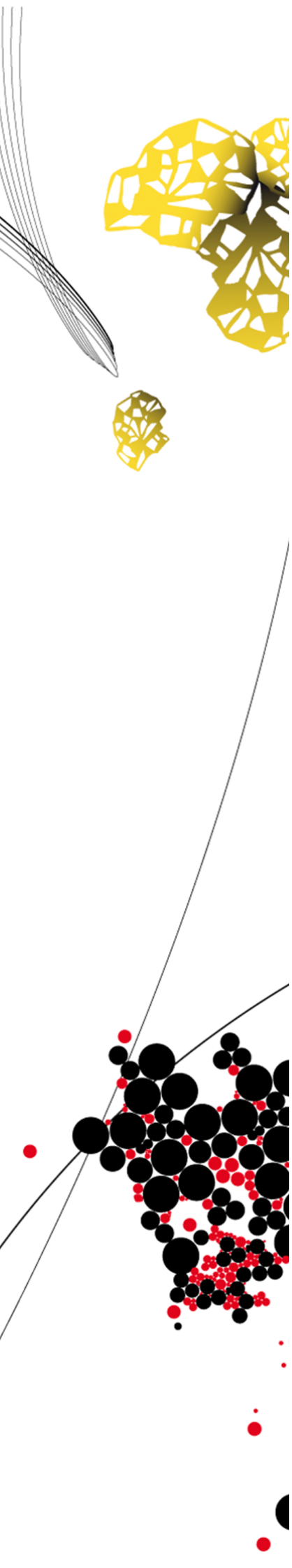

UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics & Computer Science

Multimodal post-operative complication prediction for elderly patients with hip fractures

Jorn-Jan van de Beld

M.Sc. Thesis
June 2022

Supervisors:
Prof. Dr. C. Seifert
S. Pathak Msc.
Dr. J.H. Hegeman
Dr. C.G.M. Oudshoorn
Dr.ir. M. van Keulen
ir. J. Geerdink

Computer Science
Data Science & Technology
University of Twente
The Netherlands

Contents

Part I

Research Paper

Multimodal post-operative complication prediction for elderly patients with hip fractures

Abstract*—Hip fractures are common among the elderly and they have become a major health care problem for society. Standard procedure is to have surgery, often causing complications that can lead to short-term mortality. With the help of an early warning system, we could take precautions to mitigate the consequences of these complications. Machine learning can be used to develop such a system. In this paper, we develop a multimodal deep learning model for post-operative complication prediction using both pre-operative and per-operative data from elderly hip fracture patients. We use ResNet50 to extract features from image modalities and employ LSTM units to extract features from per-operative vital signs. Features from different modalities are combined through early fusion of the features forming a single multimodal prediction model. Further, we also investigate the effect of each modality on the prediction task using SHAP. We evaluate our approach on an in-house data set with 1669 patients. We find that i) our model can predict short-term mortality and heart failure reasonably well and ii) the inclusion of per-operative features does not improve performance of the multimodal model. We use Shapley values to provide local and global explanations for our prediction task. *Our findings imply that using pre-operative static data may be enough to pre-operatively decide about the treatment option for a patient and selecting patients that need close monitoring.

I. Introduction

The absolute number of hip fractures in the older Dutch population ($\geq$ 65 years) almost doubled between 1981 and 2008 from 7,614 to 16,049 [1]. In 1997 it was estimated that in 2050 4.5 million people worldwide will suffer from a hip fracture [2]. Mortality is high during the early post-operative period with rates reported of up to 13.3% in the first 30 days after surgery [3].

An accurate risk score computed before surgery using pre-operative data could provide an objective measure, which aids treatment selection and also leads to a better informed patient [4]. On the other hand, a risk score computed after surgery, which additionally use per-operative data, could give an early warning for complications, allowing for swift measures to mitigate the consequences. In recent years, machine learning (ML) has proven to be promising for clinical practice to aid doctors in clinical decision making, such that it can reduce some their workload [5].

ML models can reliably predict risks on certain complications after surgery [6]. Deep neural networks (DNN) have been used to predict mortality after surgery using pre-operative and per-operative data [7], [8], where AUC scores up to 0.91 were reported for these multimodal models. Regarding hip fracture patients, logistic regression (LR) models have been developed to predict early mortality (<30 days) after surgery only using pre-operative static patient data, where reported AUC scores range from 0.76 to 0.82 [9]–[11]. Yenidogan et al. [12] also included pre-operative hip and chest images, in addition to pre-operative static patient data, in their multimodal model. They extracted features from the images using convolutional neural networks (CNN) and trained a random forest (RF) to predict early mortality, where they report an AUC of 0.79. In this paper, we investigate whether the addition of per-operative data improves the prediction of early mortality after hip surgery.

Specifically, we present a multimodal prediction model combining preoperative static and imaging data, and per-operative medication and time-series data. Furthermore, we do not restrict to mortality as the prediction target, but also include other common complications, specifically: heart failure, pneumonia, anaemia and delirium. We assess the added value of per-operative data, by comparing the performance of our model with those presented in literature, which only used pre-operative data. Additionally, we provide local and global explanations for the model's decisions with the aim to improve understandability for clinicians.

It is crucial that decisions made by a ML model can be understood by clinicians [13]. Some model types like decision trees are explainable simply by design, however more complex models trade performance at the cost of explainability [14]. The models we consider in this paper are not explainable by design, therefore we apply model agnostic explainability methods to understand i) contribution of each modality and ii) individual features.

Following this introduction we discuss related work in Section II, then we examine the data set in Section III. In Section IV we explain our approach and Section V describes the experimental setup. Section VI contains the experimental results. The paper ends with a discussion of the results in Section VII and our conclusions in Section VIII.

II. Related Work

In this section we discuss literature on three aspects relevant to this paper, starting with short-term complication prediction after surgery. Followed by options to connect multiple input modalities to form a single prediction model and last methods to address the explainability of ML models.

A. Short-term complication prediction

Table I summarises a sample of studies in literature featuring short-term complication prediction. It includes a range of patient populations and prediction targets, including hip fracture patients and short-term mortality prediction. The last column specifies which ML model type(s) the authors used, which shows LR models are very common. Next, we discuss a few of these papers in more detail.

Cao et al. [10] developed a model for the prediction of 30-day mortality of adult patients after surgery using static data from 134,915 patients. The authors used the synthetic minority oversampling technique (SMOTE) to counteract class imbalance, such that the ratio was 1:1 between surviving and deceased patients. They compared the performance of a convolutional neural network (CNN) with a logistic regression model (LR) and reported a large difference ($>$0.1) in AUC between the training and test set. They excluded unimportant features to successfully prevent overfitting and their final model scored an AUC of 0.76.

As stated before, this paper follows up on a study that addressed short-term mortality prediction using a similar data set [12]. The authors exploited structured and image data available before surgery and showed significant improvement compared to their baseline the Almelo hip fracture score (AHFS) developed by Nijmeijer et al. [9]. They trained two convolutional neural networks (CNN) to extract features from hip and chest x-ray images, which were fed to a random forest (RF) classifier together with the structured data features. Their multimodal model scored an AUC of 0.786 on the test set outperforming the AHFS baseline, which scored an AUC of 0.717. Thus the authors concluded, that the additional information from multiple modalities is beneficial for model performance.

B. Multimodal prediction

Clinical models that combine multiple modalities outperform models restricted to a single modality [18]. Multimodal models are commonly used for video classification tasks, where audio and image data is processed concurrently [19].

It is important to decide at which point and how information is shared between modalities within a neural network. Late fusion combines the predictions of the unimodal models with no further cross-modal information flow, while early fusion combines modalities as soon as possible [19]. In between is mid fusion, where only later layers are connected.

Early fusion allows for full information flow between modalities, however has a relatively high computational cost, due to the high number of neuron connections. On the other hand, late fusion has low computational cost, but restricts the model from learning cross-modal interactions. Bottlenecked fusion is a special kind of mid fusion, where the number of cross-modal connections is limited, which forces the model to efficiently compress cross-modal information.

C. Explainability

In order to gain the trust of clinicians models require to be explainable, where knowing what features are most important to the model for its prediction is crucial [13]. Furthermore, clinicians need to be able to justify their decision making towards patients and colleagues. ML models can either be explained locally or globally [20]. Local explanations in a clinical setting focus on justifying the prediction for a single patient, while global explanations provide insight in general prediction tendencies for a larger population.

Multiple methods are available to compute the feature importance in deep learning models, where some of the common ones are: LIME [21], deepLIFT [22], layer-wise relevance propagation [23] and Shapley values [24]. In this paper, we use Shapley values to estimate the importance of our input features.

Shapley values are especially suited in case there is multicollinearity in the data, which is common in medical data. For example, a patient might take a certain medicine, have a higher heart rate and elevated blood pressure, all pointing to the same underlying cardiovascular problem. To understand how Shapley values are

TABLE I
OVERVIEW OF LITERATURE WORK FEATURING SHORT-TERM COMPLICATION PREDICTION

Authors	Study population	Prediction target(s)	Data types	ML model(s)
Perng et al. [15]	Septic patients	Short-term mortality	Static patient data	CNN, AE, RF, KNN, SVM
Gowd et al. [16]	Total shoulder arthroplasty	Post-operative complications	Pre- and per-operative static data	LR, GBT, RF, KNN, DT, NB
Schoenfeld et al. [17]	Spinal metastasis surgery	Short-term outcomes including mortality	Pre-operative static data	LR
Lee et al. [7]	Any surgery	Post-operative mortality	Pre- and per-operative data	LR, DNN
Fritz et al. [8]	Surgery with tracheal intubation	Post-operative short-term mortality	Pre- and per-operative data	FC+LSTM+CNN
Cao et al. [10]	Adult[1]hip fracture patients	Post-operative short-term mortality	Pre-operative static data	LR, CNN
Karres et al. [11]	Adult[2]hip fracture patients	Post-operative short-term mortality	Pre-operative static data	LR
Nijmeijer et al. [9]	Elderly[3]hip fracture patients	Post-operative short-term mortality	Pre-operative static data	LR
Yenidogan et al. [12]	Elderly[3]hip fracture patients	Post-operative short-term mortality	Pre-operative static and image data	LR, XGB, RF, SVM

[1] Patients were at least 18 years old [2] Patients were at least 23 years old [3] Patients were at least 71 years old

robust against multicollinearity, we compare them with permutation importance (PI).

In case of PI, the importance equals the loss in performance when training a model with and without a certain feature. If there is multicollinearity in the data, then the absence of a certain feature is compensated by other features leading to an incorrect estimation of the feature importance. Shapley values tackles this problem, by computing the PI for all possible feature subsets and then take the weighted average for each feature. Exact computation of Shapley values is challenging and time expensive, therefore estimation methods have been developed and were made publicly available in the SHAP library [24].

III. DATA SET

The data set contains 1669 anonymized hip fracture surgery cases from the Hospital Group Twente (ZGT) between 2013 and 2021, Table II provides an overview. We included patients older than 70 years at the time of surgery collected from five modalities, which we divide in two groups **pre**-operative and **per**-operative data.

Pre-operative data encompasses information known before surgery, specifically: **static patient data (Static)**, an **axial hip x-ray image (HipImg)** and an **anterior-posterior chest x-ray image (ChestImg)**. The static patient data has 76 features, which we further subdivide in seven categories: demographics, daily living condition, nutrition, surgery information, lab results, medication and comorbidities.

Per-operative data was collected during surgery containing **vital signs (Vitals)** and **medication data (Med)**. The vitals signs are heart rate, pulse, oxygen saturation and blood pressure. We split blood pressure into dystolic, systolic and mean blood pressure leaving us with a total of six temporal features. The medication data includes 17 medication groups, which are commonly administered during hip fracture surgery.

