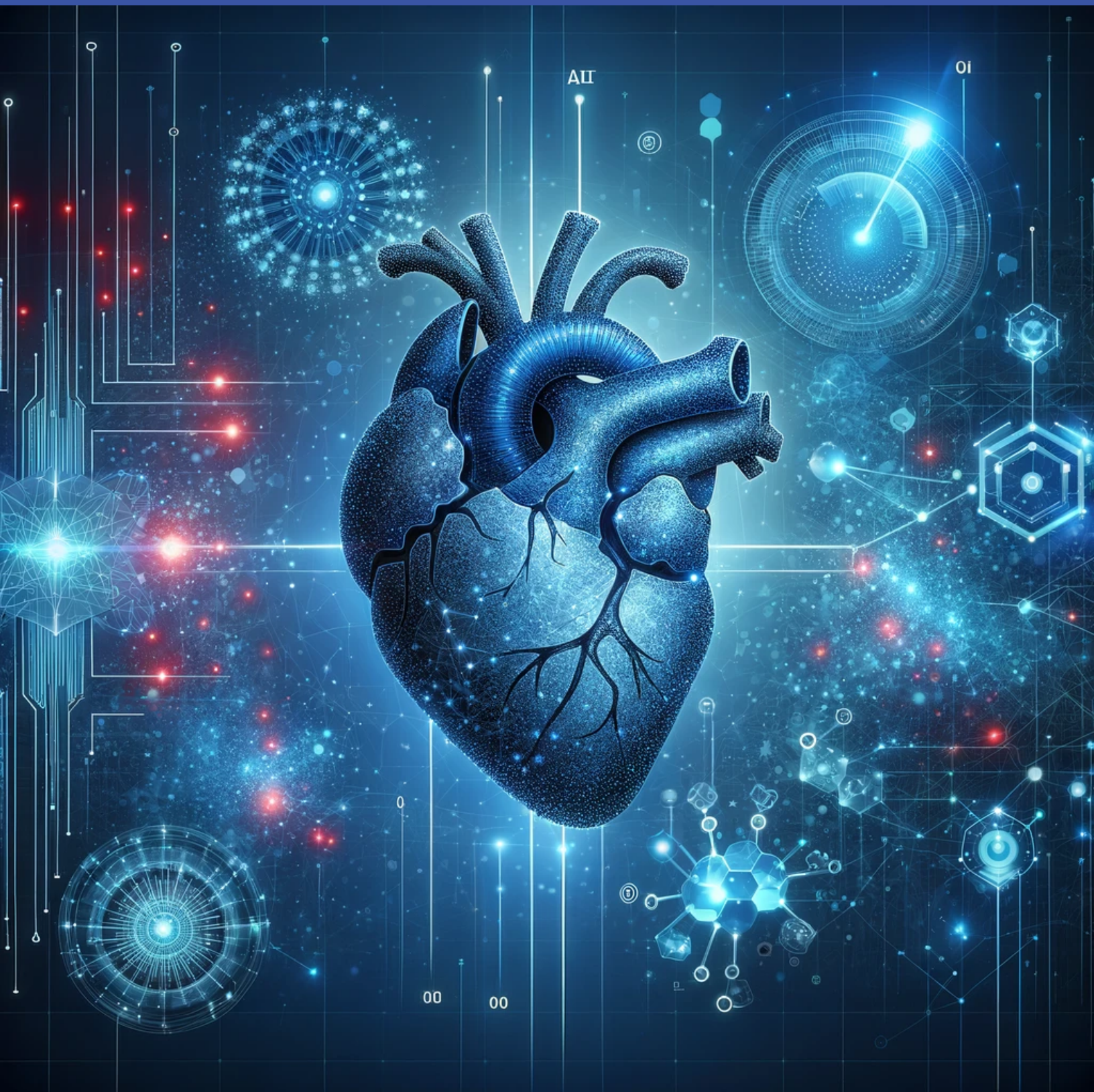


Bachelor of Artificial Intelligence and Data

Organ failure following cardiac surgery

02466 - Project Work



Bachelor of Artificial Intelligence and Data

Organ failure following cardiac surgery

02466 - Project Work

Date, 2024

By

Ditte Bjerrum Gilsfeldt (s210666)

Lucia Han Lu (s224215)

Muneer Kayali (s214642)

Viktor Skovlykke Sølvsten (s225784)

Copyright: Reproduction of this publication in whole or in part must include the customary bibliographic citation, including author attribution, report title, etc.

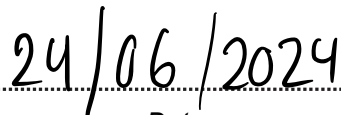
Cover photo: DALL·E (Chat GPT 4)

Unpublished by: DALL·E 2024-03-24 12.43.09, (13), Create an abstract illustration that embodies the concept of artificial intelligence surpassing traditional logistic regression methods in predicting .webp
<https://chat.openai.com/c/bd763d60-999d-4eb1-a1b7-85aa7e802882>

Approval

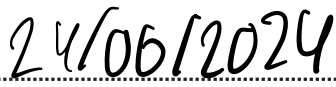
Ditte Bjerrum Gilsfeldt - (s210666)


Signature


Date


Lucia Han Lu - (s224215)


Signature

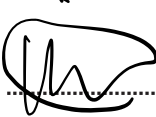

Date

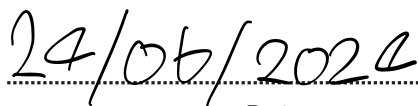
Muneer Kayali - (s214642)


Signature


Date

Viktor Skovlykke Sølvsten - (s225784)


Signature


Date

Acknowledgements

[Morten Mørup], [Professor], [Guidance counselor]

[Lars Grønlykke], [MD, PhD], [Guidance counselor]

[Chaoqun Zheng], [MD, PhD], [Guidance counselor]

[Theis Skovsgaard Itenov], [MD, PhD], [Guidance counselor]

