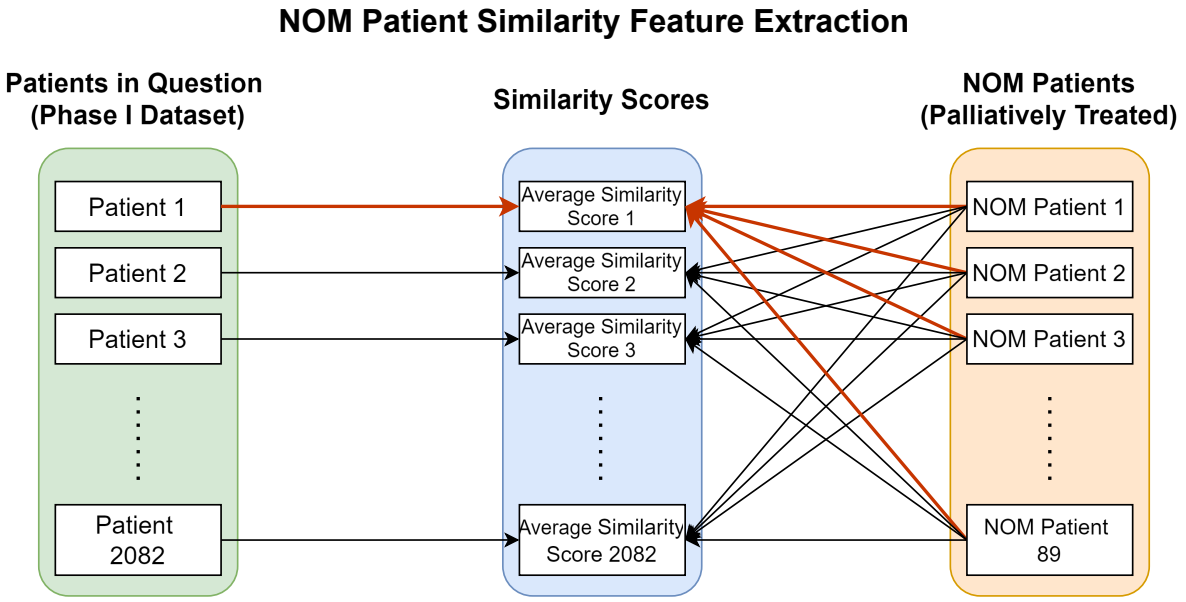


Figure 9.1: A high-level overview of the similarity score computation. For clarity, the process for the first patient is highlighted with red arrows. A similarity score between patient 1 and each NOM patient is computed, of which the average is taken. After computing the similarity score for each patient, the feature is added to the original dataset of Phase I. Only for the KNN metric the average is not computed, all NOM patients are used at once to compute the similarity score.

9.1.1 Index Dataset

The Index Group was extracted in nearly the same manner as described in Section 7. The only difference is that the patient did not receive surgery from a traumasurgeon and diseased within 100 days after suffering the hip fracture. The data was imputed and scaled in the same way as the training data of Phase I. After each filter and the handling of missing data, 89 patients remained in the Index group.

9.1.2 Selected Algorithms & Metrics

Given that there is no prior literature on the usage of patient similarity in this thesis' context, a wide variety of algorithms is tested. First, as a baseline, the average euclidean distance is used between a patient in question and each patient in the index group. This similarity score will be referred to as the Euclidean similarity score and was selected due to its simplicity. Second, the Cosine-Similarity metric was selected, given the good performance in [40]. Third, a neighborhood based algorithm was selected,

with 3-nearest neighbors. It was selected given its popularity and good performance [37] [39]. Fourth, the Jaccard Index was selected, given its good performance in [38] and the simplification of the input data. Because clustering algorithms produce a cluster as output rather than a score, they were not selected in this thesis. Furthermore, the correlation score was not included due to the poor performance and lack of benchmarking against other scores.
In addition, because many of the mentioned algorithms have reduced performance in high dimensionality settings [42], for each metric a PCA version will be computed. This implies that before computation, both the index and the Phase I dataset will become subject to PCA, reducing their dimensions to 40. This reduces the dimensions by 42% while retaining 96% of the information.

9.2 Experimental Setup

The effectiveness of the extracted similarity score will be tested similarly to the ASA score in Phase I. The model will be retrained on a 70-30 train test split, where the test set is constructed in the same way as described in Section 7.5. The only difference is the larger size of the test set, due to the validation set no longer existing, resulting in train-test v2. A baseline will be made with the best performing architecture of Phase I and train-test v2. Afterwards, the baseline will receive one of the four similarity scores (and their PCA versions) described in the previous section. The performance with the added similarity scores will be computed, producing the required results.
Next to each similarity score being individually tested against the baseline, a grid search will be performed to find the best combination of algorithms/metrics and PCA versions.

9.3 Algorithms

Euclidean Distance

The Euclidean Distance Similarity metrics is computed between the patient in question and each index patient. The mean distance is taken over all distances between the patient in question and each index patient.

Algorithm 1 Computation Euclidean Similarity Metric

$P =$ `Matrix of Patients in Question (2082 x 68)`
$D =$ `Empty Vector of average similarity scores (2082 x 1)`
$NOM =$ `Matrix of NOM patients (89 x 68)`

for `n in # rows P` **do**
  $P_n =$ `nth Rowvector of Matrix P (1 x 68)`
  $LoopDistance = 0$

  for `i in # rows NOM` **do**
    $LoopDistance \mathrel{+}= \|P_n - NOM_i\|$
  end for

  $D_n = LoopDistance$ / **#**$rowsNOM$
end for
$MinMaxNormalize(D)$

return $\mathbb{1} - D$ ▷ Invert the score, such that 0 is close and 1 is far

After the computation of each patient's Euclidean Similarity Score, the resulting vector of scores is Normalized using a MinMax scaler. Furthermore, the values are inverted by subtracting the resulting vector from the basis vector.