TABLE II
COMPARISON OF OUR DATA SET AND THE PRECEDING PAPER BY YENIDOGAN ET AL. [12]

	Yenidogan et al.	Our data set
Input modalities	Pre-operative	Pre- + per-operative
Outcomes	Mortality	Mortality + complications
Date range	2008-2020	2013-2021
Total cases	2404	1669
Mortality cases	193 (8.0%)	131 (7.8%)

Table II compares our data set to that of Yenidogan et al. [12], our data set is slightly smaller for the following reasons. First, Yenidogan et al. reported low data quality for cases before 2013, due to a high number of missing values in the static patient data, therefore we excluded cases before 2013. Second, we required data on all modalities to be available, except for per-operative medication. In case per-operative medication is missing, we act as if the patient received no medication at all during surgery.

The goal is to predict complications occurring within 30 days after surgery with mortality being the most important, where complications are not mutually exclusive. Table III shows the complications and their prevalence within the data set. The 30-day mortality rate in our data set is 7.8% and in 38.3% of the cases at least one of the considered complications occurred (including mortality). Furthermore, in 43.5% of the mortality cases there is at least one other complication, so co-occurrence of positive labels is common within our data set.

A. Preprocessing

We imputated the pre-operative static patient data iteratively with a KNeighbors Regressor[1] ($k = 10$). We

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

TABLE III
30-DAY COMPLICATIONS AND THEIR PREVALENCE WITHIN THE DATA SET
(N=1669 CASES)

Complication	# unique cases	%
Mortality	131	7.8
Heart failure	96	5.8
Pneumonia	149	8.9
Anaemia	249	14.9
Delirium	293	17.6
No complication	1030	61.7

TABLE IV
MODEL ABBREVIATIONS

Model	Input data
Pre-Static	Static pre-operative patient data
Pre-HipImg	Pre-operative hip image
Pre-ChestImg	Pre-operative chest image
Pre-All	All pre-operative data
Per-Vitals	Per-operative vitals signs
Per-Med	Per-operative medication data
Per-All	All per-operative data
Pre+Per-All	All pre- and per-operative data

processed the vitals data, such that elements are spaced 15 seconds apart, where each element represents a single time step containing six vital signs. We filled up gaps of up to 5 minutes (20 elements) using linear interpolation. Furthermore, given the close similarity of heart rate and pulse, we interchangeably replaced missing values if either one is not missing. So, if heart rate was missing at a certain time step, then we took the pulse at that time step if available and vice versa for missing pulse values. We z-normalized the vital signs for each patient separately, because this makes less assumptions about the data population [25]. If after interpolation an element (time step) still contains missing data, then we set all its values to a masking value. This masking value is recognized by our model, causing the element to be skipped during inference. Appendix B shows an example of the vital signs before and after imputation.

IV. APPROACH

We built the multimodal model, illustrated in Figure 1, by developing a model for each modality separately first and afterwards we fused the representations of the single modalities together. Table IV introduces an abbreviation for each model that we consider in this paper. We trained the unimodal models to only predict mortality and transferred the learned weights to the multimodal models. The multimodal models have multiple outputs, so they predict mortality independently from the other complications, where the complications are grouped together resulting in a multi-label prediction task. We used two separate loss functions, one for mortality prediction and one for the grouped complications, where we took the weighted sum to compute the total loss during training. Equation 1 describes how the total loss L_{total} for patient x was computed using the mortality loss L_m, complication loss L_c and complication weight w_c. This gave us another tunable parameter (w_c), which we set such that the addition of complication prediction did not harm mortality prediction performance.

$$L_{total}(x) = L_m(x) + w_c \cdot L_c(x) \tag{1}$$

The remainder of this section explains our approach to each unimodal model and concludes on how we fused them together.

A. Pre-operative models

We discuss the three pre-operative unimodal models: the Pre-Static model for the pre-operative static data and the Pre-HipImg and Pre-ChestImg models for the hip and chest images.

For our Pre-Static model, the main task was dimensionality reduction, so at a later stage it provided a similar number of features as the other modalities to the multimodal prediction model. Therefore, we used a fully connected hidden layer, from which the output was used in the multimodal models.

Convolutional neural networks (CNNs) have emerged as a powerful tool for medical image classification [26], so we used CNNs for our Pre-HipImg and Pre-ChestImg models to extract features from the hip and chest images. A wide range of CNN architectures are available, but given our small data set size we chose the relatively small ResNet50 for both image types [27].

These pre-operative unimodal model choices differ from the reference paper, where the authors used a random forest model for the pre-operative static data, a partially trained ResNet152 for the hip images and a fully trained XCeption model for the chest images [12]. Using fully connected layers instead of a random forest for the Pre-Static model made it easier to combine with the other unimodal models, in turn making it possible to be trained simultaneously. Regarding the image models

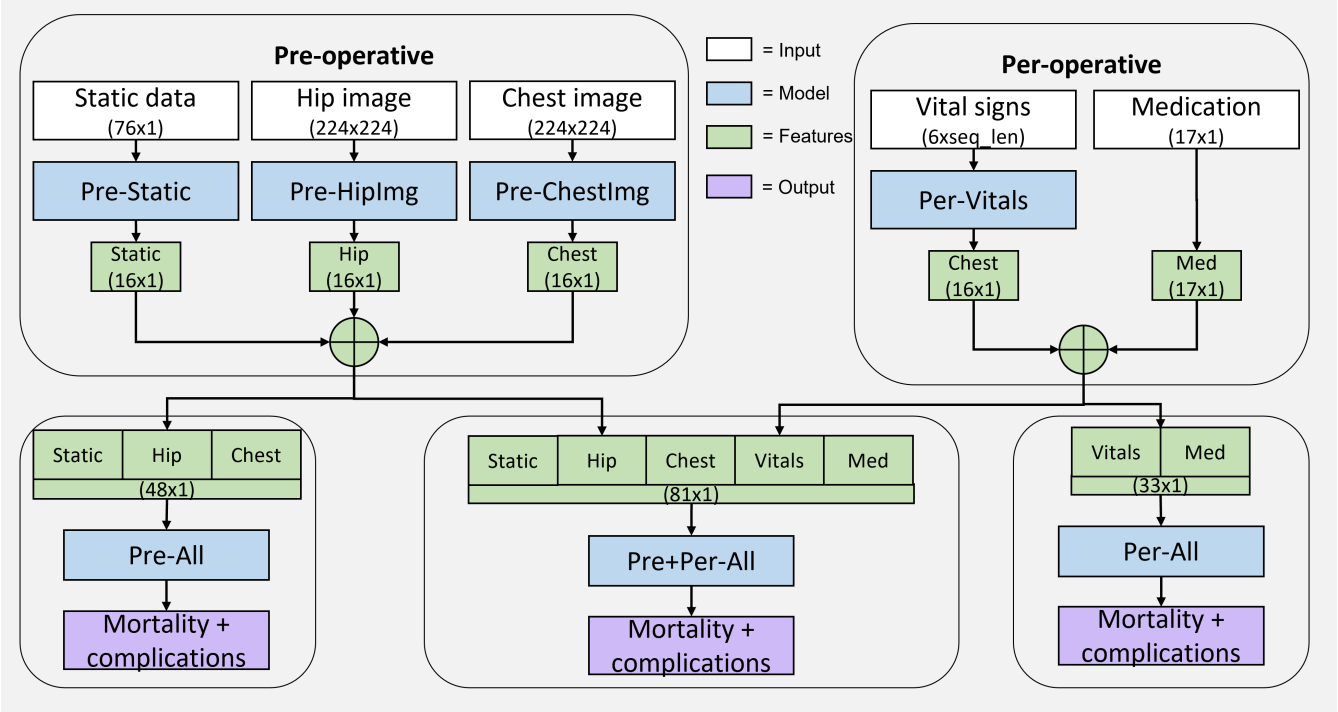

Fig. 1. Overview on how the unimodal models are fused together to form the multi-modal models. Dimensions at the input and feature extraction layers are shown in between brackets, where the sequence length (*seq_len*) for the vitals varies between patients. No feature extraction was done for the per-operative medication data.

we preferred them to have the same architecture, in order to reduce the complexity of the multimodal models.

B. Per-operative models

The Per-Vitals model takes multivariate temporal data as input, where we used bidirectional long short-term memory (LSTM) units to extract meaningful information from the vital signs. Afterwards, we added fully connected hidden layers between the LSTM output and the classification layer, such that the output of these hidden layers could be used in the multimodal model. We explored three options to increase the performance of the Per-Vitals model: multi-layer LSTM, target replication and a parallel fully convolutional network (FCN). The results of this exploration are discussed in Appendix A.

For the Per-Med model we used binary encoding for the per-operative medication data[2]. This data contained only 17 features, so contrary to the Pre-Static model it was not necessary to introduce a hidden layer for the purpose of dimension reduction. This means that the Per-Med model is just an input representation containing 17 features.

[2]In preliminary experiments we investigated ordinal and temporal encoding for the per-operative medication data, but did not find a difference

C. Multimodal model training

We fused the unimodal models together to form three multimodal models: Pre-All, Per-All and Pre+Per-All (see Table IV). These multimodal models have the same architecture, see Figure 1 and only differ in which modalities were used. We combined the pre-classification layer outputs from all unimodal models and fed it to a classifier. We chose the unimodal model architectures such, that the combination of features can be easily achieved with a concatenate layer; Ideally, each modality should supply the same number of features, such that the relative contribution can be fairly assessed at a later stage.

V. Experimental Setup

A. Data set and evaluation

We split our data set in three parts; a training (50%) validation (25%) and test set (25%), where we used stratification to ensure a similar number of positive cases in each set. Models were optimised for maximum validation AUC, which was computed after every training epoch.

We set the maximum number of epoch to 100 and used early stopping with a patience of 10, where we halved the learning rate, if there was no improvement for 5 epochs. We tuned the initial learning rate, by experimenting with a range between 10^{-2} and 10^{-5}. We used the Adam optimizer and a batch size of 32.

All models had the same classification layer for mortality prediction, which contained a single neuron with the sigmoid activation function. This function ensured the output was between 0 and 1, hence it was treated as a probability. In case of the multimodal models we also added a classification layer with 4 neurons for the prediction of the other complications, one for each complication. This layer also used the sigmoid activation function, because we required a separate probability prediction for each complication.

During training our models were tasked with minimizing the weighted binary cross entropy loss. For complication prediction loss, which is a multilabel classification task, we took the average of the individual weighted binary cross entropy losses. Weights were computed according to Equation 2, where c_i is the weight of class i, N_{total} is the total number of cases and N_{c_i} is the number of cases with class c_i [28].

$$c_i = \frac{N_{total}}{2 \cdot N_{c_i}} \tag{2}$$

Besides AUC, we also computed recall, precision and F1-score for the mortality prediction to evaluate our models. We trained each model 5 times with different initial weights to measure variability between training runs.

We used the models developed by Yenidogan et al. [12] as a baseline for evaluating the performance of our models. We refer to these baseline models with similar abbreviations as introduced in Table IV: Y-Static for their Pre-Static model, Y-HipImg for their Pre-HipImg model, Y-ChestImg for their Pre-ChestImg model and Y-All for their Pre-All model.

In the subsequent sections we discuss model specifics and design choices, starting with the unimodal models, followed by the multimodal models and last explainability.

B. Unimodal models

We start with the Pre-Static model, for which we used one fully connected hidden layer between the input and output layer[3]. Initially, we used the regular rectified linear activation function (ReLu), however performance suffered from the "dying ReLu" problem [29]. To overcome this problem we employed the leaky-ReLu activation function and for consistency fully connected layers in all our models, except for the output layers, used the leaky-ReLu activation function.