Contents

Preface	ii
Acknowledgements	iii
Abstract	5
1 Introduction	6
1.1 Motivation	6
1.2 State of the Art	7
1.3 Data description	9
1.4 Research questions	10
1.5 Brief summary of the introduction section	12
2 Methods	13
2.1 Preparing the data	13
2.2 Machine Learning and training	15
2.3 Deep Learning and training	16
2.4 Implementation	18
2.5 Performance metrics	18
2.6 Model evaluation on extracted features	19
3 Results	20
3.1 Results from pre-operative data	20
3.2 Models trained on post-operative data	22
3.3 Models trained on subsets from different groups	24
4 Discussion	25
4.1 The global goals of UN	27
4.2 The ethics of AI usage	27
4.3 Perspectives on future research	28
5 Conclusion	29
A Appendix	30
A.1 Pictures from slide	30

Preface

This paper, Organ failure following cardiac surgery – the use of AI for prediction modeling, is written by a group of four bachelor students from the Technical University of Denmark (DTU) with the field of study being Artificial Intelligence (AI), interested in Machine Learning. Moreover, this paper is written upon the interest regarding how such technology is used in the vast healthcare sector, as it is found fascinating. This journey is, therefore, about exploring the potential of using artificial intelligence for the prediction modeling of organ failure following cardiac surgery.

The paper was written in the periods of a 5 ECTS 13-week course semester and a 3-week intense course, with a due date of 24/06/2024.

The intended audience of this research paper is fellow students of the AI study line, and anyone interested in the possibilities of AI related to medical and operational problems.

The importance of this paper focuses on showing how the technology of Machine- and Deep Learning can be useful for predicting morbidity and mortality, as a result of cardiac surgery. In addition to this, the hospitals are highly interested in optimizing the number of patients surviving the surgery, because, in worst-case scenarios, patients will end up in the intensive care unit (ICU), or even deceased. Hence, the hospitals are concerned with assessing if there exists a correlation between certain medical factors, which is to determine if a patient is suited for surgery or not in an early manner. Therefore, it is of hope to build models that can guide the hospital when making clinical decisions. Such models can provide information about the risk of the surgeries for different patients, and, further, to optimize the ICU so that hospitals can attain correct resources post-surgery. In summary, this research paper aims to provide hospitals with tools for predicting morbidity and mortality, which can help optimize the survival rate, and usage of certain resources and minimize complications post-surgery (3).

Furthermore, this paper intends to give an understanding of how AI models can help the health sector make better decisions for patients. Through the Discussion section in this report, the ethical aspect and the responsibility can help to illuminate how to use AI in the medical field.

This paper was the first time that the group ever worked with unprocessed real-world data, therefore, a big challenge was to sort the dataset and make a new dataset with features and variables that could be used to train the model.

Moreover, through this paper, the interest in working with AI in healthcare has grown bigger, and it would be interesting to work with it, for instance, in a future bachelor project. For this reason, the thoughts of this paper must be reconsidered, as a result of the inexperience working with real-world data, and the group must also recognize that the time was not sufficient for the ambitions that were set.

Lastly, we want to thank the guidance counselors who helped with the direction of the paper. Their names can be found in the acknowledgment section.

Bibliography

- [1] Andrei V Konstantinov, L. V. U. *A Generalized Stacking for Implementing Ensembles of Gradient Boosting Machines*. Studies in Systems, Decision and Control, 2021, Volume 350, pp. 4,.
- [2] Brownlee, J. How to configure xgboost for imbalanced classification, August 21, 2020.
- [3] Grønlykke, L., and Theis Skovgaard Itenov (Rigshospitalet), M. M. Organ failure following cardiac surgery – the use of ai for prediction modelling.
- [4] Herlau, T., Schmidt, M. N., and Mørup, M. *Introduction to Machine Learning and Data Mining*. Polyteknisk Kompendie, 2023, pp. 25, 26, 139, 140, 174, 175, 177, 245, 247, 248, 265.
- [5] Kanade, V. What is logistic regression? equation, assumptions, types, and best practices, 2022.
- [6] Kobayashi, Y., Peng, Y.-C., Yu, E., Bush, B., Jung, Y.-H., Murphy, Z., Goeddel, L., Whitman, G., Venkataraman, A., and Brown, C. H. Prediction of lactate concentrations after cardiac surgery using machine learning and deep learning approaches.
- [7] Lek, S., and Park, Y. S. Multilayer perceptron. 2455–2462.
- [8] (Lægemiddelstyrelsen), D. M. A. Faq om ai i medicinsk udstyr, 28. april 2021.
- [9] Manaswi, N. K. *Deep Learning with Applications Using Python Chatbots and Face, Object, and Speech Recognition With TensorFlow and Keras*. Apress Berkeley, CA, 2018, pp. 47 – 49, 115 – 126.
- [10] Molina, R. S., Molina-Rodríguez, M. A., Rincón, F. M., and Maldonado, J. D. Cardiac operative risk in latin america: A comparison of machine learning models vs euroscore-ii. *The Annals of Thoracic Surgery* 113, 1 (2022), 92–99.
- [11] Nashefa, S. A., Roquesb, F., Sharplesc, L. D., Nilssond, J., Smitha, C., Goldstonee, A. R., and Lockowandtf, U. Euroscore II, 6 January 2012.
- [12] Nguyen Thai-Nghe, Zeno Gantner, L. S.-T. Cost-sensitive learning methods for imbalanced data, August 21, 2020.
- [13] OpenAI. Dall-e 2024-03-24 12.43.09 - create an abstract illustration that embodies the concept of artificial intelligence surpassing traditional logistic regression methods in predicting.
- [14] Paluszek, Michael, Thomas, Stephanie, Ham, and Eric. *Practical Matlab Deep Learning*. Apress, 2022, ch. What Is Deep Learning?, pp. 1–24.
- [15] Panesar, and Arjun. *Machine Learning and Ai for Healthcare*. Apress, 2021, pp. 62–83.
- [16] Pierluigi, C., Oriol, P., and Petia, R. *BOOK CHAPTER · JOURNAL ARTICLE - Embedding random projections in regularized gradient boosting machines*. Springer Berlin Heidelberg, 2011, p. 203.
- [17] R, S. E. Understand random forest algorithm with examples.
- [18] Rosenberg, D. Unbalanced data? stop using roc-auc and use auprc instead, Jun 7, 2022.