Cosine Similarity

The cosine similarity metric is computed in a similar fashion as the Euclidean distance metric. For reach patient in question, the average is taken over each cosine similarity with the NOM patients. Afterwards

the results are scaled.

Algorithm 2 Computation Cosine Similarity Metric

$P =$ `Matrix of Patients in Question (2082 x 68)`
$C =$ `Empty Vector of average Jaccard scores (2082 x 1)`
$NOM =$ `Matrix of NOM patients (89 x 68)`

for `n in # rows P` **do**
  $P_n =$ `nth Rowvector of Matrix P (1 x 68)`
  $CosineSimilarity = 0$

  for `i in # rows NOM` **do**
    $CosineSimilarity \mathrel{+}= S_c(P_n, NOM_i)$
  end for

  $D_n = CosineSimilarity$ / #$rowsNOM$
end for
$MinMaxNormalize(D)$

return D

Here the formula S_c represent the cosine similarity which is formulated in the following equation:

$$S_c(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{9.1}$$

Here vectors A and B represent the patient in question and an individual NOM patient.

K-Nearest Neighbors

The K-nearest neighbors was developed using the Scikit learn NearestNeighbors [1] function. To achieve the goal of finding a single similarity score, the patients in question and NOM patients were combined into a single dataset. Here the patients in question and NOM patients received target value 0 and 1 respectively. For each patient in question, its k closest neighbors were found. Parameter k = 3 was selected to make the score more sensitive to close NOM patients. Furthermore, the score was computed by summing the target values and multiplying them with their inverse distance to the patient in question.

Algorithm 3 Computation Knn Similarity Metric

$P =$ `Matrix of Patients in Question (2082 x 68)`
$M =$ `Matrix of all Patients in Question and NOM (2171 x 68)`
$D =$ `Empty Vector of Knn similarity scores (2082 x 1)`

for `i in # rows P` **do**
  $P_i =$ `ith Vector of Matrix P (1 x 68)`
  $Neighbors_i =$ `Vector of` k `target values; 1 if NOM 0 if not`
  $Weights_i =$ `Vector of` k `inverse distances scaled over total distance`
  $D_i = Weights_i^T Neighbors_i$
end for
return D

Jaccard Index

The Jaccard Index was computed using the Scikit learn Jaccard score function [2]. It is defined as the size of the intersection divided by the size of the union of the sets. This function is originally designed to test set similarity between predicted and true binary labels. However, given its proven efficiency in patient

[1] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html#sklearn.neighbors.KNeighborsRegressor
[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.jaccard_score.html

similarity [38], it was adopted in this thesis. To utilize the Jaccard index, the data must be transformed to binary. This was done by taking the median of each feature, subsequently setting each value higher than the median to 1 and lower to 0. This was done for both the patients in question and all NOM patients. The following pseudo-code describes the computation of the Jaccard similarity score, given this binary data.

Algorithm 4 Computation Jaccard Similarity Metric

$P =$ `Matrix of Patients in Question (2082 x 68)`
$D =$ `Empty Vector of average similarity scores (2082 x 1)`
$NOM =$ `Matrix of NOM patients (89 x 68)`

for `n in # rows of P` **do**
  $P_n =$ `nth Vector of Matrix P (1 x 68)`
  $LoopJaccard = 0$

  for `i in # rows NOM` **do**
    $LoopJaccard \mathrel{+}= J(P_n, NOM_i)$
  end for

  $D_n = LoopJaccard$ / **#**$rowsNOM$
end for

return D

Here the formula $J()$ represent the Jaccard Index which is formulated in the following equation:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{9.2}$$

Here vectors A and B represent the patient in question and an individual NOM patient.

10 RESULTS PHASE II

Here, the best-performing model from Phase I (based on precision) was selected. This model was trained on a new train-test split (70%-30%). The similarity scores are individually added to the model to observe the effect on the metrics AUCROC, AUCPR, Max. Precision given Min. Recall. The results are described in the following sections.