We restricted our Pre-HipImg and Pre-ChestImg models to have the same CNN architecture (see Section IV), furthermore we preferred smaller networks given our small data set. Therefore, we chose a fully trained ResNet50 for the image classification tasks, where we used the pre-trained weights from ImageNet. Note that we trained a separate model for both image modalities.

During training we augmented training images with random shift (0.2), shear (0.2), zoom (0.2) and rotation (20°) to mimic a more diverse training set. Also, we used bicubic interpolation to fit images to the 224x224 shape required for ResNet50. We added two fully connected layers with 256 and 16 neurons before the classification layer. The image models were trained with a relatively small learning rate of 10^{-5}, because higher values led to very low precision ($<$0.01). To prevent overfitting we added a dropout of 0.3 between fully connected layers and a L2 regularization factor of 10^{-3}.

Our Per-Vitals model contain one bidirectional LSTM layer to extract features from the vital signs. During preliminary experiments we varied the number of units between 64 and 256 and found that a layer with 2x128 units worked best. Data was passed through a masking layer before the LSTM layer, which caused time steps with any missing data to be completely skipped. For the Per-Vitals model we set the learning rate to $5 \cdot 10^{-4}$ and appointed a dropout of 0.5 for the LSTM units. The LSTM layer is followed by two hidden layers with 128 and 16 neurons, and a dropout rate of 0.3.

C. Multimodal model training

We fused the unimodal models together by concatenating the pre-classification layers, where each modality contributed 16 features, except for the per-operative medication modality, which contributed 17 features. Thus, the concatenate layer of the Pre+Per-All model consists of 81 features, which is followed by one fully connected layer with 64 neurons. Furthermore, we used the pre-trained weights of the unimodal models, that were trained to predict mortality. Last, we added two classification layers, where the first contains 1 neuron for mortality prediction and the second layer consists of 4 neurons to predict the other complications. We found a complication loss weight (See Section IV) of 0.2

[3]We found no difference in performance compared to deeper networks during preliminary experiments

gave a good balance between complication and mortality prediction performance.

At first, we froze the weights of the pre-trained unimodal models, while we experimented with the hidden layers (after concatenation) of the multimodal models. At a later stage, we found that unfreezing the weights of the Pre-Static and Per-Vitals models increased performance, while we found no difference for the image models. Therefore, we trained our multimodal models with unfrozen weights for the Pre-Static and Per-Vitals models, and frozen weights for the Pre-HipImg and Pre-ChestImg models.

We trained the multimodal models with a learning rate of $5 \cdot 10^{-3}$. We used the same model architecture to perform the modality ablation tests leading to the Pre-All and Per-All models.

D. Explainability

We improved global interpretability of the Pre+Per model by computing the relative importance of each modality. To achieve this we iterated through all cases in the validation and test set and computed the Shapley value for each feature [24]. The Shapley values were calculated, such that for each patient it sums up to the complementary predicted mortality probability. Equation 3 describes this summation, where $f(x)$ is the predicted mortality for patient x, M is the number of features and ϕ_i is the Shapley value for feature i. The null prediction is denoted by ϕ_0, which is the average predicted mortality probability within the training set.

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i(x) \tag{3}$$

Equation 4 describes how we computed the importance of a single feature Φ_i across all samples in the test set with size N.

$$\Phi_i = \frac{1}{N} \sum_{x=1}^{N} |\phi_i(x)| \tag{4}$$

1) Relative importance: For each modality we summed up the respective importances of its features and divided this sum by the total sum of all feature importances, which yielded us the relative importance of each modality.

2) Global explanation: We used the computed Shapley values to generate a beeswarm plot for the Pre+Per-All model on the test set, which shows the top 20 most important features in descending order. The plot shows how the value of a feature impacts the model output, where a positive impact means higher mortality.

3) Single neuron explanation: We found features from the Pre-Static model to be most important for prediction, so we further investigated these features. We repeated the same procedure to compute the Shapey values, but in this case $f(x)$ (Equation 3) was the output of a single neuron from the Pre-Static model, which we explained with all 76 pre-operative static features. This provided us insight in how specific pre-operative static features affect the prediction.

4) Local explanation: Last, we computed a local explanation by generating a waterfall plot for a single positive case. This plot shows how the model builds up its prediction from the initial starting point ϕ_0.

VI. Results

A. Model performance

Table V presents the mortality prediction performance for all unimodal and multimodal models, additionally we included the results of the original paper for comparison.

*1) **Does the inclusion of per-operative features improve 30-day mortality prediction?**:* The Per-All model achieves decent performance on the training set, but scores poorly on the test set. The same thing holds for the Per-Vitals model, while the Per-Med model scores poorly on both data splits. The discrepancy in performance between the training and test set could indicate overfitting, still the decent performance of the Per-All model on the training set suggests that meaningful features can be extracted from the per-operative data. However, direct comparison of all metrics regarding the test set for the Pre-All and Pre+Per-All models implies that the inclusion of per-operative data does not improve 30-day mortality prediction.

*2) **To what extent can complications be predicted by the Pre+Per-All model?**:* Table VI shows how well the final multimodal model can predict complications, where 30-day mortality is the most severe complication. Only for heart failure prediction the model achieves reasonable performance on the test set, while other complications in the test set remain difficult to predict. We experimented with a weight for the complication loss ranging from 0 to 1 and observed strong improvement up until 0.1-0.2, but for higher values the performance stagnated. Yet, higher complication loss weight was not found to negatively affect mortality prediction.

*3) **How do our multimodal models compare to state-of-the-art?**:* Our unimodal image models (Pre-HipImg and Pre-ChestImg) score worse compared to the image models (Y-HipImg and Y-ChestImg) developed by Yenidogan et al. [12]. We foresaw some loss in

TABLE V
AVERAGED PERFORMANCE OVER 5 RUNS OF OUR MODELS COMPARED TO PRECEDING RESEARCH. THRESHOLD FOR RECALL AND PRECISION IS 0.1 FOR MODELS BY YENDIDOGAN ET AL. AND 0.5 FOR OUR MODELS. N.A.: VALUES NOT REPORTED IN ORIGINAL PAPER

	AUC		**Recall**		**Precision**		**F1-score**	
Yenidogan et al.	Train	Test	Train	Test	Train	Test	Train	Test
Y-All	n.a.	0.79	n.a.	0.71	n.a.	0.17	n.a.	0.28
Y-Static	n.a.	0.73	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Y-HipImg	n.a.	0.67	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Y-ChestImg	n.a.	0.70	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Our models								
Pre-All	0.93 (0.02)	0.75 (0.01)	0.96 (0.04)	0.66 (0.05)	0.22 (0.01)	0.16 (0.01)	0.36 (0.02)	0.25 (0.01)
Pre-Static	0.89 (0.02)	0.76 (0.01)	0.90 (0.03)	0.68 (0.06)	0.21 (0.02)	0.17 (0.01)	0.35 (0.02)	0.27 (0.01)
Pre-HipImg	0.52 (0.05)	0.53 (0.02)	0.32 (0.18)	0.25 (0.14)	0.08 (0.03)	0.12 (0.07)	0.13 (0.05)	0.12 (0.04)
Pre-ChestImg	0.58 (0.02)	0.54 (0.05)	0.34 (0.16)	0.21 (0.13)	0.11 (0.01)	0.08 (0.03)	0.16 (0.03)	0.12 (0.04)
Per-All	0.77 (0.10)	0.56 (0.03)	0.70 (0.16)	0.34 (0.07)	0.18 (0.05)	0.09 (0.01)	0.28 (0.07)	0.15 (0.02)
Per-Vitals	0.67 (0.03)	0.57 (0.05)	0.50 (0.07)	0.38 (0.10)	0.13 (0.01)	0.10 (0.02)	0.20 (0.01)	0.16 (0.03)
Per-Med	0.49 (0.03)	0.53 (0.06)	0.40 (0.05)	0.47 (0.17)	0.07 (0.01)	0.08 (0.02)	0.12 (0.01)	0.14 (0.04)
Pre+Per-All	0.93 (0.01)	0.75 (0.01)	0.96 (0.03)	0.65 (0.13)	0.23 (0.02)	0.15 (0.02)	0.37 (0.03)	0.25 (0.03)

TABLE VI
COMPLICATION PREDICTION PERFORMANCE OF THE MULTIMODAL MODEL

Complication	**Training AUC**	**Test AUC**
Mortality	0.91	0.75
Heart failure	0.70	0.67
Pneumonia	0.69	0.60
Anaemia	0.67	0.57
Delirium	0.61	0.51

TABLE VII
RELATIVE IMPORTANCE FOR MORTALITY PREDICTION OF INPUT MODALITIES

Modality	**Relative importance**
Pre-operative	**87.4%**
Static patient data	51.7%
Hip image	20.9%
Chest image	14.8%
Per-operative	**12.6%**
Vitals	2.4%
Medication	10.2%

performance due to the smaller data set and smaller CNN architecture, yet the difference is larger than expected.

On the other hand, our Pre-Static model performs on par with state-of-the art and slightly outperforms the Y-Static model [9], [10], [12]. We observe strong indication for overfitting given the large discrepancy in performance between the train and test set for the Pre-Static model. We did take measures (dropout and L2 regularization) to prevent overfitting, however stronger measures might need to be taken, or redundant features could be removed.

B. Explainability

Table VII shows the relative importance of each individual modality as well as the pre-operative and per-operative groups. The features extracted from the per-operative modalities barely contribute to mortality prediction, while features from the static patient data contributes the most. Especially, the per-operative vital signs appear to be of no value to the multimodal model with a relative importance of only 2.4%, the medication data are a little more valuable with a relative importance of 10.2%. For the medication data this is more than expected, given the very poor performance of the Per-Med model. Therefore, there might be some predictive value in the interaction between medication administration during surgery and pre-operative factors.

Figure 2 shows two beeswarm plots and one waterfall plot providing more insight in the importance of specific features. Starting with Figure 2a, which shows how the top 20 features impact the model decision. Features extracted from the static patient data dominate with 10 out of 16 features present in the top 20. Furthermore, the image modalities are also well represented, with 4 features for both. The remaining two features come from

the per-operative medication modality.

These features are vaguely defined and require further inspection, if we want to extract knowledge that is understandable from a clinical perspective. Figure 2b shows the top 10 pre-operative static features contributing to the *Static-2* neuron. Importantly, from Figure 2a we learn low *Static-2* values positively affect mortality prediction. This means that low values in Figure 2b mean higher mortality, for example older patients have a higher predicted mortality.

Last, Figure 2c shows the most important features contributing towards a prediction for a single patient. This patient did not survive the first 30-days after surgery, where the model predicted a value of 0.787. Positive values (red) indicate the value of that feature increased the mortality probability prediction, while negative values (blue) indicate the value of that feature decreased mortality probability prediction. The values of the features are shown in gray left to the feature name. We observe that the *Static-2* is most important in this case with a contribution of 0.09, also features from the hip and medication modalities are important for this case. For clinical applicability plots like Figure 2c could be crucial, especially if the features are less vaguely defined.

VII. Discussion

The addition of per-operative data did not yield a significant improvement in performance compared to an earlier study by Yenidogan et al. [12], however our data set is slightly smaller and still achieves similar performance. Moreover, due to the combination of class imbalance and a small data set, the AUC is very dependent on a few randomly selected cases. Specifically, there are only 33 positive cases in the test and validation set and chances are that those do not generalise the elderly hip fracture population well.