- [19] Saini, A. Gradient boosting algorithm: A complete guide for beginners, 25 May, 2024.
- [20] Singh, H. Understanding gradient boosting machines, Nov 3, 2018.
- [21] Styrelsen, S. Sks-browser, vers 4.06, 01-04-2024.
- [22] Taud, H., and Mas, J. *Multilayer Perceptron (MLP)*. Springer International Publishing AG 20, 2018, pp. 452, 454.
- [23] Trust, R. P. H. N. F. about.
- [24] Tuychiev, B. A guide to the gradient boosting algorithm, Dec 2023.
- [25] Yang, T., Yang, H., Li, Y., Lui, X., Ding, Y.-J., Li, R., Mao, A.-Q., Huang, Y., Li, X.-L., Zhang, Y., and Yu, F.-X. Postoperative delirium prediction after cardiac surgery using machine learning models, February 2024.
- [26] Yu, Y., Zhang, C. P. Z., Shen, K., Zhang, Y., Xiao, J., Xi, W., Wang, P., Rao, J., Jin, Z., and Wang, Z. Machine learning methods for predicting long-term mortality in patients after cardiac surgery.
- [27] Zheng, C., Itenov, T., , and Grønlykke, L. *Cardiac surgery prediction modelling*. 2024, pp. 3, 4.
- [28] Zhou, H. Learn data mining through excel. 88 – 89.

Abstract

This project explores the application of Artificial Intelligence (AI) in regards to predicting organ failure following cardiac surgery, which is a critical issue in healthcare. The study investigates the potential of various Machine Learning (ML) and Deep Learning (DL) models to improve the accuracy of predictions for short-term and long-term mortality as well as morbidity, particularly focusing on kidney failure. Utilizing a dataset of 8000 patients from Rigshospitalet, collected over five years, the project aims to address the limitations of traditional prediction models like EuroSCORE II by employing the Logistic Regression (LR) model, Random Forest (RF), Gradient Boosting Machine (GBM), and Multilayer Perceptron (MLP) models.

The methodology involves extensive data preparation, including cleaning and preprocessing, followed by training and evaluating the models mentioned before by using performance metrics such as ROC AUC and PR AUC. The ML and DL models are implemented using the scikit-learn library in Python, with hyperparameters tuned via `GridSearchCV()`. The DL model, an MLP, is hypothesized to outperform traditional ML models due to its ability to handle complex, non-linear relationships within the data.

Despite challenges such as data imbalance and the limited size of the dataset, which restricted the complexity of the neural networks, the project underscores the potential of AI in enhancing clinical decision-making. The results indicate that while the models did not achieve optimal performance, primarily due to the data issues, they provide a foundation for future research. The study also highlights the ethical considerations of using AI in healthcare, emphasizing the need for responsible implementation aligned with the global goals of UN for healthcare and well-being.

In conclusion, this research contributes to the growing body of knowledge on the use of AI for medical predictions, suggesting that with more refined and balanced data and advanced techniques, AI can significantly aid in predicting post-surgical complications and optimizing patient outcomes.

1 Introduction

1.1 Motivation

When patients undergo cardiac surgery, a serious number of risks are known to follow. This includes declining health, complications during surgery, sicknesses, and in the worst cases, mortality. Hence, numerous research fields are focused on measuring such risks aiming to reduce and, ultimately, prevent them. The most commonly used tools for estimating complication rates and risks are the EuroSCORE II model and the STS risk score, already existing Machine Learning (ML) models for short-term mortality. However, these risk scores have shortcomings, as they are designed to predict mortality but neither morbidity nor mortality in longer terms, so, they lack deeper precision in regards to estimating surgery risks for the individual patient. Hence, it is relevant to research the possibilities of using Artificial Intelligence (AI) to create prediction models that compensate for the weaknesses observed in the current models. In such research, a logistic regression (LR) model will be developed to represent the EuroSCORE II along with other learning models, and comparisons will be made between all of them (3).

Ultimately, these models will be trained to detect both short-term and long-term complications in patients aiming to help prevent mortality and possibly morbidity, too. It is known from the project description that out of 1800 yearly patients who undergo cardiac surgery at Rigshospitalet in Copenhagen, a crucial amount of these individuals will experience complications post-surgery, keeping them hospitalized longer, and sometimes their health becomes fatal. Therefore, the investigation done in this project is essential, as this newfound knowledge can serve as guidance for doctors and patients deciding on the surgery (3).

To summarize, the possibilities of using AI to create accurate prediction models will be investigated in this project, and the investigation will contain ML and Deep Learning (DL), both representational for AI. It is hypothesized that AI can provide a dynamic predictive tool useful for aiding decision making regarding whether a patient should undergo surgery (25). Therefore, the objective of this project is to apply different AI models for the prediction of both short- and long-term mortality as well as morbidity, with the ultimate goal of achieving high accuracies. For morbidity prediction, this project will specifically focus on the risk of kidney failure post-surgery. To achieve these objectives, mortality and morbidity risks will be determined using an LR model, a Random Forest (RF) model, and a Gradient Boosting Machine (GBM). Additionally, a Multi-layer Perceptron (MLP) model will be employed for further investigation. It is expected that the DL and ML models (other than the LR model) will be particularly effective in predicting the risks of kidney failure (morbidity) and short- and long-term mortality after cardiac surgery. Given that these AI models are specialized in identifying complex relationships, they are likely to uncover subtle correlations and non-linear dependencies within the dataset that might be overlooked by an LR model. To test this hypothesis, three research questions have been formulated to guide the investigation.

However, the results of the research will not be trusted blindly. It is necessary to discuss the ethical perspectives when using AI as a tool in healthcare organizations as well as other settings. An important consideration is identifying who holds responsibility if something goes wrong. In addition to this, the goals of the United Nations (UN) for healthcare and well-being must be considered to ensure that these AI tools and models do not conflict with these objectives.