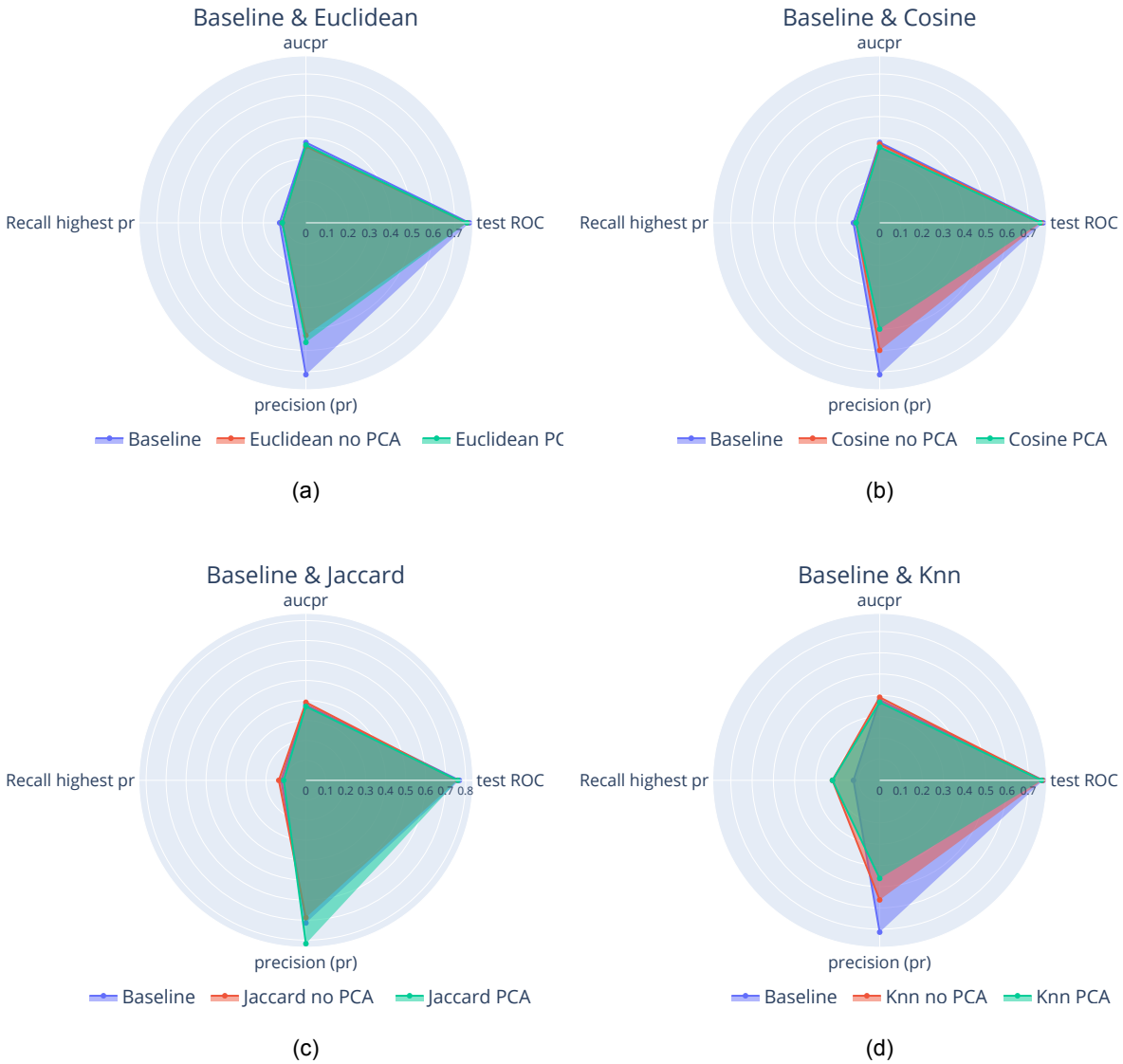


Figure 10.1: Results of individual NOM similarity scores addition to the baseline model. Also includes the Similarity Score computed on a dimensional reduced matrix.

Added Similarity Features	Selection Criteria	Test - AUCROC	Test - AUCPR	Test - Max Precision	Test - Recall Given Max Precision
Baseline Model	-	0.77	0.37	0.71	0.12
Cosine PCA + Jaccard Standard	Highest Precision & Highest AUCPR	0.77	**0.43**	**0.91**	0.12
Euclidean + Euclidean PCA + Cosine + Jaccard PCA	Highest Recall	0.77	0.39	0.5	**0.32**
Euclidean + Euclidean PCA + Cosine PCA + Jaccard + Jaccard PCA + Knn	Highest AUCROC	**0.78**	0.4	0.63	0.12

Table 10.1: This table showcases the models with the highest metric scores. The column ”Added Similarity Features” described the added similarity features to the original baseline. For each metric, the best performing model is shown. Euclidean (PCA) = Euclidean Similarity Score (with PCA), Cosine (PCA) = Cosine Similarity Score (with PCA), Jaccard (PCA) = Jaccard Index Score (with PCA), and Knn (PCA) = weighted 3-NearestNeighbours score (with PCA).

10.1 Effects Individual Similarity Scores

The results from the individual addition of NOM similarity scores are presented in a Radar plot (Figure 9.1), such that a quick comparison between the baseline and addition is possible.
From these radarplots, it may be observed that PCA does not have the same effect on every similarity metric. For the Euclidean distance it improves the performance, while it does the opposite for the cosine similarity metric. However, it seems that both these similarity metrics slightly improve the recall and AUCPR (Average Precision), compared to the baseline model.
For the Jaccard score, it seems that precision increased especially compared to the baseline. Even more when PCA is applied. The Knn similarity metric seems to increase the recall score compared to the baseline, while reducing the precision. Due to these findings and their apparent different characteristics per similarity metric, each combination of the similarity metrics was applied to the model. The results of the best performing model, compared to the baseline, are shown in Figure 10.2 and Table 10.1.

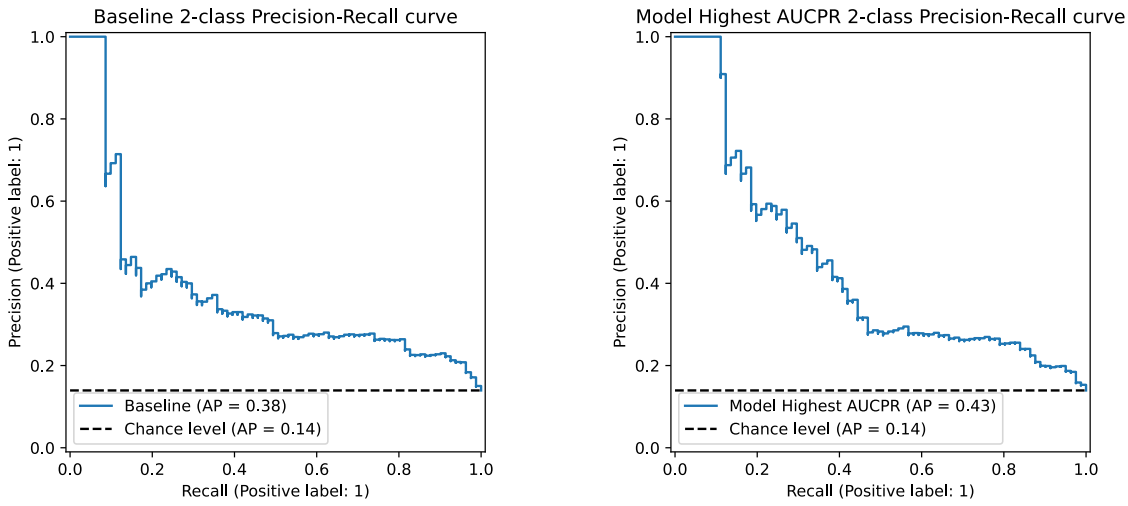


Figure 10.2: Left: Baseline Precision-Recall Curve. Right: Precision-Recall Curve with added PCA Cosine Similarity and Jaccard Index scores.