We employed the sample normalisation strategy for the vital signs, yet this does eliminate the ability for the model to take into account cross case differences in average vital sign value. Multi-resolution normalisation has been proposed as a solution to the latter [8]. This normalisation still uses sample normalisation for the per-operative vital signs, however the mean and standard deviation are z-transformed across all cases and added to the pre-operative static variables.

The models that include static patient data or vital signs are overfitting the data. The addition of dropout layers and L2-regularization did not solve this problem, therefore a different approach is required. It has been shown that reducing the number of pre-operative static features based on their importance can prevent overfitting [10]. We could use the Shapley values to iteratively select the most important feature, up until we reach a certain subset size. Additionally, if we prevent the strong overfitting on the pre-operative static data, this could incentivize the multimodal models to focus more on the other modalities for information.

Minimizing binary cross entropy loss does not directly mean that AUC is maximized and this discrepancy could lead to inferior results [30]. Different loss functions have been proposed that directly incorporate AUC in the loss function, consequently improving model training.

Our multimodal model is not robust against missing data, only the per-operative medication data is allowed to be missing. In clinical practice this would mean patients are excluded, if they are missing pre-operative hip and thorax images or per-operative vitals. We impute the pre-operative static patient data, so having some missing values there does not lead to exclusion. Therefore, for clinical applicability future models should be robust against missing modalities, in order to avoid patient exclusion.

Our fusion method could be described as mid fusion, because we used the pre-classification layer of each unimodal model, however we did do dimension reduction before concatenation. This method ensured each modality contributed the same number of features to the classification layer, however this might not be optimal for post-operative complication prediction. Future work could include a deeper investigation of fusion methods, like late fusion and bottleneck fusion. Bottleneck fusion restricts cross-modal information flow by using a very limited amount of neurons for information to pass through. The idea is that the model is forced to condense the most important cross-modal features leading to better performance at negligible computational cost [19].

The aim to immediately predict mortality might have been overly ambitious and starting with predicting if any complication occurs within 30-days could improve assessment of the data set predictive capabilities. Coincidentally, this resolves the class imbalance issue, because in our data set 49% of the patients experience at least one complication within 30 days after surgery. Furthermore, as an intermediate step complications could be grouped by severity or cases could be scored on a scale from no complication to mortality. Clinically this could help determine, whether a patient is at risk after surgery and requires more attention.

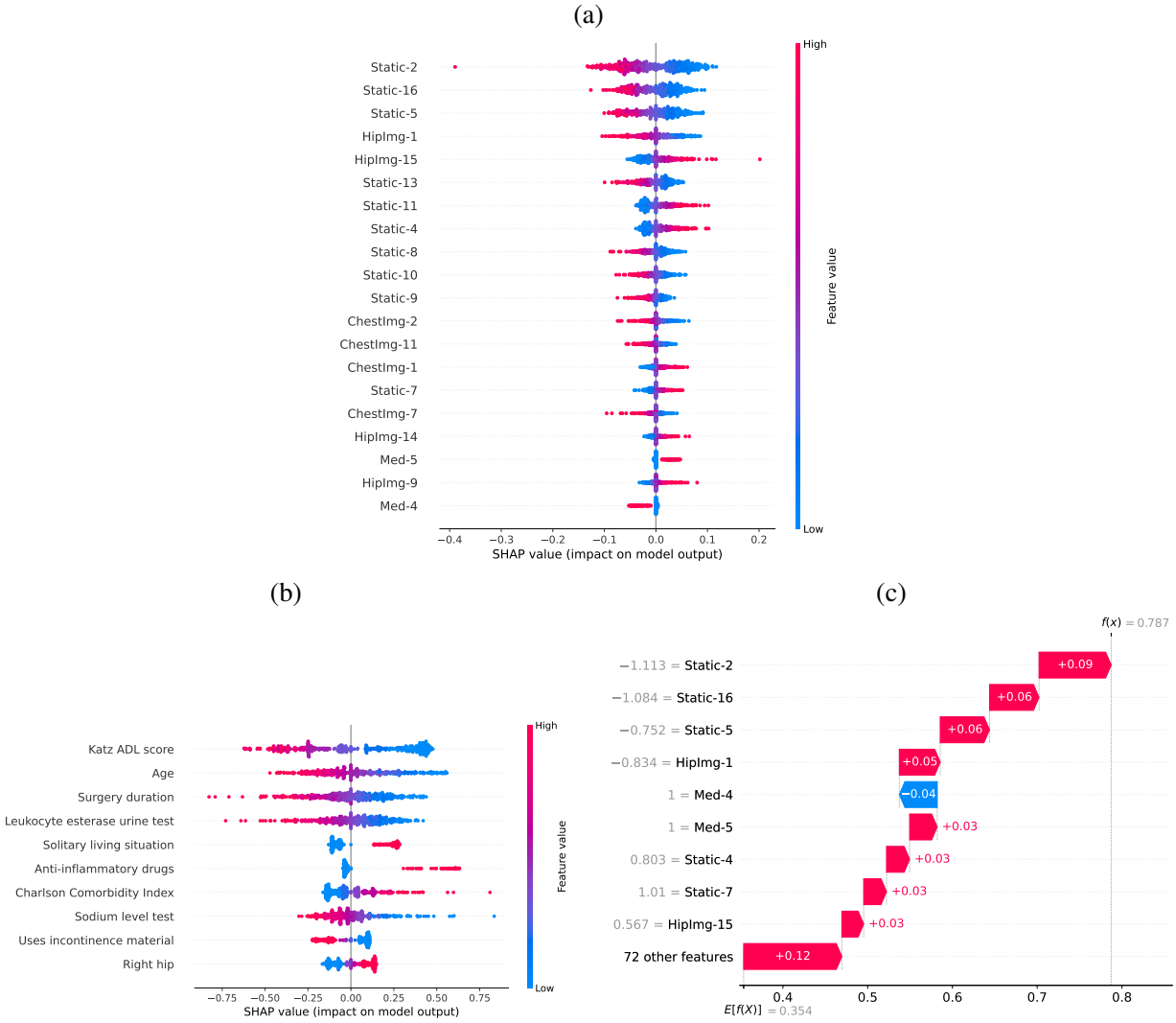

Fig. 2. (a) Beeswarm plot of the 20 most important values for the full multimodal model, the number behind each modality specifies a specific neuron. On the horizontal axis is the Shapley value, where positive means the feature value increased mortality prediction and negative means it lowered mortality prediction. The color refers to the feature value itself: red if high and blue if low. (b) Beeswarm plot for the *Static-2* neuron. (c) Waterfall plot showing how features contribute to the prediction for a specific patient. Here a red color means the feature value increased mortality prediction and blue means it decreased mortality prediction. The actual feature value is shown in gray left to the feature name.

VIII. Conclusion

We conclude that per-operative data in addition to pre-operative data does not significantly improve 30-day mortality prediction, when compared to an earlier study. However, our data set is relatively small and a bigger data set should give a more definitive answer, especially introducing more deceased patients should improve model performance. The other complications remain difficult to predict with the exception of heart failure, for which our model achieves reasonable performance. Further investigation confirmed that pre-operative features are most important for mortality prediction, while per-operative features contribute little with the exception of a few per-operative medications. We used Shapley values to explain model predictions, in order to make our multimodal model more understandable for clinical practitioners. Furthermore, we encourage future work to prevent overfitting, by using these Shapley values to reduce the number of pre-operative static features. Alternatively, the prediction task could be altered to prediction of any complication or complication severity.

REFERENCES

[1] K. A. Hartholt, C. Oudshoorn, S. M. Zielinski, P. T. P. W. Burgers, M. J. M. Panneman, E. F. v. Beeck, P. Patka, and T. J. M. v. d. Cammen, "The Epidemic of Hip Fractures: Are We on the Right Track?" *PLOS ONE*, vol. 6, no. 7, p. e22227, Jul. 2011, publisher: Public Library of Science.

[2] B. Gullberg, O. Johnell, and J. Kanis, "World-wide Projections for Hip Fracture," *Osteoporosis International*, vol. 7, no. 5, pp. 407–413, Sep. 1997.

[3] F. Hu, C. Jiang, J. Shen, P. Tang, and Y. Wang, "Preoperative predictors for mortality following hip fracture surgery: A systematic review and meta-analysis," *Injury*, vol. 43, no. 6, pp. 676–685, Jun. 2012.

[4] H. J. Jones and L. de Cossart, "Risk scoring in surgical patients," *The British Journal of Surgery*, vol. 86, no. 2, pp. 149–157, Feb. 1999.

[5] G. Briganti and O. Le Moine, "Artificial Intelligence in Medicine: Today and Tomorrow," *Frontiers in Medicine*, vol. 7, 2020.

[6] B. Xue, D. Li, C. Lu, C. R. King, T. Wildes, M. S. Avidan, T. Kannampallil, and J. Abraham, "Use of Machine Learning to Develop and Evaluate Models Using Preoperative and Intraoperative Data to Identify Risks of Postoperative Complications," *JAMA Network Open*, vol. 4, no. 3, pp. e212 240–e212 240, Mar. 2021.

[7] C. K. Lee, I. Hofer, E. Gabel, P. Baldi, and M. Cannesson, "Development and Validation of a Deep Neural Network Model for Prediction of Postoperative In-hospital Mortality," *Anesthesiology*, vol. 129, no. 4, pp. 649–662, Oct. 2018.

[8] B. A. Fritz, Z. Cui, M. Zhang, Y. He, Y. Chen, A. Kronzer, A. Ben Abdallah, C. R. King, and M. S. Avidan, "Deep-learning model for predicting 30-day postoperative mortality," *British Journal of Anaesthesia*, vol. 123, no. 5, pp. 688–695, Nov. 2019.

[9] 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*, vol. 47, no. 10, pp. 2138–2143, Oct. 2016.

[10] Y. Cao, M. P. Forssten, A. Mohammad Ismail, T. Borg, I. Ioannidis, S. Montgomery, and S. Mohseni, "Predictive Values of Preoperative Characteristics for 30-Day Mortality in Traumatic Hip Fracture Patients," *Journal of Personalized Medicine*, vol. 11, no. 5, p. 353, Apr. 2021.

[11] J. Karres, N. Kieviet, J.-P. Eerenberg, and B. C. Vrouenraets, "Predicting Early Mortality After Hip Fracture Surgery: The Hip Fracture Estimator of Mortality Amsterdam," *Journal of Orthopaedic Trauma*, vol. 32, no. 1, pp. 27–33, Jan. 2018.

[12] B. Yenidogan, S. Pathak, J. Geerdink, J. H. Hegeman, and M. van Keulen, "Multimodal Machine Learning for 30-Days Post-Operative Mortality Prediction of Elderly Hip Fracture Patients," in *2021 International Conference on Data Mining Workshops (ICDMW)*, Dec. 2021, pp. 508–516, iSSN: 2375-9259.

[13] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use," in *Proceedings of the 4th Machine Learning for Healthcare Conference*. PMLR, Oct. 2019, pp. 359–380, iSSN: 2640-3498.

[14] U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly, "Explainable AI in Healthcare," in *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, Jun. 2020, pp. 1–2.

[15] J.-W. Perng, I.-H. Kao, C.-T. Kung, S.-C. Hung, Y.-H. Lai, and C.-M. Su, "Mortality Prediction of Septic Patients in the Emergency Department Based on Machine Learning," *Journal of Clinical Medicine*, vol. 8, no. 11, p. 1906, Nov. 2019, number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[16] A. K. Gowd, A. Agarwalla, N. H. Amin, A. A. Romeo, G. P. Nicholson, N. N. Verma, and J. N. Liu, "Construct validation of machine learning in the prediction of short-term postoperative complications following total shoulder arthroplasty," *Journal of Shoulder and Elbow Surgery*, vol. 28, no. 12, pp. e410–e421, Dec. 2019, publisher: Elsevier.