To extend the introduction, background knowledge of this project will be presented in a State of the Art section, which includes previous studies done in this field of research. Furthermore, the three research questions will be presented along with the corresponding methodology, and then a Data Description section will be formulated to provide an understanding and overview of the data at hand.

1.2 State of the Art

This section describes existing academic theory concerning previously used methods for creating the EuroSCORE II model, which predicts risks of mortality of patients undergoing cardiac surgery, along with other improved ML models based on the same field. These existing theories motivate the methods that this project uses for research.

1.2.1 The EuroSCORE II model

Firstly, the development of the EuroSCORE II model, based on the belonging EuroSCORE II article (11), will be described. The goals of this article were to update the European System for Cardiac Operative Risk Evaluation (EuroSCORE) risk model that originated from the year 1995. This model is mentioned as the first EuroSCORE model. However, in this article, the newer EuroSCORE II model, from 2012, was developed and trained based on data collected from 22381 patients who had undergone major cardiac surgery. The data was collected over 12 weeks, from May to July, in 2010. It was also mentioned that along with the risk factors already considered in the existing EuroSCORE model, new risk factors had been identified through the article's research, providing a more comprehensive understanding of the risks involved in cardiac surgery when creating EuroSCORE II. Moreover, for training this model, data was split into a subset for training a series of single variable logistic regression models and a subset for testing the models. The article concluded that EuroSCORE II was superior, as it was better calibrated while preserving powerful discrimination (the model was more accurate), and this conclusion was based on plotted ROC AUC curves. From the areas under the ROC curves, EuroSCORE II was seen to have a value of 0.8095, while the first LR model had a value of 0.7896. In addition to this, the article argued that this superiority was based on the fact that the models were trained on current data, thus, being better at reflecting current cardiac surgical practice. Finally, it was also deduced, from the models, that cardiac surgical mortality has been significantly reduced in the last 15 years despite including older and sicker patients.

1.2.2 Improving the EuroSCORE II model

EuroSCORE II is currently one of the world's leading cardiac risk models (23), but it is limited to only predicting the probability of a patient dying either during or shortly after their proposed surgery. But with improving technology and growing knowledge, older models like EuroSCORE II itself are bound to require improvements. Therefore, it has become a growing interest to create newer and stronger AI models. This implies developing models, using both ML and DL techniques, that can cover what EuroSCORE II is unable to. The article (10) is an example of how researchers have tried to apply more advanced ML models for predicting mortality within 30 days. In this study, 6 different ML models were trained, and they were LR, Naïve Bayes (NB), MLP, Support Vector Machine (SVM), RF and GBM, and their models were all implemented in Python using SciPy, XGBoost, Keras, and Sci-kit learn libraries. Moreover, it is explained that hyperparameter tuning for all the models was done using a grid search on the test set. The best performing models were found by comparing both the ROC AUC and PR AUC of each model. Then, based on the ROC and PR analyses, the study concluded that the GBM was the most precise model and, too, more precise than the EuroSCORE II model. However, it was mentioned that the difference was not statistically significant.

The contents of this article will provide guidance for the possible ML models that will be investigated in this project.

1.2.3 Using ML for long-term mortality prediction

In addition to predicting short-term mortality, other studies have attempted to create models for long-term morbidity prediction as well. The article (26) has done exactly this, where it aimed to construct and validate several ML models and a DL model when trained and used to predict long-term mortality along with identifying risk factors in patients post-cardiac surgery. Eight different methods were used, and these were LR, Artificial and Recurrent Neural Networks (ANN), NB, Gradient Boosting machine (GBM), adapting boosting (Ada), RF, bagged trees (BT), and eXtreme Gradient Boosting (XGB). These models were compared based on their accuracies and their ROC AUC along with other metrics. Based on the results, the study concluded that Ada performed best in predicting mortality within 4 years after cardiac surgery.

Results from the study show that the Ada model had an accuracy of 0.834, however, the RF model had the highest accuracy, being 0.841, and the LR model is also noteworthy, having an accuracy of 0.835 both being higher than the Ada model. Though, the study weighted ROC AUC the most in regard to performance evaluation. Here, the Ada model had the highest probability with a value of 0.801, whereas the RF and LR models had values of 0.789 and 0.797 respectively.

In sum, the findings from the two mentioned studies, particularly what ML models were used and their evaluation metrics (for instance, ROC AUC and PR AUC) will serve as significant inspiration for this project. By using similar methodologies, this project aims to develop precise prediction models for mortality risks.

1.2.4 Predicting post-operative complications

It is evident that newer studies have shown to the utilization of DL and ML methods regarding the cardiac field. The article(6) is another instance of this since it highlights how researchers have used the ML and DL models RF, ANN and a Multivariate Linear Regression (MLR) model with a focus on predicting lactate concentration in patients post-cardiac surgery by using data collected during surgery. Higher levels of lactate concentration can be damaging to a patient's organs and body tissues, and it is, therefore, described as a complication post-surgery. In addition to this, the article mentions that lactate concentration levels are used as a monitoring factor during surgery, as it ensures that all body tissues and organs receive sufficient blood flow and oxygen during the surgery. From this, the study found that it is indeed possible to predict levels of lactate in the blood after surgery with moderate accuracy by using data collected both before and during the surgery. Therefore, in opposition to the previously mentioned articles, which used AI to predict mortality risks, this study primarily focused on predicting morbidity risks.

One of the key factors that this article highlights is how it argues that numerous prediction models in cardiac surgery settings only used static patient characteristics, hence, tabular data, and a limited number of intraoperative variables (variables that are present while undergoing surgery). However, it was known that during cardiac surgery, particularly during cardiopulmonary bypass, minute-level data was available on key parameters such as flow, hemoglobin concentration and mean arterial pressure. Because conventional prediction models often ignored such dynamic data, DL approaches were better suited to incorporate time-varying and nonlinear data that had more complex interactions. This knowledge of DL being able to take advantage of minute-level data is highly important for this project, although, this project does not aim to train models using minute-level data, it will discuss the importance and potential of such data, acknowledging that implementing it requires high complex models with numerous parameters, which exceeds the current scope.