When observing table 10.1 that each metric may be increased with the right combination of similarity scores. It appears that the addition of PCA cosine similarity and the Jaccard Index can increase the

AUCPR and the Maximum precision that the model can achieve. Given the high importance of these metrics, this model is chosen for further investigation. Therefore, a comparison between this model and the baseline is made, using a Precision-Recall Curve (Figure 10.2.
It may be observed that the increase in AUCPR comes from the higher precision in the lower recall range. This implies that both models still seem to have difficulties to reach high precision at a recall above 50%. Nonethless, the model with added similarity scores appears to be more confident for a larger number of inferences. Additionally, the model with added similarity score seems to have 0.9 precision around the 10% recall range, while the baseline drops to a precision of 0.7 much sooner.

11 DISCUSSION

In this section, the results of Phase I and Phase II will be put into context. The meaning, implication, and limitations of the results will be discussed per experiment. Finally, the discussion will be summarized in clinical and academic recommendations regarding future research towards the topic of post-surgery mortality prediction of geriatric patients with a hip fracture.

11.1 Phase I: Literature and Study Baselines

Prior to the experiments that will answer the research question, the results of (Yenidogan, 2021) and partially those of (Nijmeijer, 2016) were reproduced. The exact reproduction is difficult, as is the intepretation of a difference in performance. This is due to a multitude of reasons.
First, the data collection process within ZGT has changed over the years, resulting in discrepancies between studied datasets. Second, the architectures used in (Yenidogan, 2021) for reproducing the AHFS and for multimodal learning (MM-Y) were directly addopted. This implies that no hyperparameter tuning was applied in this study, which could decrease the reported performance. Regarding the hyperparameters of the Image modality models, merely a comparison was made between the architectures of (Yenidogan, 2021) and (Van de Beld, 2022). Ultimately, the latter was selected.
Third, the data quality of the dataset used in this study is believed to be higher than in the compared studies. This could in turn increase the performance of the model.
Fourth, the ASA-score is excluded from the baselines. This is known to be an important feature, especially for the AHFS. Fifth, the splitting of the data was performed differently, with an extra validation set to avoid data leakage between the multimodal models. This means that the model has less data to train on and that it is certain that no data leakage takes place between different multimodal modalities. Knowing this, when interpreting the results in Table 8.2, it is clear that the AUCROC dropped compared to related literature. As described before, this may be caused by many factors or a combination thereof. Nonetheless, there is still a clear increase in AUCROC, precision, and recall from AHFS-b (logistic regression) to the MM-Y, as reported in [13]. Nonetheless, the overall performance for the 30-day mortality prediction has dropped for both the AHFS and MM-Y architecture.
Additionally, it is clear that the 100-day mortality prediction yields a better performance. This is to be expected, given that the target group becomes larger. In the case of deciding treatment, the clinical relevance of 100-day mortality is comparably to the 30-day mortality. Therefore, it is chosen to continue with the 100-day mortality prediction, given that it is proven to increase the maximum precision.

11.2 Phase I: Effects of Comorbidity Features

Before interpreting the results, described in Table 8.3, the origin of the comorbidity data should be recalled. A One-versus-rest Random Tree Classifier was developed to extract comorbidities from embedded textual data. This model was developed internally within ZGT, and hence trained on an overlapping dataset. This implies that in some cases, data leakage may have occurred. Note that this has no effect on testing whether or not comorbidities contribute to a better performing model. However, it weakens any claims made regarding this specific extraction method.
Taking this into account, when looking at Table 8.3 only a slight increase in AUCROC (0.01) and AUCPR (0.01) is seen. Furthermore, the Max. precision given Min. recall decreases (by 0.05) and the recall given this precision increases slightly (by 0.01). Already it may be observed that the AUCROC and AUCPR do not show the same behaviour as the maximum precision and subsequent recall. Although, the AUCPR increases, it cannot be said that the addition of comorbidities has a large impact on the performance of the MM-Y baseline architecture and does not make a big step towards clinical implementation.

This result could be explained, due to the correlation between the comorbidities and already present variables. The increase of complexity could be disproportionate to the amount of new information that the comorbidities provide. For this specific architecture, it might not have the desired effects, however it may increase the performance of more simple architectures (without extra modalities). Therefore, further experimentation was conducted to answer these questions.

11.3 Phase I: Analyzing Different Modalities

As mentioned before, the addition of the comorbidities on the original MM-Y architecture, does not have a large effect and even reduces the maximum precision. Therefore, to each other combination of modalities, the comorbidities are added. These results are described in Figure 8.1.

First, adding comorbidities to the data seems to have a positive effect on the maximum precision and AUCPR, especially in combination with a single X-ray modality (either Hip or Thorax). However, in combination with both X-ray modalities, these metrics drop. This could be explained by the increased complexity and high amount of multicolinearity. The comorbidities could explain similar information as the image modalities. However, this causal relation is not proven in this experiment and further research would be required. A feature importance analysis could shine more light on this matter, as well as the utilizing a different and more transparent model.

Second, we can observe a clear difference in behavior between the variables. The AUCROC remains the same, while changing the included modalities. Here the static data with added comorbidities comes out on top (AUCROC 0.77). However, the difference, compared to the other models, is minimal. The AUCPR reveals a slightly different behaviour. Because precision is incorporated in this metric, a higher precision for lower recall levels is translated into a higher AUCPR. For the AUCROC, this effect is reduced, due to the higher False Positive Rate. Furthermore, the maximum precision and subsequent recall at that threshold differ per model. This behaviour is not represented well by the AUCROC and AUCPR.