[17] A. J. Schoenfeld, H. V. Le, Y. Marjoua, D. A. Leonard, P. J. Belmont, C. M. Bono, and M. B. Harris, "Assessing the utility of a clinical prediction score regarding 30-day morbidity and mortality following metastatic spinal surgery: the New England Spinal Metastasis Score (NESMS)," *The Spine Journal*, vol. 16, no. 4, pp. 482–490, Apr. 2016.

[18] L. de Munter, S. Polinder, K. W. W. Lansink, M. C. Cnossen, E. W. Steyerberg, and M. A. C. de Jongh, "Mortality prediction models in the general trauma population: A systematic review," *Injury*, vol. 48, no. 2, pp. 221–229, Feb. 2017.

[19] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention Bottlenecks for Multimodal Fusion," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 14 200–14 213.

[20] A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," arXiv, Tech. Rep. arXiv:2006.11371, Jun. 2020, arXiv:2006.11371 [cs] type: article.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144.

[22] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*. PMLR, 2017, pp. 3145–3153.

[23] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLOS ONE*, vol. 10, no. 7, p. e0130140, Jul. 2015, publisher: Public Library of Science.

[24] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[25] F. Karim, S. Majumdar, and H. Darabi, "Insights Into LSTM Fully Convolutional Networks for Time Series Classification," *IEEE Access*, vol. 7, pp. 67 718–67 725, 2019, conference Name: IEEE Access.

[26] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Annals of Translational Medicine*, vol. 8, no. 11, pp. 713–713, Jun. 2020, number: 11 Publisher: AME Publishing Company.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, iSSN: 1063-6919.

[28] G. King and L. Zeng, "Logistic Regression in Rare Events

Data," *Political Analysis*, vol. 9, no. 2, pp. 137–163, 2001, publisher: Cambridge University Press.

[29] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," Feb. 2019, arXiv:1803.08375 [cs, stat].

[30] E. E. Eban, M. Schain, A. Mackey, A. Gordon, R. A. Saurous, and G. Elidan, "Scalable Learning of Non-Decomposable Objectives," *arXiv:1608.04802 [cs, stat]*, Mar. 2017, arXiv: 1608.04802.

[31] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to Diagnose with LSTM Recurrent Neural Networks," arXiv, Tech. Rep. arXiv:1511.03677, Mar. 2017, arXiv:1511.03677 [cs] type: article.

[32] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate LSTM-FCNs for time series classification," *Neural Networks*, vol. 116, pp. 237–245, Aug. 2019.

APPENDIX A
PER-VITALS OPTIMIZATION

In order to improve on the single layer Per-Vitals model we explored three options: A multi-layer LSTM, target replication [31] and a parallel convolutional network (LSTM-FCN) [25].

A multi-layer LSTM might be able to learn more intricate transition rules, where each LSTM layer returns a sequence to be processed by the next LSTM layer, except for the last LSTM layer. Target replication has been proposed to help the model pass information over longer sequences, by generating a prediction at each time step and compute the complementary loss [31]. Where the final target replication loss is the weighted sum of the final prediction loss and the average of all pre-final prediction losses. The last option is to add a parallel convolutional network to create a LSTM-FCN, for which it has been shown to achieve higher performance on a broad range of datasets [25]. We use the one-layer LSTM model as a baseline for comparing the performance of the three mentioned options.

The implementation of the multi-layer LSTM is straightforward, since we simply insert additional bidirectional LSTM layers containing 2x128 units. For target replication we introduce a parameter α, which denotes the weight for the average pre-final prediction loss. Consequently, the weight for the loss of the final prediction is $1 - \alpha$, as we require the weights to sum up to one, so loss magnitude remains similar. The LSTM-FCN contains three convolutional layers with a number of filters (128, 246 and 128) and a kernel size of 8,5 and 3 respectively [32]. These convolutional layers run parallel to the baseline LSTM and outputs are concatenated before the classification layer.

Table VIII shows the results for the three options compared to the one-layer LSTM baseline, where we report the average AUC on the validation set over five runs. The results indicate that none of the options improve performance compared to the baseline.

We do note that the target replication for $\alpha = 0$ is rather low, while it should be close to the one-layer LSTM. Essentially there is no difference, because in this case the loss function only takes the final prediction into account. Therefore, we can not exclude a possible implementation problem and we encourage further research using target replication. Incorporating earlier predictions in the loss function could help focus the model on defining time steps within a sequence, which is promising in a medical context. Specifically, it is fair to assume patients enter and leave the operation room in stable condition and events that are critical to outcome prediction tend to happen in the middle of surgery. We think target replication could help to divert the attention of the model to these events during training and correctly predict post surgery complications

TABLE VIII
RESULTS OF DIFFERENT OPTIONS FOR THE PER-V MODEL, PERFORMANCE IS THE AVERAGE OVER 5 RUNS

Model	Mean validation AUC
One-layer LSTM	0.754
Multi-layer LSTM	
2 layer	0.748
3 layer	0.731
Target replication	
$\alpha = 0$	0.645
$\alpha = 0.5$	0.649
LSTM-FCN	0.712

Furthermore, it is curious that the LSTM-FCN performs worse, where we expected it to achieve at least similar performance to the one-layer LSTM. If the convolutional layers are unable to extract meaningful values, then the model should be able to replicate the baseline performance, by only considering the features extracted by the LSTM.

In conclusion, we found that none of the three options improved 30-mortality prediction performance compared to the one-layer LSTM. Therefore, we used the one-layer LSTM for our multimodal experiments in our study, where we found that features extracted form the per-operative vitals signs barely contributed to the prediction. The reason we found none of the three options improved performance might as well be due to the poor predictive power of the per-operative vital signs and not due to the options.

APPENDIX B
VITALS EXAMPLE

Figure 3 and 4 show the per-operative vitals signs before and after pre-processing, respectively.

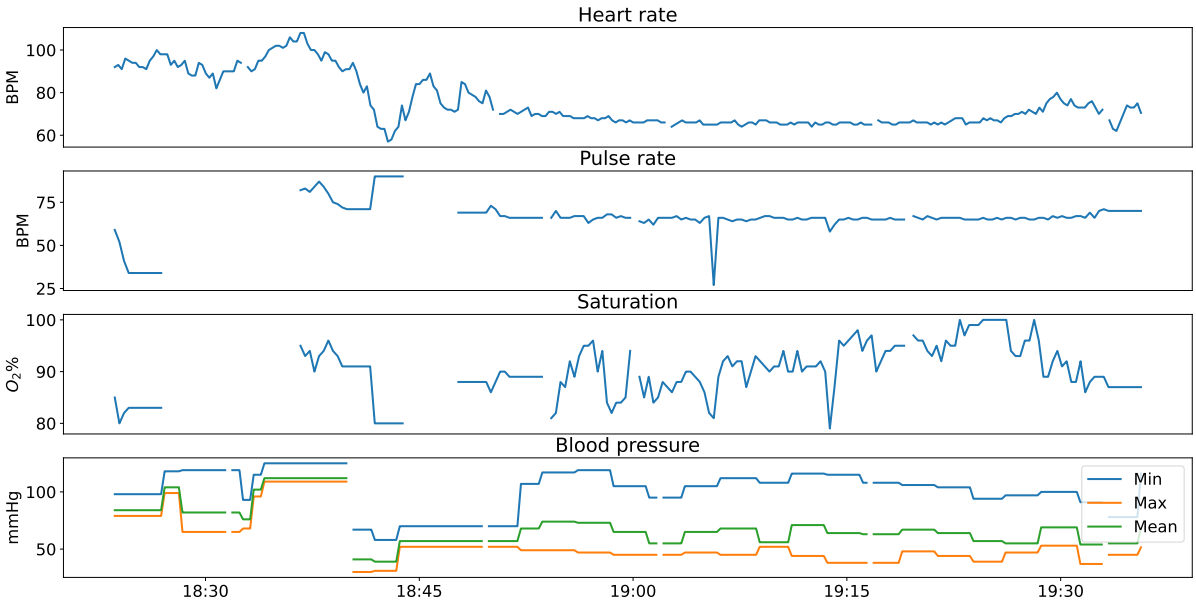

Fig. 3. Unimputed example of the vitals signs of a patient

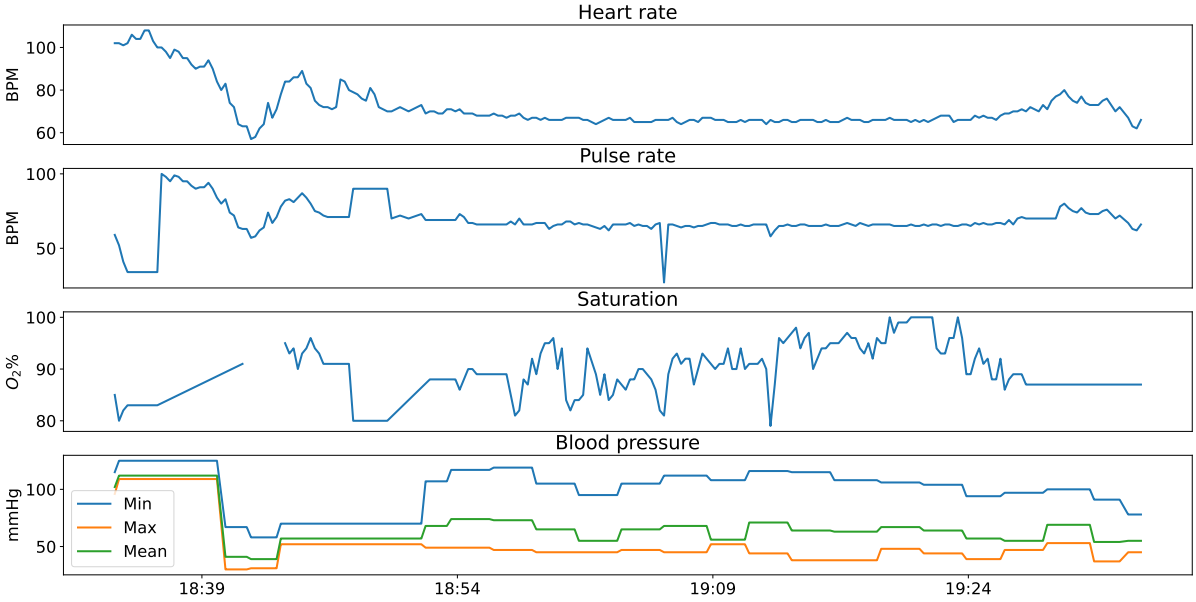

Fig. 4. Imputed example of the vitals signs of a patient

Part II

Appendix

Chapter 1

Dataset analysis

In this chapter we first take a closer look at the dataset in Section 1.1, this includes all the different modalities. Afterwards, we conduct a statistical analysis in Section 2 to assess the predictive power of the dataset.

1.1 General inclusion criteria

We used the following inclusion criteria to gather the data from a database maintained by ZGT:

- Patients were at least 71 years old at the time of surgery
- Patients had surgery for a hip fracture
- Patients were admitted between January 1 2013 and July 21 2021

Figure 1.1 shows an overview of the three main data types accompanied by their modalities. First, we gathered pre-operative data containing general information about the patient and x-ray images from the hip and chest. Second, we collected per-operative data containing information about medication and vitals (e.g. heart rate) during surgery. Last, we have post-operative data about complications within 30 days with mortality being the most severe. In Sections 1.1.1-1.1.3, we delve deeper into each of the data modalities, which includes pre-processing steps and a missing analysis.