To conclude, the results from this article have highlighted the strengths of DL, and despite not using DL on minute-level data, this project will still research methods within DL and train such models on tabular data. On the whole, the different approaches from all of the mentioned articles will serve as guidance for creating accurate predictions of both mortality and morbidity, ultimately leading to improved clinical decision making and resource optimization in the ICU.

1.3 Data description

The data used in this project is collected from 8000 patients over five years, who have undergone cardiac surgery at Rigshospitalet. The Cardiac patient data journey has a pre-operative section such as *sex, age, medical- and surgical history* and more, peri-operative such as *blood pressure, heart rate, pressure* and more, ICU such as *blood pressure, heart rate, dialysis* and more, and, finally, Ward such as *blood pressure, heart rate, medication*. Everything is further shown in appendix A.1.

Furthermore, there will be a restriction when working with the data, due to the GDPR rules, thus, the data can only be used on one specific device, where it is located on a locked drive, only with access for the staff who needs access, when training and creating the AI models. This also means that all personal information has been anonymized, removing information like CPR numbers.

To compensate for the removed CPR numbers, each measurement and test result is linked to a specific identification number on the form *ID 1, ID 2* and so on in all data files, assuring the tracking of correlation between the samples. None of the observations is specifically labeled as pre-operative, peri-operative (during the time of surgery), or post-operative in practice. Measurements like blood pressure are often sampled once under all three categories. Instead, each observation has, along with the ID, a date and time that describes when each observation was noted. In addition to this, the data is structured in multiple tables that are distributed across 27 different CSV files, and they all contain duplicate observations making the data highly disorganized. Meanwhile, it is still possible to keep track of when in the surgery process of a patient the observations are made (by assessing the date and time). For a general understanding of the dataset, the following categories have been made:

1.3.1 Profile

The profile describes information about the patient, that is already known before the process of operation begins. The variables documented are presented below.

Profile variables	
Biological gender	Male or Female
Age	Years
Diagnosis	Name and SKS
Alcohol	Drinks per week
Smoking	Frequency and previous history

Table 1.1: Profile variables

1.3.2 Measurements

This section of the data describes all the information that is collected during the patient's hospital duration. Hence, these data contain the most information about why and what future complications the patients will experience. In other words, the variables presented below will most likely make up the majority of what is used for prediction.

Measurement variables	
Blood pressure	Millimetres Mercury(mmHg)
Mean Arterial Pressure	mmHg
Invasive Arterial blood pressure	mmHg
Central venous pressure	mmHg
Pulmonary Artery Pressure	mmHg
Oxygen saturation in the blood	Percentage
respiratory rate	Breaths per minute
Blood and body temperature	Celsius
Pulse	Beats per minute
Weight	In Kilograms, measured several times over the process
Urine	Millilitres
Laboratory test results	Very broad, but most important is liver health and infections
Medicine administered	Both name and amount
EKG results	Several measurements
Echocardiography	Several measurements
Coronary angiography	Examination of Coronary arteries
Spirometer test	Ventilation and general health of the lungs
Other operations performed	A week after the cardiac surgery
Total incrition	Food, liquid and IV drop, etc.
Total excretion	Defecation, urine, vomit, bloodloss, etc.
Ventilator data	Duration of the assisted breathing and settings
Duration of anesthesia	Describes the duration of cardiac surgery
anaesthesia data under the operation	Several measurements
Blood loss during operation	Milliliters and cause of bloodloss
CV bypass and closing of aorta	Duration of the manoeuvre

Table 1.2: Measurement variables

The most important variables to understand from the dataset are the CV bypass and closing of the aorta, which is how the blood is kept circulating through the body during the operation. The duration of this can often describe how successful the operation was and how much the patient will bleed during and after the operation, which is a key factor in recovery time.

1.3.3 Admission and follow-up process

The Admission process describes how long each patient spends in the different locations. This is documented by specifying the location, for example, *Thorax intensive* and also the start of admission and end of admission. Sometimes a patient is transferred from one ICU to another for logistical reasons but as the care in the different ICUs is the same this can be ignored. The duration spent in the ICU will be one of the main target variables.

The follow-up information with the patients is also documented. This includes variables such as if death has occurred within a year and if so when. This is a very important variable for describing the overall success of the operation.

1.4 Research questions

The main focus of this project is to assess the risk of mortality for patients who have undergone cardiac surgery, and this specifically includes the investigation of mortality within 30 days up to 90 days post-surgery (short-term) and up to a year (long-term). In addition to this, another point of interest is the prediction time spent in the ICU, done by investigating the risk of kidney failure, hence, non-mortal complications following a patient's cardiac operation, namely morbidity.

To achieve these objectives, the following research questions have been formulated below.

First research question:

To what extent is a Machine Learning (ML) or Deep Learning (DL) model, other than the Logistic Regression (LR), capable of predicting the mortality risks of a cardiac surgery patient?

Objective:

The goal of this question is to create a model that can give a high accuracy similar to the EuroSCORE II model, or conclude that an LR model, representative of the EuroSCORE model, is better suited for predicting mortality following cardiac surgery.

Methodology:

This will be operationalized by first choosing and designing suitable ML and DL models that can predict the probability of mortality risks. It is important to create a common ground for comparison for a valid result. Therefore, apart from comparing newfound results to results displayed in the original EuroSCORE II article, a new LR model will be trained as a representation, which will act as the EuroSCORE II model (because the original EuroSCORE II model is inaccessible, and the data used in this project differs from data that was used to train the EuroSCORE II model). Then to obtain a model with similar performance strength or even to outperform the LR, an RF model could be a powerful method as it is useful for handling data that is non-linear. Furthermore, a Gradient Boosting Classifier (GBC), also known as the GBM, and a Multilayer Perceptron (MLP) could also be useful, since they can capture even more complex patterns that may have been overlooked by models like the LR and RF.