Third, it may be observed that the Maximally achieved precision does not reach above 60%. In other words, none of the models analyzed is more confident than 60% when making mortality predictions. Additionally, the corresponding recall given the maximum precision hovers around 0.15-0.2. As stated previously, clinical implementation requires at least a higher precision to be useful in real life applications and preferably higher recall.

Ultimately, we may conclude that adding comorbidities has a positive effect on the desired behavior (higher maximum precision), but it remains hard to make a generalized statement for all different modality options of the MM-Y architecture. Because the maximum precision does not reach a high enough threshold (at least 90%), it is chosen to continue with the least complex and highest AUCPR model (the model with modalities: Static + Comor.).

ASA Score Addition

With the selected model in the previous section, it was tested whether the ASA score adds valuable information if the comorbidities are added to the dataset. The comparison is describe in Table 8.4. The addition of the ASA score is proven in other studies to have a positive impact on predictive performance. However, this is not reflected in the results of this study. This may be caused by the amount of other variables that are highly correlated with the ASA-score. For example, the KATZ ADL score reports on frailty, comorbidities report on underlying diseases, and survey data reports on current physical capabilities. These results cannot produce a conclusion on the general relevance of the ASA-Score; however, for this selected architecture (Static + Comorbidity data), it does not seem to have a positive effect on prediction.

11.4 Phase II: Effects of Similarity Scores

In Phase II, a novel feature extraction strategy was introduced to increase the performance of the chosen model in Phase I. As a reminder, this model is the Random Forest developed in [13], without the image modalities and with the NLP extracted comorbidity data. Each similarity score was individually added, resulting in different outcomes per score. The Euclidean distance does not improve any of the metrics of interest. The cosine similarity score also does not increase any of the metrics individually. The Jaccard index with PCA improves the Maximum achieved precision by 10%, given a minimum recall of 10%. The Jaccard index without PCA improves the recall by 3 percentage points, given the highest precision. The Knn score with PCA greatly increases the recall (10%), but achieves a much lower maximum precision. In summary, in the given context the Jaccard index and Knn score have proven to increase max precision and recall subsequently.

While these results do not mathematically prove the positive effect of similarity scores, it is clear that in the given circumstances the effect is positive on the precision. Although the AUCROC does not increase in any of the cases, the Max. precision and the AUCPR increase.
Given these positive results, a grid search was carried out to combine the different behaviors of the similarity scores. The best resulting combinations of similarity scores are given in Table 10.1. This results in the following findings.

First, similarly to previous findings, the highest AUCROC does not imply the highest AUCPR and highest Max. precision. Second, by adding the Cosine PCA and the Standard Jaccard Index, the Maximum precision increases to 0.91 (20 percentage points) compared to the baseline. As a result, the AUCPR increases to 0.43 (6 percentage points). Note that no cross-validation was performed due to the test set being selected on the most recent cases (hence, not subject to change). This reduces the strength of the overall claim, however a quite large improvement is achieved over the baseline model.

Second, when analyzing the PR curves described in Figure 10.2, a clear improvement in recall can be observed compared to baseline. After adding similarity score a precision of 0.8 or higher as achieved, for a longer range of recall scores. Additionally, the proposed model is able to achieve a precision of 0.5 or higher up until a recall of 0.5. Although such precision levels are not clinically relevant, it is still a large improvement compared to baseline. Here, the precision falls below 0.5 around a recall of 0.2. However, both models still struggle with achieving precision scores of 0.3 or higher, for recall scores of 0.5 or higher. This implies that many patients are still hard to correctly identify for this architecture. Although the exact effects of each similarity score are not know, an improvement in precision is evident.

In summary, the similarity scores derived from NOM patients clearly improve the clinically important metrics (precision and AUCPR). Due to the lack of parameter tuning and cross-validation, more research is required to validate these claims.

11.5 Recommendations

This section will summarize previous discussion into recommendations. Clinical and academic recommendations are given separately. The purpose of this section is to give ideas for future work, and to recommend certain considerations for developing a model that is ready for clinical implication.

11.5.1 Clinical

The clinical recommendations that arise from the findings in this thesis, revolve mainly around the CvGT treatment process. Instead of altering clinical processes and monitoring the developments, the model is used to evaluate whether the required information is present at the required time. Given this, the following recommendations are made.

1. In the case of the MM-Y architecture (with added comorbidities), the ASA score is not required to achieve a higher precision score. Therefore, with this architecture, the ASA score does not have to be determined before the moment of model inference. It can remain during the pre-operative screening phase.

2. Given that comorbidities have been proven to increase model performance, it is important to act upon this knowledge. Although the current NLP extraction method is a very good starting point, more information can be obtained. Per example, a distinction between severity of a comorbidity could increase the power of the data. Thus, more detailed comorbidity reporting can aid the power of these features.

3. In the current context (available data) it is not recommended to focus on multimodal methods when clinical implementation is the main goal. The complexity of multimodal learning seems to not outweigh the reported performance gains.

4. At the moment of writing, the ZGT database does not poses a clear indication for NOM (for patients with hip fractures) in the EHR system. Adding additional tags could help expedite the data extraction process.

11.5.2 Academic

1. The first recommendation for future work is to focus on 100-day mortality, instead of 30-day mortality. This holds especially when the model is being developed to aid treatment decision. Naturally, due to the increased size of the target group the classification problem becomes easier. Similarly, the clinical relevance remains high, because 100- and 30-day mortality are equally as important for treatment selection.

2. In related work, the most common metrics is AUCROC. However, in this study it is proven that models with similar AUCROC can have different underlying behaviour. Many models had similar AUCROC, but different AUCPR and Precision results. Therefore, it is recommended to analyze the AUCPR for a generalized overview, and to refer to the Precision-Recall curve for selecting the final model. Additionally, it is recommended to include metrics such as f1-score and increase the focus on precision when creating a model fit for medical implementation.