1.1.1 Pre-operative data

Using the criteria stated earlier, we selected 1966 unique cases concerning 1911 patients. The plausibility to have multiple hip surgeries within the chosen time span, explains the disparity between the number of cases and patients. Besides a slightly

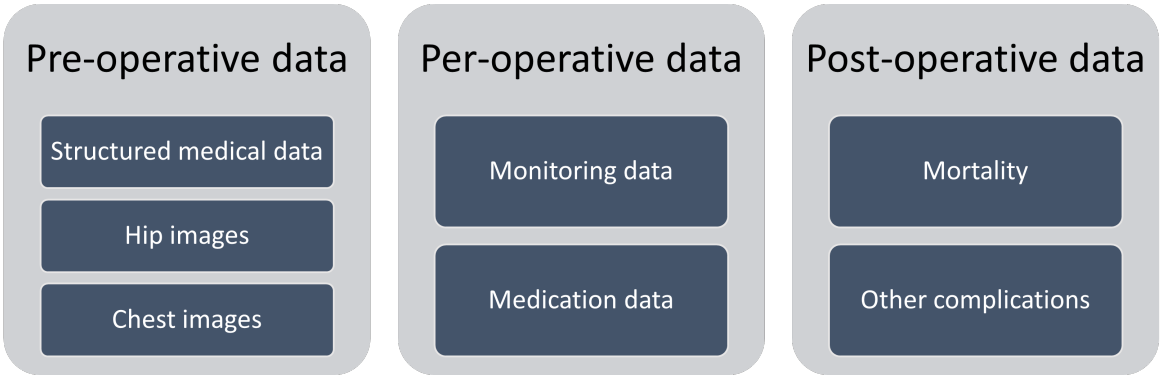

Figure 1.1: Overview of data types

shifted and shorter time span the dataset is mostly the same with respect to previous research. We shifted the time frame to include recent cases and excluded cases before 2013, because these had a relatively high amount of missing data. Similarly to earlier research, we merged the data with another data set, which reduced the number of missing values. We divide the pre-operative data in 8 categories: demographics, daily living activities, nutrition, surgery information, lab results, medication, commorbidities, and image data. All available features grouped by category are shown in Table 1.1 with the exception of image data, which we discuss in Section 1.1.1.

We checked all binary and categorical features for class imbalance, because extreme cases could cause predictions to be unreliable. Additionally, removing redundant variables decreases the complexity of the model. We identified the smallest class for each binary/categorical feature, as well as the number of unique classes.

For binary features, where the number of classes is two, we required the smallest class to occur in at least 100 cases, which resulted in the exclusion of ten features. Only 18 patients took immunosuppressants (L04) and only 45 patients got a positive result from the IRAI (blood related) lab test. The remaining eight excluded binary features concerned comorbidities, which were: lymphoma (2), leukemia (5), peptic ulcer disease (14), liver disease (19), prior myocardial infarction (83). The number in between brackets corresponds to the size of the smallest class.

For categorical variables exclusion is less trivial, because we could combine subclasses to obtain acceptable class sizes. A certain type of fracture, a subtrochanteric fracture, was recorded for only 41 patients compared to 896 and 744 recordings for medial column fractures and pertrochanteric fractures, respectively. Also, for 286 patients we have no information about the type of fracture, but due to the limited importance of this feature in earlier research, no further action was taken. In eleven cases the type of surgery was documented as “other”, since this provides no information, we treat these values as missing. Also, we treated a small class, with only

7 cases concerning the patient's living situation, as missing.

We explored the data to get more insight in the amount of missing data, but also to gather some general information. The average age is 84 ($\pm$5.4) and there is an unequal distribution of gender with 1399 (71.2%) women to 567 (28.2%) men. Specifically for this dataset 7.68% of the patients did not survive the first 30 days after surgery. We only removed the "Vitamine D" column, because it was missing in 73.5% of the cases.

Image data

Prior to surgery at least two x-ray images should be available, one from the hip and one from the thorax. Although, cases commonly contain images from different directions, we selected one direction for each image type. For the hip images we selected the axial direction and for the thorax images we selected the anterior to posterior direction (front to back). Currently, 204 cases are missing a hip image and 85 cases are missing a thorax image. It should be possible to retrieve some of these by specifically looking into them, however it would have costed too much time.

1.1.2 Per-operative data

The per-operative (or intra-operative data) holds information about the patient during surgery and consists of two modalities. First, we discuss the **monitoring data** that contains information about the vital signs of the patient Second, we take a closer look at the **medication** a patient received during surgery.

Monitoring data

When a patient is connected to a anesthesia machine vital signs are regularly measured and stored, not only during surgery, but also before and after. In order to get the vital signs during surgery, we first gathered all available data from machines around the day of surgery. Then using the planned start and end times of the surgery, we specifically select the data measured during surgery.

Before we delve deeper into the data, we first check if there are cases with no monitoring data at all. The raw monitoring dataset contains 2056 unique cases, which is more compared to the 1966 cases in the pre-operative data. Only two cases from the pre-operative data were missing monitoring data, thus leaving us with 1964 recordings of vital signs.

Next, we assess the quality of the vital signs, more specifically how much data is missing. Before computing the percentage of missing data, we first enforce a regular time interval between the start and end times of the surgery. There are

Table 1.1: Available pre-operative features grouped by category

Demographics	**Daily living activities**	**Nutrition**	**Surgery information**
Age	Help with transfer from bed to chair	Malnutrition risk	Fracture type
Surgery start/end	Help with showering	Unintended weight loss	Surgery type
Falling risk	Help with dressing	Drink or tube feeding	Fracture laterality
Fall last year	Help with going to toilet	Decreased appetite	ASA score
Pre-fracture mobility	Help with eating	SNAQ score	
Living situation	Help with self-care		
Prone to delirium	Katz ADL score		
Memory problems	Incontinence material used		
Delirium in the past			
CCI score			

Lab results	**Medication (reason/effect)**	**Comorbidities**
HB	Blood thinners	Chronic pulmonary disease
HT	Vitamin D	Congestive heart failure
CRP	Polypharmacy	Peripheral vascular disease
LEUC	A02 (acid related disorders)	Cerebrovascular disease
THR	A10 (diabetes)	Dementia
BLGR	B01 (antithrombotic)	Renal disease
IRAI	B02 (antihemmorrhagics)	Rheumatological disease
ALKF	B03 (antianemic)	Cancer
GGT	C01 (cardiac therapy)	Cerebrovascular event
ASAT	C03 (diuretics)	Liver disease
ALAT	C07 (beta blockers)	Lymphoma
LDH1	C08 (calcium channel blockers)	Leukemia
UREU	C09 (renin-angiotensin system)	Peptic ulcer disease
KREA	C10 (lipid modification)	Diabetes
GFRM	L04 (immunosuppressants)	Prior myocardial infarction
NA	M01 (anti-inflammatory)	
XKA	N05 (psycholeptics)	
GLUCGLUC	R03 (airway obstruction)	

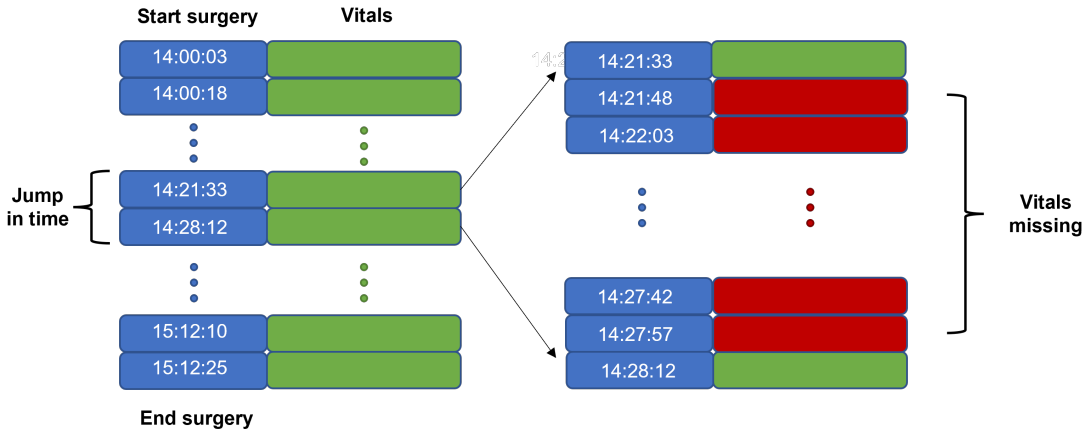

Figure 1.2: Shows how enforcing a regular time interval exposes gaps in the data. Green indicates monitoring data is available and red indicates data is not available at a certain time during surgery.

multiple reasons to do this, but first and foremost it is important for temporal machine learning that the data points are evenly spaced. Also, gaps are common in the data, meaning that for several minutes not a single vital sign was registered. Furthermore, if we do not enforce a regular time interval, then in some cases it would seem there is no data missing, because the corresponding time stamps will just jump in time. Figure 1.2 illustrates how it may seem no data is missing on the left side, however there is a gap of about five minutes with no data. After enforcing the interval, the missing data is exposed. We set the interval to 15 seconds, because the health monitoring machines of ZGT take measurements at this interval.

Another common reason for missing data is, that at the start of surgery it takes a while for all vitals to be available. Similarly, at the end surgery it takes a while for all data to become unavailable. The most probable explanation is that it takes a while to fully connect and disconnect a patient to an anesthesia machine. To counteract these start and end artifacts we only kept the data in between the first and last time all vitals were non-missing. Additionally, we linearly interpolate gaps in the monitoring data of up to five minutes, where we think gaps bigger than five minutes require further investigation and should not be interpolated.

In summary the pre-processing steps up until now are:

1. Only select data measured during surgery
2. Enforce regular time interval
3. Trim start and end of surgery
4. Interpolate gaps ($\leq$ 5 minutes)

After these pre-processing steps we can assess the quality of the data at a global as well as an individual scale. For each patient we computed the percentage of missing values for each vital. Before we judge the quality at a global scale, we first

Table 1.2: Mean percentage of the amount of missing data after pre-processing

Vital	Mean % missing
Heart rate	0.00
Pulse	0.00
Saturation	0.27
Dystolic blood pressure	0.14
Systolic blood pressure	0.14
Mean blood pressure	0.12

inspect specific cases that missed a lot of data. This led to several case specific corrections, where in two cases we shifted the date of surgery by one day, while all further corrections concerned the start and end of surgery. Big gaps (¿10min) often occur here, which, with the method of enforcing a regular time interval, results in an inflated missing percentage. Therefore, we manually altered the start and end time of surgery to fit the available data for those cases. However, for six cases the gaps are either too big or right in the middle of surgery, so we remove these cases from the dataset.

The last step done during pre-processing we combined the heart rate and pulse, since these values tend to be the same. If for example the heart rate is missing, but the pulse is available, we set the heart rate equal to the pulse. For the global assessment of quality we computed the average amount of missing data per vital, Table 1.2 shows the results. Patient specific corrections were done for all cases that still had some missing heart rate data after pre-processing, eventually lowering the mean percentage to zero. We did not perform these time expensive corrections for the other vitals, if necessary we could do this at a later stage during the study. Last, Figure 1.3 shows an example of the data after pre-processing for a patient throughout the whole surgery, this illustrates the input to the deep learning model.

Medication data

We collected all data about what medications a patient received during surgery, specifically the features are volume, medication group, dose and time of administration. Similar medications are already grouped and the raw dataset contains 103 unique medication groups. For training purposes we only select medications that were used in at least 100 distinct cases, Table 1.3 shows these medications together with their occurrence count and effect. At a later stage during this research we might leave out more medication groups, if either they show no contribution to prediction performance or a clinical expert deems them irrelevant.