Second research question:

If the performance of the models between different attributes for instance, age, sex, pulse and previous diseases will vary, how can they be detected using AI, and what influence would they have on helping hospitals predict mortality and morbidity?

Objective:

The goal of this question is to determine if there is a difference in a model's performance between the different predicting variables and the outcome of the surgery. A potential conclusion could be that there is a stronger model performance when tested between people aged above 60 and risk of mortality compared to people below the age of 60 as well as other features.

Methodology:

This will be investigated by comparing the calculated performance metrics of each feature, which will be called groups, across the different AI models. When training the models, the certain groups used will be chosen by logical reasoning. To then compare the performances between the different groups and the outcome, precision and recall can be used to provide insight into how the models performed depending on the groups. Moreover, the results can be plotted for clear illustration, which will make it easier to analyze the impact of the different groups and if there exists differences between them.

Third research question:

How well can ML or DL models predict post-operative complications like organ failure and more specifically kidney failure? Can the findings help doctors take further precautions and, thus, help prevent more patients from staying in the ICU?

Objective:

The goal of this question is to create an AI model that can predict the risk of a patient's kidney failing post-surgery, hence contributing to the possibility of being able to determine the expected length that a patient will spend in the ICU.

Methodology:

This question will be researched by first creating some relevant metrics to predict, which in this case is kidney failure. Afterward, a DL Network will be fitted to the data to find linear or nonlinear correlations between measurements and the chosen post-operative metric. ML models like Gradient Boosting Machines (GBM) and RF will also be used.

1.5 Brief summary of the introduction section

To summarize, this project aims to explore the potential of using AI in developing prediction models for assessing the risks associated with post-cardiac surgery settings. Current models that are in use in the medical field, such as the EuroSCORE II and the STS risk score, primarily focus on short-term mortality and have limitations in predicting long-term outcomes and specific morbidities. This study will, therefore, develop AI models, including an LR model, an RF, a GBM, and an MLP model, to evaluate if they succeed in predicting both short-term and long-term complications. In other words, it aims to investigate if such models can make accurate predictions in areas where models like the EuroSCORE II fall short.

Three research questions were formulated to achieve this. The first is "To what extent is a Machine Learning (ML) or Deep Learning (DL) model, other than the Logistic Regression (LR), capable of predicting the mortality risks of a cardiac surgery patient?". Here, it is hypothesized that advanced ML and DL models will yield high accuracies when predicting mortality risks due to their ability to capture complex patterns in data. The second question was "If the performance of the models between different attributes for instance, age, sex, pulse and previous diseases will vary, how can they be detected using AI, and what influence would they have on helping hospitals predict mortality and morbidity?", which was based on the hypothesis that different attributes might significantly influence a model's performance, providing valuable insights into risk factors. At last, the third research quest was formulated as "How well can ML or DL models predict post-operative complications like organ failure and more specifically kidney failure? Can the findings help doctors take further precautions and, thus, help prevent more patients from staying in the ICU?". It is hypothesized that AI models, especially DL models, will accurately predict post-operative complications, thereby helping in clinical decision-making and improving patient outcomes.

Following the introduction, the contents of this project will then be elaborated through a section on Methods, which includes relevant theory, and a Results section presenting the findings. The results will be analyzed in a Discussion section, which, too, will discuss ethics and how and why the findings are relevant. Finally, everything is rounded up by a conclusion.

2 Methods

Research methods that have been used throughout this project along with associated theory are described in this section.

2.1 Preparing the data

Before training any ML or DL models, the data was prepared, which included cleaning, prepossessing and then selecting the most important features.

2.1.1 Cleaning/feature selection

The first thing to do was to clean the data, which has been done by handling missing values, fixing values that are formatted differently, and removing unwanted observations. This was important since the data was distributed across 27 different CSV files that were unstructured and within the files, the data had a varying number of both columns (variables) and rows (ID numbers) since the data held duplicated observations. Moreover, the data suffered from unconventional column naming and inconsistent metrics. Therefore, duplicates or observations that were deemed either unnecessary, irrelevant or lacking sense have been removed to reduce noise and improve the overall quality of the data. For instance, the profile variable *smoking* found in Table 1.1, a chosen feature from one of the CSV files, which recorded the frequency and previous history, had a measurement named *never recorded*. These were seen as missing values and, hence, removed.

The most important change that was made to the data was that all IDs that did not have recordings of an induction event, a stop event, or a recording of forceps having been on and off the aorta were removed. This was to guarantee that for the remaining IDs, there existed a start and stop timestamp on the operation itself. Therefore, these timestamps were saved for the remaining ID numbers such that for each of these patients, it was recorded when their surgery had started and ended.

Using these timestamps, pre-, peri- and post-operative data could be found. For example, in one of the CSV files, the dataset with Laboratory test results (one of the measurement variables in Table 1.2, namely the *lab* dataset), a subset of entries were made to fulfill the condition that the timestamp of the lab samples was between the start and stop timestamps for a patient's operation. This made sure that all of the entries were peri-operative data samples. All of this was to further ensure that each of the IDs left had survived a surgery, hence, if morbidity or mortality has happened, it was not during the surgery.

In general, the *lab* dataset was very messy and included more than 300 different unique tests. When selecting features from this dataset, it must be done with great consideration of missing values. For instance, for the pre-operative lab data, only *hemoglobin*, *leukocytter* and *trombo-cytter* were chosen as features, because these tests were done on most patients before the surgery, so they provided minimal data loss. This was another example of how missing values would be handled. Moreover, the average value for each patient was taken and saved as a feature. For the peri-operative lab data, *pulse average* and its standard deviation as well as *saturation average* and its standard deviation were chosen. The reason the standard deviation was also taken was because these values tend to swing much more than some of the other features.

Two other CSV files called *bleeding* and *urine* seemed to be useful as well via consultation with the supervising doctors (these files recorded the measurement variables Blood loss during operation and Urine, also from Table 1.2). Here, for each patient, there were records of measurements of how much blood they have lost and how much urine they have exerted, both before, meanwhile, and after the surgery. However, most patients were measured under the surgery, which was why these measurements were filtered such that only peri-operative measurements were outputted. For these two measurements, the sum (in ml) of blood lost and urine exerted for each patient was taken and used as a feature. These features (especially bleeding) could be great indicators of whether the surgery is going in a positive or negative direction (under the opinions of the doctors supervising this project).