3. The third recommendation revolves around the inclusion of NOM patients. Note that it is both relevant for clinical and academic purposes. When the model intervention rests upon the decision to opt for palliative treatment, this information is shown to be highly relevant. Although some machine learning models might be able to detect such patterns by itself, it is shown that in this study (where a Random Forest model is used), this extracted feature can increase predictive performance. Furthermore, such similarity scores are easy to interpret for clinicians, which adds value to another important factor for clinical implementation, interpretability.

11.6 Limitations

The limitations subject to this thesis are listed below.

1. A limitation of the comparisons drawn between this thesis and the compared research of Yenidogan and Nijmeijer is the size of the train set. The extra validation set in the multimodal training process, resulted in a 20% smaller training set. The effects of this change cannot be measured in the given circumstances, and hence is a factor that cannot be accounted for when interpreting the results.

2. The training of the NLP algorithm, used for extracting the comorbidities, was partially trained on the same dataset. This means that the quality of the extracted comorbidities may be higher for this study than for future data due to inference on train data.

3. Only a small amount of hyperparameter tuning was performed in this thesis. Although hyperparameter tuning usually does not result in a large change in performance, it could lead to an unfair comparison with models who were tuned (e.g. models in related literature).

4. The hyperparameters of the similarity scores were not tuned. This was deemed outside the scope of this thesis, where the influence of similarity scores was question by itself. Altering the hyperparameters may change the individual behaviour of the similarity scores and hence the magnitude of their effects.

12 CONCLUSION

To conclude this study, its findings are summarized and the main research questions are recalled and answered.

12.1 Research Question 1

Main: Does the current state of the ZGT database and the comorbidity data extracted from NLP improve the performance of the multimodal model developed in [13]?

It cannot be concluded that the current state of the database (higher data quality) has a positive effect on the model created in (Yenidogan, 2021). The predictions made for this dataset have resulted in a decrease in AUCROC.
However, the differences between the dataset and methods used in [13] and those of this study are too large to ignore. A conclusion cannot be made on the effect of the new dataset.
Nonetheless, it was found that changing the target feature to 100-day mortality (compared to 30-day mortality, used in the baselines), increased all performance metrics greatly.

subquestion: To what extent do the new features improve the performance of the multimodal architercture [13] and how does this performance compare to the AHFS [11]?.

Furthermore, the addition of NLP extracted comorbidity features does not have a positive effect on the original MM-Y architecture. However, when one or more image modalities is removed from the MM-Y architecture, the comorbidities are shown to increase performance. Adding comorbidity features to only the static data modality increases the Max. precision by 3 percentage points and AUCPR by 1.5 percentage points. Adding comorbidity features to a combination of the Static data and Hip modality increases the Max. precision by 6 precentage points and AUCPR by 0.5 percentage points.
Compared to the AHFS preformance, reported in the original literature, the AUCROC remains lower. However, compared to the reproduction of the AHFS (AHFS-b), the performance increases significantly. The AUCROC increases with 10% compared to this baseline. However, it is not possible to give a statement on the exact amount of performance increase that was caused by the addition of comorbidity features.

12.2 Research Question 2

Main: Related literature has pointed out numerous future research directions that require additional data. However, gathering additional data is a costly process. Given the readily available data, can a gap in the literature be defined and exploited to improve post-surgery mortality prediction in the given context?

These research questions were broken down into a literature research and an experimentation phase. The following two subquestions give answer to the problem.
Subquestion 1: What gaps in the literature may be identified, and which of those may be exploited without changing the current dataset?

The Research Limitations (opportunities) and Literature Gaps that were found are: Data heterogeneity between research, insufficient sample size, selection of irrelevant features, missing features, large research population and metric usage. Note that these limitations hold only for the given context, where 100-day mortality is predicted for geriatric patients with a hip fracture. These opportunities are described

in more detail in Section 4.3. In conclusion, multiple gaps in the literature were defined, that did not require an extension of the current dataset (e.g. more detailed description, expansion of dataset etc.). First, the inclusion of 100-day mortality as target value, which was proven to be beneficial in Phase I. Second, there is little research on the replacement of features in case of unavailability. This could make models more robust when used across multiple health institutions. Third, the inclusion of NOM (Non Operatively Managed) patients, was found to be a novel aspect in postoperative mortality prediction for geriatric patients with hip fracture. The positive effect of the latter was proven by answering the following subquestion.

Subquestion 2: Non-treated patients are often excluded in related work. To what extent does extracting features, such as a similarity score from patients who received palliative treatment, improve the predictive performance?

The inclusion of similarity metrics positively impacts the performance of the multimodal architecture (MM-Y) [13]. The AUCPR (Average Precision) increases with 6 percentage points. The maximum Precision given a minimum recall of 0.1, increases with 20 percentage points. This shows that the inclusion of NOM patients can increase the performance of the Random Forest model, developed in [13].

In summary, this study shows that the inclusion of NLP-extracted comorbidities can improve the AUCPR and precision. The inclusion of the ASA-score is not required for this specific architecture. It is recommended to select 100-day mortality as the target value, since it increases each metric that quantifies model performance. Furthermore, it was found that the AUCROC does not always present the metrics that are of importance for clinical implementation. Finally, the addition of a patient similarity score, derived from palliatively treated patients, greatly improves the AUCPR and Max. Precision.

REFERENCES

[1] Eleni Kanasi, Srinivas Ayilavarapu, and Judith Jones. The aging population: demographics and the biology of aging. *Periodontology 2000*, 72(1):13–18, 2016.

[2] Inspectie Gezondheidszorg en Jeugd Ministerie van Volksgezondheid, Welzijn en Sport. Personeelstekorten in de zorg, Jan 2022.

[3] Johnathan R. Lex, Joseph Di Michele, Robert Koucheki, Daniel Pincus, Cari Whyne, and Bheeshma Ravi. Artificial intelligence for hip fracture detection and outcome prediction: A systematic review and meta-analysis. *JAMA Network Open*, 6:E233391, 3 2023.