Furthermore, there are multiple ways we could feed this data to the model, we

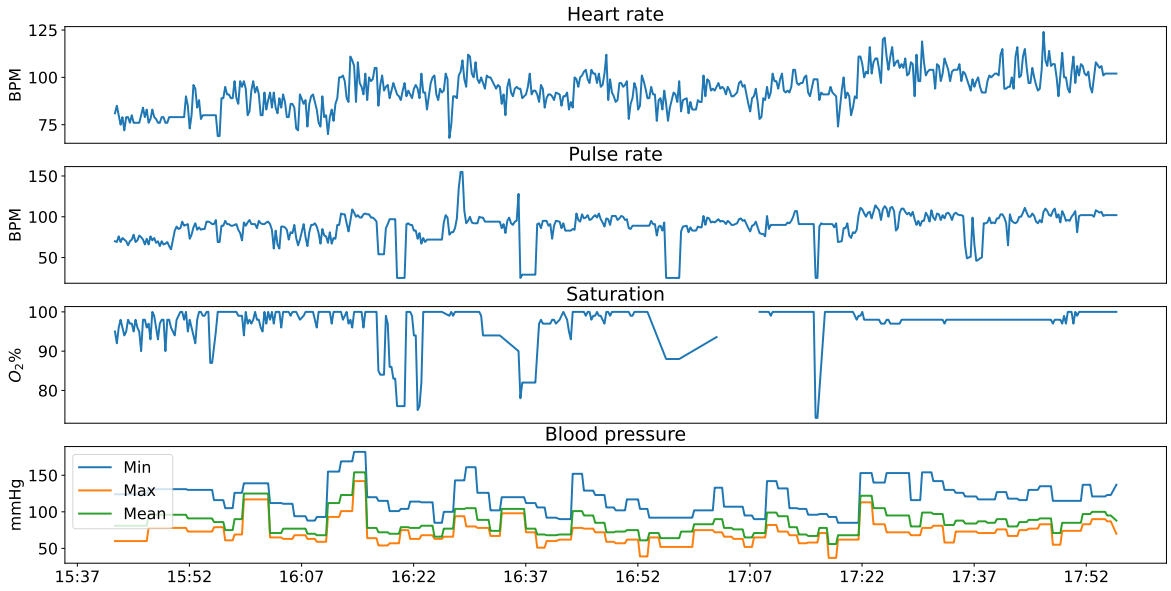

Figure 1.3: Example of the monitoring data after pre-processing of patient throughout surgery

Table 1.3: List of medications that were at least administered in 100 unique cases, also includes the general reason for usage. Percentages are with respect to a total of 1616 cases for which medication data is available.

Medication	**# unique cases**	**Effect**
Bupivacaine	649 (40.2%)	Anesthetic
Cefazolin	692 (42.8%)	Antibiotic
Dexemethasone	198 (12.3%)	Anti-inflammatory
Efedrine	445 (27.5%)	Increase blood pressure
Elektrolytes	468 (29.0%)	Minerals
Esketamine	561 (34.7%)	Anesthetic
Lidocaine	405 (25.1%)	Anesthetic
Metamizole	187 (11.6%)	Painkiller
Midazolam	475 (29.4%)	Anesthetic
Noradrenaline	887 (54.9%)	Increase blood pressure
Ondansetron	282 (17.5%)	Counter post-operative nausea
Piritramide	446 (27.6%)	Painkiller
Propofol	648 (40.1%)	Anesthetic
Rocuronium	260 (16.1%)	Muscle relaxant
Sufentanil	746 (46.2%)	Painkiller
Sugammadex	112 (6.9%)	Reverse muscle relaxant
Tranexamic acid	465 (28.8%)	Prevent blood loss

discuss three ways in order of increasing complexity. The least complex manner is in a binary format, which only indicates if a patient received a certain medication during surgery. Another way would be to transform the data into a ordinal format that indicates how often a patient received a certain medication. Finally, we could use the total amount of each medication given to a patient during surgery, however currently there is too much information missing to make this feasible. Although, it is tempting to add the data in the most complex way that is still feasible, it could also harm prediction performance. Moreover, less complex models are easier to understand.

1.1.3 Post-operative data

The post-operative data contains the prediction targets for the model with the most important being 30-day mortality. Also, we collected information about less severe post-operative complications with the goal to predict these simultaneously with mortality. Furthermore, if the predicted probability for 30-day mortality and a certain complication are both high, then that complication might be the reason for high mortality. Ideally, if we know a patient has a high risk for a certain complication during surgery, then preparations could be made preemptively to reduce the consequences. Figure 1.4 shows the incidence rate of each of the complications as a percentage of the total amount of cases in the initial dataset. To ensure that there was sufficient data available to reliably train a model to predict complications, we set a minimum incidence rate of 5.1% corresponding to 100 unique cases. This left the following four complications: delirium, anemia, pneumonia and heart failure.

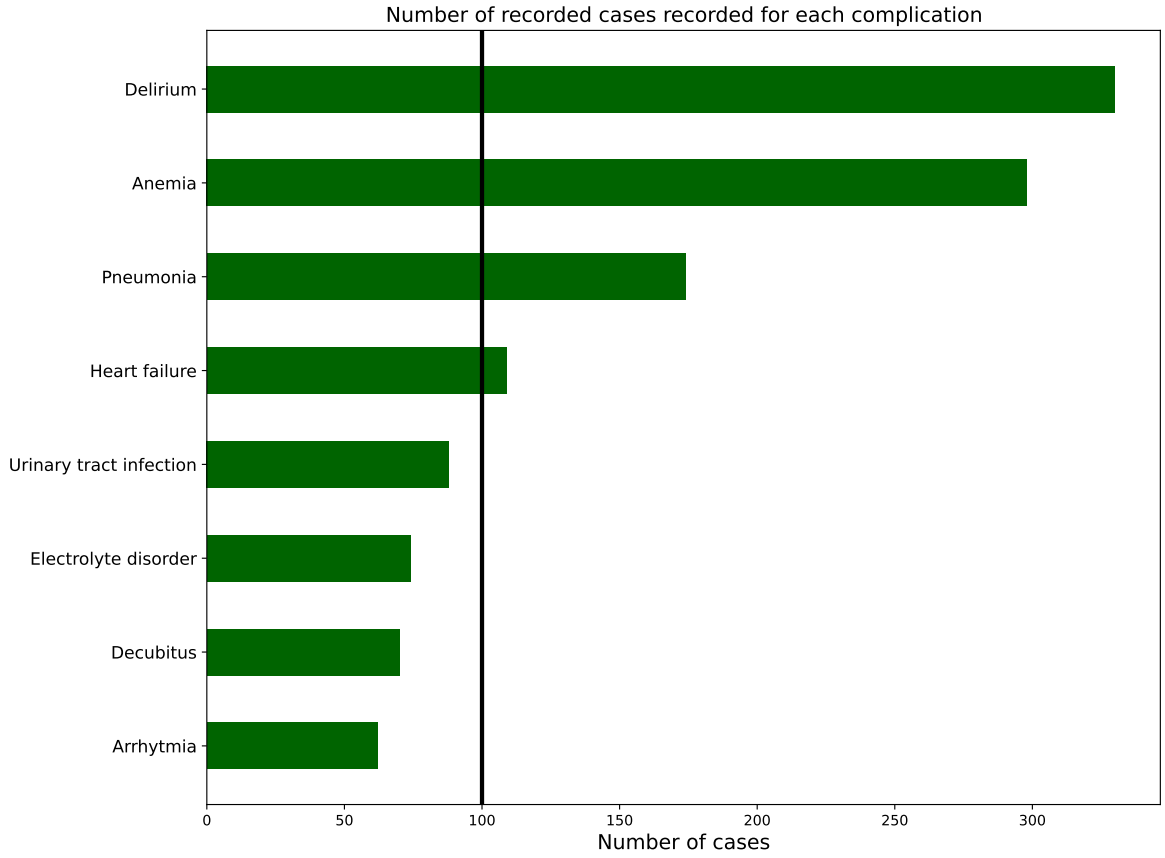

Figure 1.4: Complication occurences as a percentage of the the total amount of cases in the initial dataset (n=1966). The vertical line denotes a cutoff at 5.1%, which correspons to 100 cases.

Chapter 2

Statistical analysis

Incorporating per-operative data to predict mortality is the novelty of our study, consequently little is known about the predictive power of this data. Therefore, we conduct statistical analyses, so we can set fair expectations for the performance of our per-operative models. We conduct all tests in a two-sided fashion, although one-sided tests provide more information on how the population relate to each other, they also easily lead to incorrect conclusions and usage is generally discouraged. This section contains three subsections each examining a part of the novel data, Section 2.1 is about the effect of complications on mortality, Section 2.2 investigates the effect of the vitals on complications and Section 2.3 studies the effect of medication on complications. For each analysis, we justify our method choice and check its assumptions. Then we discuss a a single example extensively, before showing the full results.

2.1 Complications

Besides mortality we add some other complications as prediction targets for the model, therefore it is interesting to know how these complications affect mortality. This could be helpful for interpreting the model's decisions at a later stage. For example, if the model predicts high mortality in conjunction with a certain complication, then we want to know if there is a causal relationship between them. If this is the case, then it would be fair to expect treatment of the complication would lower mortality. If not, then separate treatments might be necessary.

The post-operative complication data contains a binary complication values for each patient, therefore we elect the two-sided z-test for the difference between two proportions as the appropriate method. The test will determine, whether the proportion of patients not surviving the first 30 days after surgery is different between patients who do and do not suffer a certain complication after surgery. The test as-

sumes patients are independently distributed. Also, the sample size needs to be big enough, this assumption automatically holds, because we set the minimum occurrence frequency to 100. Next the method is applied to the specific case of delirium.

In the specific case of delirium the hypotheses become:

$$H_0 : \pi_1 = \pi_2$$

$$H_1 : \pi_1 \neq \pi_2$$

Where,

π_1: The proportion of patients without delirium that did not survive the first 30-days after surgery.

π_2: The proportion of patients with delirium that did not survive the first 30-days after surgery.

The z-test statistic is defined as follows:

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{2.1}$$

Where,

p_1: The sample proportion of patients without delirium that did not survive the first 30-days after surgery.

p_2: The sample proportion of patients with delirium that did not survive the first 30-days after surgery.

p: The sample proportion of patients that did not survive the first 30-days after surgery.

n_1: Number of patients without delirium.

n_2: Number of patient with delirium.

For the whole dataset the values are $p_1 = 0.072$, $p_2 = 0.109$, $p = 0.078$, $n_1 = 1376$, $n_2 = 293$. We plug these values into Equation (2.1) and find a test statistic of $z = -2.154$, corresponding to a p-value of 0.0313. Thus, we can not reject the null hypothesis, because for a two-sided test at a 95% confidence level the p-value has to be smaller than 0.025, in other words:

> There is no significant difference in mortality during the first 30-days after surgery between patients with delirium and patient who do not experience delirium.

Table 2.1 shows the results after repeating this statistical test for all complications. For *pneumonia* and *heart failure* the p-value is statistically significant. Therefore, we expect that the model will pick up signs of these complications to better predict mortality.

Table 2.1: Resulting p-values after a two-sided z-test for proportions, values in bold are significant at a 95% confidence level.

Complication	p-value
Delirium	0.0361
Anemia	0.1275
Pneumonia	$<$**0.0001**
Heart failure	$<$**0.0001**

2.2 Monitoring data

We conduct an extensive statistical analysis regarding the monitoring data to assess its quality. As discussed in Appendix 1, the data consists of a number of repeatedly measured vitals during surgery. Our goal is to evaluate the *general* capability of the monitoring data to predict post-surgery complications, therefore we only use the patient mean and standard deviation for each vital. Here we treat mortality as one of the complications.