For the post-operative lab data, the measurements hemoglobin, kalium, natrium, pulse and saturation were chosen as features as these, too, provided the smallest data loss. For all these measurements, both the average and standard deviation were also used as features for the postoperative data.

Then, a CSV file with the diagnoses (the Diagnosis profile variable from Table 1.1) for each patient were used to construct the *disease* feature. Here, each patient had a list of diseases they were diagnosed with. This feature is later and further worked on in the preprocessing stage. Another feature that was extracted was the operation SKS code, also from the Diagnosis variable. This column recorded what type of surgery each patient had gotten, which was valuable information.

Finally, as a result of this cleaning process, the entirety of the data would be construed as two separate datasets with no missing values, one being for all pre-operative data, so all data recorded before the start timestamp of a patient's operation, named the pre-operative dataset. The other dataset was for both pre-, peri-, and post-operative data, named the post-operative dataset. Hence, the post-operative dataset included all data recorded before the start of the operation but also all data that was recorded after the start of the operation, meaning data from both during the operation and after. The pre-operative dataset had a total of 24 features, while the post-operative dataset had a total of 40 features. In addition to this, the target variable that was chosen for the pre-operative dataset was *dead within a year*. Here, from the *population* CSV file, the death date was taken for each patient (if dead) and the binary feature *dead within a year* was constructed by comparing to the surgery date of the patient. For the post-operative dataset, the feature *kidney failure* was constructed and chosen as a target variable. This was done by looking at the *dialysis* CSV file, which is a treatment for people whose kidneys are failing. Here, the binary feature was constructed such that if a patient from the population appeared in the dialysis dataset, then that patient had kidney failure.

After having cleaned the data and extracted the needed features, the new datasets would be ready for the next step.

2.1.2 Preprocessing

The next step was about data preprocessing, meaning that the data was transformed such that it was a better fit for analysis. The preprocessing was done by changing the metrics or formats of some of the features. One example was changing the format of the features that included dates and time as hours because these were observed as strings, hence, they were changed to a date and time data type using a feature in Pandas. In addition to this, to handle the numerous disease types recorded as different names and SKS codes, the diseases were split into multiple groups, namely *Respiratory Disease*, *Circulatory Organ Disease*, *Type 1 diabetes* and *Type 2 diabetes*, *Other metabolic disease* and *Genital or urine related diseases*, and all groups took in binary variables, so False for not having such disease, and True for having such disease.

This was done by looking at the first two letters of each unique SKS code. For instance, the ones starting with DI means circulatory diseases, and codes starting with DJ means respiratory diseases (21), and then a loop was ran through the SKS codes for each patient, and binary values were appended to the new lists accordingly. A similar was done for the different types of operations, where the most frequently occurring types were made into new groups, and the rest were gathered under a group called *Other operations*, and all still being binary. By doing so, the whole dataset could be transformed such that nonnumerical values were avoided as much as possible, wherever it is found relevant. This would help the AI models learn more efficiently with faster convergence. Furthermore, considering the messy format of the data at hand, ensuring homogeneity is also highly important. For instance, in the profile variable *alcohol*, it was seen that some measurements were a single integer, while others were an interval. This was taken care of by taking the mean to all encountered intervals within *alcohol*, hence, achieving homogeneity. Once reprocessing was done for all the needed variables, the data was ready for feature selection.

The last step that was done for data preparation was standardization of all the continuous instances within both datasets. This standardization step was crucial for RF, as it ensured that the regularization of the LR model would be uniformly applied. For instance, if standardization was not done, then features with larger scales would dominate the regularization process. In addition to this, MLP would also use the standardized data, since it was known that its activation functions, too, were sensitive to the scale of the inputs. However, the RF model and GBM were still trained on the regular dataset, since the scale of the input features did not affect the nature of those algorithms.

2.2 Machine Learning and training

Following cleaning and preprocessing of the data, it was then ready for model implementation. It is evident from the previous sections that ML is a commonly used tool in this field of research. ML is a subset of AI where computer models are trained to learn from their environment over time with the intention of improving their performances (15). Such models are, therefore, able to find certain patterns in given data, allowing them to perform better, the more they learn.

Hence, in this step, in order to answer the first research question, the three chosen ML models, LR, RF and GBM, have been trained. Each of the three ML models has been trained twice, once on the standardized version of both the pre-operative dataset and then the post-operative dataset, and they had 24 input features and 40 respectively. For the pre-operative dataset, the target variable was mortality, meanwhile, the models would target the risk of kidney failure when trained on the post-operative dataset, and in both scenarios, the threshold for predicting a positive class was set to be above 0.5. This meant that a probability over 0.5 meant the model would predict 1, and for anything under, the model would predict 0.

In addition to this, how the models were trained along with hyperparameter tuning will be explained in further detail in the Implementation section.

2.2.1 Logistic Regression (LR)

One of the models within the ML branch that was trained was the Logistic Regression (LR) model. In general, the LR model is used for binary classification tasks by predicting the probability of an outcome, event, or observation. This model analyzes the relationship between one or more independent variables to classify data into discrete classes (5). The goal of the LR function is to obtain values for a set of coefficients by fitting a logistic function to the data. This is achieved by maximizing the likelihood of the observed data, which involves finding the coefficients that make the observed data most probable under the model. (28).

The LR model trained to predict mortality risks was used as the representation of EUROSCORE II. The creation of a new LR model was due to the reason that the original EuroSCORE II model was inaccessible, and the EuroSCORE II model took some variables that the dataset investigated in this project did not contain as input. The performance of the different models will be compared by assessing ROC AUC and PR AUC values, which, too, will be explained later.