[4] Maximilian Peter Forssten, Gary Alan Bass, Ahmad Mohammad Ismail, Shahin Mohseni, and Yang Cao. Predicting 1-year mortality after hip fracture surgery: An evaluation of multiple machine learning approaches. *Journal of Personalized Medicine*, 11, 8 2021.

[5] E. A. Murphy, B. Ehrhardt, C. L. Gregson, O. A. von Arx, A. Hartley, M. R. Whitehouse, M. S. Thomas, G. Stenhouse, T. J.S. Chesser, C. J. Budd, and H. S. Gill. Machine learning outperforms clinical experts in classification of hip fractures. *Scientific Reports*, 12, 12 2022.

[6] Carmen A. Brauer, Marcelo Coca-Perraillon, David M. Cutler, and Allison B. Rosen. Incidence and mortality of hip fractures in the united states. *JAMA*, 302:1573–1579, 10 2009.

[7] Jorma Panula, Harri Pihlajamäki, Ville M. Mattila, Pekka Jaatinen, Tero Vahlberg, Pertti Aarnio, and Sirkka Liisa Kivelä. Mortality and cause of death in hip fracture patients aged 65 or older - a population-based study. *BMC Musculoskeletal Disorders*, 12, 2011.

[8] Nitchanant Kitcharanant, Pojchong Chotiyarnwong, Thiraphat Tanphiriyakun, Ekasame Vanitcharoenkul, Chantas Mahaisavariya, Wichian Boonyaprapa, and Aasis Unnanuntana. Development and internal validation of a machine-learning-developed model for predicting 1-year mortality after fragility hip fracture. *BMC Geriatrics*, 22, 12 2022.

[9] Fei Xing, Rong Luo, Ming Liu, Zongke Zhou, Zhou Xiang, and Xin Duan. A new random forest algorithm-based prediction model of post-operative mortality in geriatric patients with hip fractures. *Frontiers in Medicine*, 9, 5 2022.

[10] Michael Bui, Wieke S. Nijmeijer, Johannes H. Hegeman, Annemieke Witteveen, and Catharina G.M. Groothuis-Oudshoorn. Systematic review and meta-analysis of preoperative predictors for early mortality following hip fracture surgery, 2023.

[11] W. S. Nijmeijer, E. C. Folbert, M. Vermeer, J. P. Slaets, and J. H. Hegeman. Prediction of early mortality following hip fracture surgery in frail elderly: The almelo hip fracture score (ahfs). *Injury*, 47:2138–2143, 10 2016.

[12] Yi Li, Ming Chen, Houchen Lv, Pengbin Yin, Licheng Zhang, and Peifu Tang. A novel machine-learning algorithm for predicting mortality risk after hip fracture surgery. *Injury*, 52:1487–1493, 6 2021.

[13] Berk Yenidogan, Shreyasi Pathak, Jeroen Geerdink, Johannes H. Hegeman, and Maurice Van Keulen. Multimodal machine learning for 30-days post-operative mortality prediction of elderly hip fracture patients. volume 2021-December, pages 508–516. IEEE Computer Society, 2021.

[14] Sverre A. I. Loggers, Hanna C. Willems, Romke Van Balen, Taco Gosens, Suzanne Polinder, Kornelis J. Ponsen, Cornelis L. P. Van de Ree, Jeroen Steens, Michael H. J. Verhofstad, Rutger G. Zuurmond, Esther M. M. Van Lieshout, Pieter Joosse, and FRAIL-HIP Study Group. Evaluation of Quality of Life After Nonoperative or Operative Management of Proximal Femoral Fractures in Frail Institutionalized Patients: The FRAIL-HIP Study. *JAMA Surgery*, 157(5):424–434, 05 2022.

[15] E. C. Folbert, J. H. Hegeman, M. Vermeer, E. M. Regtuijt, D. van der Velde, H. J. ten Duis, and J. P. Slaets. Improved 1-year mortality in elderly patients with a hip fracture following integrated orthogeriatric treatment. *Osteoporosis International*, 28:269–277, 1 2017.

[16] MD Wiles, CG Moran, O Sahota, and IK Moppett. Nottingham hip fracture score as a predictor of one year mortality in patients undergoing surgical repair of fractured neck of femur. *British journal of anaesthesia*, 106(4):501–504, 2011.

[17] Berk Yenidogan and dr ir Maurice van Keulen dr Jasper Reenalda Shreyasi Pathak. Thesis berk: 30-days post-operative mortality prediction of elderly hip fracture patients, 2020.

[18] B.E> Emmerson, M Varacallo, and D Inman. *Hip Fracture Overview*. StatPearls Publishing, 2023.

[19] Mohamed Daabiss. American society of anaesthesiologists physical status classification, 3 2011.

[20] Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1. Springer-Verlag London, UK, 2000.

[21] Yang Cao, Maximilian Peter Forssten, Ahmad Mohammad Ismail, Tomas Borg, Ioannis Ioannidis, Scott Montgomery, and Shahin Mohseni. Predictive values of preoperative characteristics for 30-day mortality in traumatic hip fracture patients. *Journal of Personalized Medicine*, 11, 5 2021.

[22] Machine learning algorithms to predict mortality and allocate palliative care for older patients with hip fracture. *Journal of the American Medical Directors Association*, 22:291–296, 2 2021. see abstract.

[23] Malcolm R. Debaun, Gustavo Chavez, Andrew Fithian, Kingsley Oladeji, Noelle Van Rysselberghe, L. Henry Goodnough, Julius A. Bishop, and Michael J. Gardner. Artificial neural networks predict 30-day mortality after hip fracture: Insights from machine learning. *Journal of the American Academy of Orthopaedic Surgeons*, 29:977–983, 11 2021.