We use the student's t-test to find out whether two populations of sample data have a significant difference in their means. However, for some complications the difference in sample size is too big for the student's t-test. For these complications we employ the Welch's t-test instead, because it is robust against differences in sample sizes. The assumption is that values are independently normally distributed. Hence, we created histograms for all vitals to see if it would be fair to make that assumption, Figure 2.1 shows two of these histograms. Regarding the heart rate the data seems normally distributed, although it is a bit right skewed. In contrast, the values for saturation follow an unusual distribution, due to the fact that throughout surgery saturation tends to be equal to its upper limit of 100%, therefore we can not assume normality. For this specific case we choose the Mann-Whitney U test instead, which does not assume normality. All other distributions, including the standard deviation of the saturation, are similar to that of the heart rate shown in Figure 2.1, so we apply the Welch's t-test to all data except the saturation mean.

First, we apply the test to a single case to clarify how the significance test is conducted. In this example we investigate, whether there is a significant difference in the mean heart rate of patients who develop pneumonia compared to those who do not. The equations below define our hypotheses:

$$H_0 : \mu_1 = \mu_2$$

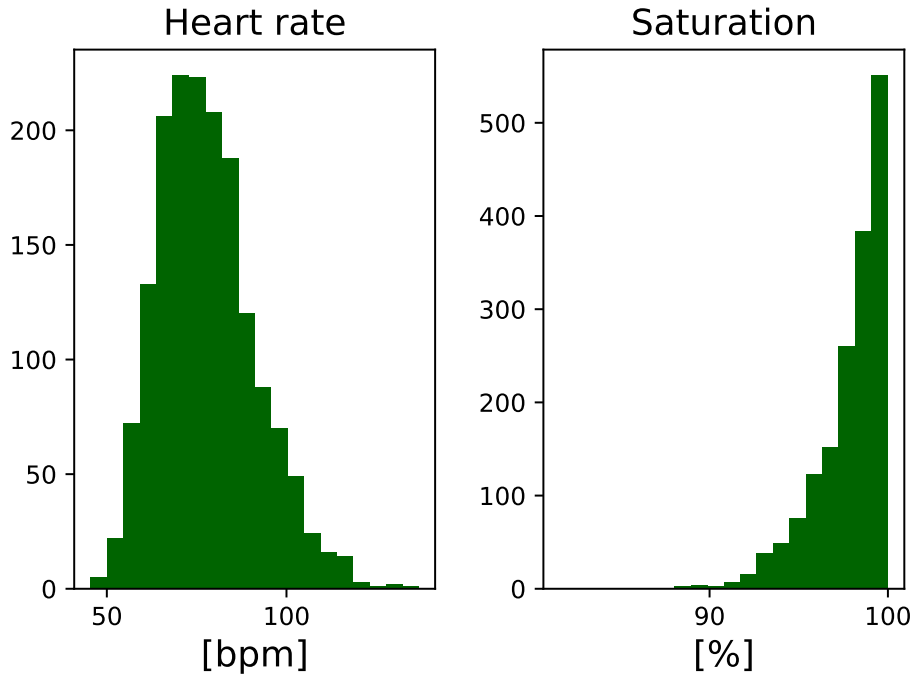

Figure 2.1: Distribution of the mean heart rate and saturation within the dataset. The heart rate seems normally distributed, while saturation is clearly non-normal.

$$H_1 : \mu_1 \neq \mu_2$$

Where,

μ_1: The mean of the average heart rate for patients that do not develop pneumonia after surgery

μ_2: The mean of the average heart rate for patients that do develop pneumonia after surgery

The definition of the test statistic:

$$t = \frac{m_1 - m_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{2.2}$$

Where,

m_1: The sample mean of the average heart rate for patients that do not develop pneumonia after surgery

m_2: The sample mean of the average heart rate for patients that do develop pneumonia after surgery

S_1: Standard deviation of the average heart rate for patients that do not develop pneumonia after surgery

S_2: Standard deviation of the average heart rate for patients that do develop pneu-

monia after surgery
n_1: Number of patients without pneumonia.
n_2: Number of patient with pneumonia.

In this particular example $m_1 = 77.6$, $m_2 = 80.2$, $S_1 = 13.5$, $S_2 = 13.1$, $n_1 = 1520$, $n_2 = 149$. Using Equation (2.2) we get a test statistic of $t = -2.501$, which matches a p-value of 0.013. With the same criteria as before the null hypothesis can be rejected at a 95% confidence level, in other words:

> There is sufficient evidence to reject the null hypothesis, that the mean average heart rate of patients suffering delirium after surgery is equal to that of patients who do not suffer delirium.

We repeat this test for all but one combination of complication and average vital sign/standard deviation of vital sign, Figure 2.2 and 2.3 show the results, respectively. As explained earlier, we apply the Mann-Whitney U test to the mean saturation data, though the Welch's t-test would yield the same conclusions. Combinations for which the difference in means is statistically significant are highlighted dark green.

Despite the idea that the standard deviation might be able to capture the level of fluctuation in vitals, whereas the mean can not, the latter has more significant combinations. Moreover, all but one cell in Figure 2.3 is also significant in Figure 2.2, thus vital sign fluctuation might not be important for complication prediction, however most probably standard deviation is inadequate to capture these fluctuations. Most importantly, on a general level there seems to be evidence of differences in vital signs between patients with a certain complication compared to patients without that complication. Thus, we assume that a machine learning model can learn about these differences and improve the prediction of post-surgery complications.

2.3 Medication data

We conclude our statistical analysis with tests on the medication data described in Section 1.1.2. In this case, we add another objective besides general assessment of predictive power, which is to identify how to add the data to the model. We explore two possible ways to do this: add as a binary variable or add the number of times administered as an ordinal variable.

For the first case both the input and output variables are binary, therefore we use the same test as in Section 2.1: the two-sided z-test for proportions. We apply the method in exactly the same manner, but with the complications added as outcome variables and medication groups as input variables. Figure 2.4 shows the resulting p-values, although the figure can be overwhelming, once again for easier navigation significant p-values are highlighted in dark green.

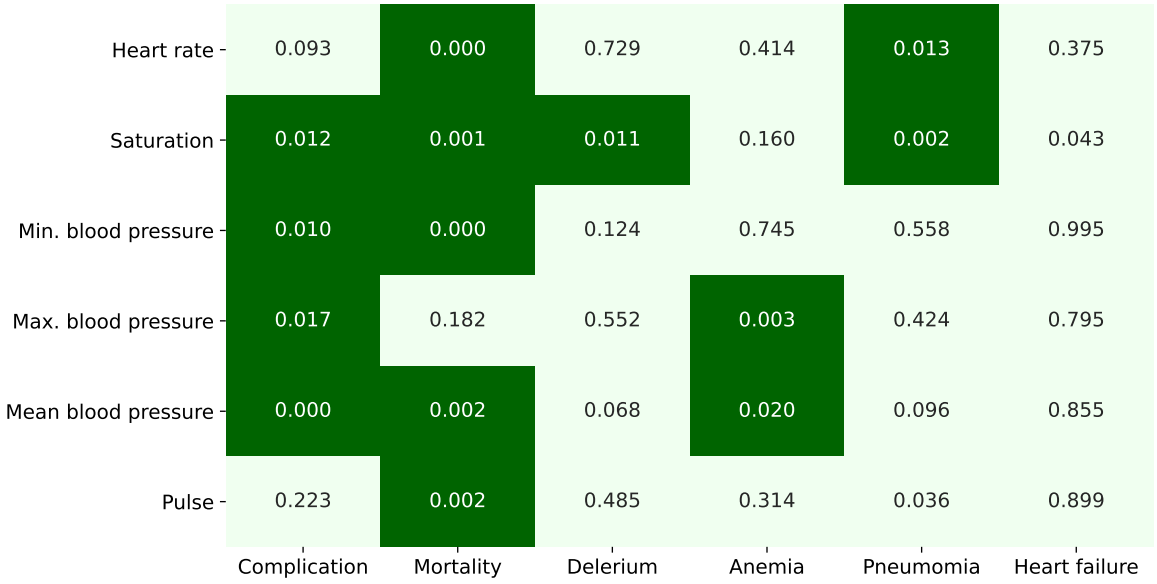

Figure 2.2: Resulting p-values of statistical tests to investigate the effect of the mean of the vitals on complications. Significant values have been highlighted in dark green.

Remarkably, there are not a lot of significant combinations. None of the medications significantly affect mortality or anemia, furthermore only seven out of the seventeen medication groups significantly affect at least one complication. We expect the binary medication data to be of limited value to the model, although these tests do not cover possible interactions with other parts of the data. For example, a certain medication may not necessarily increase mortality, but in combination with a high heart rate it does.

Next, we apply a different significance test to the medication data in ordinal format, whose values indicate how often a patient received a certain medication. We chose the Mann-Whitney U test to be appropriate, since this test does not require normally distributed data and can be used for ordinal data. Once again we discuss a simple example to help understand the results of every combination of medication and complication. Below we apply the method to the combination of bupivacaine and pneumonia; starting with the following hypotheses:

> H_0: The distribution of the number of times bupivacaine is given during surgery is the same for patients suffering pneumonia afterwards compared to patients who do not.

> H_1: The distribution of the number of times bupivacaine is given during surgery is not the same for patients suffering pneumonia afterwards compared to patients who do not.

Figure 2.3: Resulting p-values of statistical tests to investigate the effect of the standard deviation of the vitals on complications. Significant values have been highlighted in dark green.

The definition of the test statistic:

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_j) \tag{2.3}$$

with,

$$S(X,Y) = \begin{cases} 1, & \text{if } X > Y, \\ \frac{1}{2}, & \text{if } X = Y, \\ 0, & \text{if } X < Y. \end{cases}$$

Where,
n: sample size from population X
m: sample size from population Y

In this case, $n = 1142$ and $m = 132$, where the smaller group developed pneumonia after surgery. Using Equation (2.3) yields a U-value of 66698, which after conversion to a z-statistic corresponds to a p-value of 0.011. Therefore, we reject the null hypothesis at a 95% confidence level, or more precisely:

> There is sufficient evidence to reject the null hypothesis, that the number of times bupivacaine is given is not equally distributed between patients who develop pneumonia and patient who do not.

A more complex representation of the medication data does not yield better results, see Figure 2.5, there is one new significant combination, but another was lost. Also, the p-values are similar to Figure 2.4, which is in line with expectation after taking a closer look at the data, Most medications were given only once or twice, so the ordinal values still acted as if they were binary. Although, we use a weaker significance test, the variables itself remain very similar, thus resulting in similar p-values. Therefore, statistically speaking there is no reason to use ordinal values instead of binary values for the medication data.

Figure 2.4: Resulting p-values of statistical tests to investigate the effect of medication in binary form on complications. Significant values have been highlighted in dark green.

Figure 2.5: Resulting p-values of statistical tests to investigate the effect of medication in ordinal form on complications. Significant values have been highlighted in dark green.

2.4 Summary

Returning to our original goal of the statistical analysis, which was to judge whether there is predictive power within the per-operative dataset. Adding other complications as prediction targets might not only give a more complete post-surgery prognosis, but might also help explain the accompanying mortality prediction in the case of pneumonia and heart failure. Furthermore, the addition of monitoring data seems to be promising based on just the mean and standard deviation, since we design a model that can also learn patterns in the time series, expectations are high for this data. Finally, the medication data shows moderate statistical significance, where there is to be no reason to make it any more complex than a binary variable.