[24] Christopher Q. Lin, Christopher A. Jin, David Ivanov, Christian A. Gonzalez, and Michael J. Gardner. Using machine-learning to decode postoperative hip mortality trends: Actionable insights from an extensive clinical dataset. *Injury*, 55, 3 2024.

[25] Thomas E. Cowling, David A. Cromwell, Alexis Bellot, Linda D. Sharples, and Jan van der Meulen. Logistic regression and machine learning predicted patient mortality from large sets of diagnosis codes comparably. *Journal of Clinical Epidemiology*, 133:43–52, 5 2021.

[26] Hui Shan Lin, J. N. Watts, N. M. Peel, and R. E. Hubbard. Frailty and post-operative outcomes in older surgical patients: A systematic review, 8 2016.

[27] Patrick Haentjens, Jay Magaziner, Cathleen S. Colón-Emeric, Dirk Vanderschueren, Koen Milisen, Brigitte Velkeniers, and Steven Boonen. Meta-analysis: Excess mortality after hip fracture among older women and men, 3 2010.

[28] Scott Schnell, Susan M. Friedman, Daniel A. Mendelson, Karilee W. Bingham, and Stephen L. Kates. The 1-year mortality of patients treated in a hip fracture program for elders. *Geriatric Orthopaedic Surgery Rehabilitation*, 1:6–14, 2010.

[29] Jorn-Jan Van De Beld. Feature importance to explain multimodal prediction models. a clinical use case, 2024.

[30] Lu Lu. Dying relu and initialization: Theory and numerical examples. *Communications in Computational Physics*, 28(5):1671–1706, June 2020.

[31] Jorn-Jan Van De Beld, M Sc Thesis, M Van Keulen, and J Geerdink. Multimodal post-operative complication prediction for elderly patients with hip fractures, 2022.

[32] European Conference and Ecml Pkdd. Machine learning and knowledge discovery in databases, 2013.

[33] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

[34] Jonathan Cook and Vikram Ramadas. When to consult precision-recall curves. *The Stata Journal*, 20(1):131–148, 2020.

[35] health it. What is an electronic health record (ehr)?, Sep 2019.

[36] Sang Ho Oh, Seunghwa Back, and Jongyoul Park. Measuring patient similarity on multiple diseases by joint learning via a convolutional neural network. *Sensors*, 22, 1 2022.

[37] Anis Sharafoddini, Joel A. Dubin, and Joon Lee. Patient similarity in prediction models based on health data: A scoping review, 1 2017.

[38] Zheng Jia, Xudong Lu, Huilong Duan, and Haomin Li. Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity. *BMC Medical Informatics and Decision Making*, 19, 4 2019.

[39] Hao Sen Andrew Fang, Ngiap Chuan Tan, Wei Ying Tan, Ronald Wihal Oei, Mong Li Lee, and Wynne Hsu. Patient similarity analytics for explainable clinical risk prediction. *BMC Medical Informatics and Decision Making*, 21, 12 2021. Uses KNN with weights defined by experts (we do not have this so we just implement KNN).

[40] Joon Lee, David M. Maslove, and Joel A. Dubin. Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PLoS ONE*, 10, 5 2015.

[41] D. J. Lowsky, Y. Ding, D. K.K. Lee, C. E. McCulloch, L. F. Ross, J. R. Thistlethwaite, and S. A. Zenios. A k-nearest neighbors survival probability prediction method. *Statistics in Medicine*, 32:2062–2069, 5 2013.

[42] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In Jan Van den Bussche and Victor Vianu, editors, *Database Theory — ICDT 2001*, pages 420–434, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

A APPENDIX DATASET

Feature	Group
Female	Characteristics
AGE	Characteristics
ALAT	Lab
ASAT	Lab
Alkalische Fosfatase (AF)	Lab
CRP	Lab
GGT	Lab
Hematocriet	Lab
Hemoglobine	Lab
Kalium Heparine plasma	Lab
Kreatinine Heparine plasma	Lab
LD	Lab
Leucocyten	Lab
Natrium Heparine plasma	Lab
Trombocyten	Lab
Ureum	Lab
GFR	Lab
Glucose	Lab
BLGR recoded is O	Lab
IRAI recoded is pos	Lab
A02	Medication
A10	Medication
B01	Medication
B02	Medication
B03	Medication
C01	Medication
C03	Medication
C07	Medication
C08	Medication
C09	Medication
C10	Medication
L04	Medication
M01	Medication
N05	Medication
R03	Medication

Table A.1: This table displays all variables that are used throughout this study.

Feature	Group
Non-invasive Blood Pressure	Vitals
Heartrate	Vitals
SNAQ	Nutrition
onbedoeld afgevallen	Nutrition
Verminderde eetlust	Nutrition
drink of sondevoeding	Nutrition
gevallen afgelopen 6mnd	Independance
hulp bij transfer bed stoel	Independance
hulp bij douchen	Independance
hulp bij aankleden	Independance
hulp bij toiletgang	Independance
hulp bij eten	Independance
Katz adl score	Independance
geheugen probl	Independance
ASA full	Independance
Valrisico	Independance
Delier	Independance
Hemiplegia/paraplegia	Comorbidities
Peripheral vascular disease	Comorbidities
Metastatic solid tumor	Comorbidities
Dementia	Comorbidities
Renal disease	Comorbidities
Myocardial infarction	Comorbidities
Malignancy, except skin neoplasms	Comorbidities
Chronic pulmonary disease	Comorbidities
Mild liver disease	Comorbidities
AIDS/HIV	Comorbidities
CHF	Comorbidities
Peptic ulcer disease	Comorbidities
Cerebrovascular disease	Comorbidities
Moderate/Severe liver disease	Comorbidities
Diabetes, with chronic complications	Comorbidities
Diabetes, without chronic complications	Comorbidities
Rheumatic Disease	Comorbidities
CCI	Comorbidities
hip	Xray
thorax	Xray

Table A.2: This table displays all variables that are used throughout this study.