2.2.2 Random Forest (RF)

The next model that was implemented and trained was a Random Forest (RF) model. The RF model is able to combine numerous outputs from individual decision trees into a single, aggregated output. This ensemble approach allows RF to handle complex datasets effectively (17), as it uses and combines the strengths of multiple decision trees to improve the overall predictive performance. The RF model can manage both regression and classification problems, making it versatile for various types of data analysis (1).

The hypothesis was that both the pre-operative and post-operative datasets contained complex non-linear relationships. Consequently, an RF model could potentially outperform the LR model due to its ability to handle such complexity. By constructing multiple decision trees and collecting all of their predictions, the RF model can capture intricate patterns and interactions within the data that might be missed by simpler models like logistic regression.

2.2.3 Gradient Boosting Machine (GBM)

The last ML model that was trained, also known to be a strong ML model for handling complex patterns, was a Gradient Boosting Machine (GBM), and it was hypothesized to have robust predictability since it combined predictions of multiple weak learners. To explain further, The GBM is applicable for both regression and classification problems, and it aims to find a function $F^*(x)$ that minimizes the expected value of the loss function when mapping x to y over a joint distribution for all (y, x) -values. Normally the function $F(x)$ will be a parameterized class of the function $F(x; P)$ where $P = \{\beta_m, a_m\}$ is a parameter set (16). That is why, a GBM is a combination of the predictions of multiple weak learners, and these are used to obtain a more precise outcome.

Technical details:

To thoroughly understand GBM, one must be familiar with the concept of boosting. Boosting is a method that transforms multiple weak learners into a single strong learner by fitting new trees to an adjusted version of the original dataset. In addition to this, Gradient Boosting (GB) can also be compared to the well-known AdaBoost algorithm. However, the difference is that Adaboosting uses data points with large weights to identify possible shortcomings, whereas the gradient boosting algorithm uses the gradient in the loss function to achieve something similar (20). Moreover, the GB algorithm takes tabular data with a set of features (X), and a target (y), and the algorithm aims to generalize unseen data points based on the training data (24). To minimize both bias and variance errors, the GBM model iteratively improves upon errors on previous models. For continuous variables, a gradient boosting regressor is used, with mean squared error as the loss function. For classification problems, a gradient boosting classifier is employed, using log-likelihood as the loss function (19).

2.3 Deep Learning and training

To expand on the first research question and to answer the third one, an MLP, within the DL branch, has also been trained. Previously, as presented in the State of the Art section and other sources such as the book (14), DL has proven to be a promising strategy for data analysis in medical settings.

DL is a subset of machine learning that refers to neural networks with more than one layer of neurons. Neural networks behave similarly to the neurons in the human brain, therefore, the name *deep learning* is taken to imply that a learning system is a deep thinker. Since deep learning networks usually have multiple layers before the final output, they can solve more complex problems, especially regarding predictions and classifications (14). Hence, via DL, it is possible to create models that can handle the complexity of the dataset explored in this project. The models within the DL branch that are used will be described below.

MLPS can model complex non-linear relationships and automatically learn feature representations through backpropagation, as known from the theory section. Hence, it was hypothesized that the MLP would be more powerful than traditional ML models like the trained LR, RF, and GBM. The ability of an MLP to handle various data types and use advanced optimization techniques further enhances its predictive capabilities. Therefore, this Neural Network (NN) was trained to predict both kidney failure and risk of mortality.

2.3.1 Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) model is a type of artificial neural network(ANN) that is used for non-linear modeling (7), enabling the capture of complex relationships. MLP is structured with an input layer, at least one hidden layer and an output layer. Neurons in the MLP are all the nodes in each layer except the input layer, and they use a non-linear activation function, for instance, sigmoid or ReLU. Furthermore, the MLP is trained via backpropagation, a supervised learning technique, which minimizes the loss function like entropy, and it uses an optimizer for controlling the parameters which are weight and bias (9).

Explaining backpropagation:

The process of training the MLP is also important to mention, and as known from above, it uses backpropagation. Here, the backpropagation algorithm is done through the application of two stages. The first stage is simply the process of forward passing through the network. The second stage is where the available desired responses get compared to the responses that are already produced. In this stage, the output layer of n neurons will have n deviations (errors) between the desired responses and those produced by the output neurons, and these errors are calculated and then used to adjust the weights and thresholds of all neurons (22). In summary, backpropagation means that errors will go from the output to the hidden layer and then to the input layer (working backward), and then the learning rate and optimizer will be regulated by adjusting the parameters accordingly.

Technical details:

To address further technical details of the MLP model, it is relevant to understand the structure of the MLP network. The network consists of an input layer, a number of hidden layers and an output layer, and due to this structure, each input layer will be propagated layer-by-layer toward the output layer (22). For the input layer, the number of neurons within this will equal the number of chosen input features. Moreover, the number of hidden layers, each having a number of neurons usually depends on the complexity of the patterns within the dataset. However, due to the nature of the dataset, the MLP was set to have only one hidden layer. This based on the nature of both dataset (pre- and post-operative), where there number of entries were insufficient when taking the number of features in consideration. Therefore, by using one hidden layer with a moderate number of neurons the risk of overfitting is reduced. For instance, if the model were to have too many layers, it would introduce an excessive number of weights compared to the available data, thereby increasing the likelihood of overfitting. In addition to this, the possible activation functions to be considered for the hidden layer are Rectified Linear Activation (ReLU), Logistic (Sigmoid) and Hyperbolic Tangent (Tanh).

Lastly, the output layer consists of one neuron with a Sigmoid activation function (4) when used to predict the risk of both mortality and morbidity (kidney failure), since this function is best suited for binary classification, as it ensures that the neuron outputs a value between 0 and 1, which was interpreted as the risk probability.

It was hypothesized that the MLP would be more powerful than traditional ML models like the trained LR, RF, and GBM, and it would yield the highest ROC AUC and PR AUC values. The ability of an MLP to handle various data types and use advanced optimization techniques further enhances its predictive capabilities.

2.4 Implementation

The models implemented in this project were developed using the scikit-learn library in Python, which helped ensure consistency along the different implementations. Additionally, each model was confined within a function and all of them followed a similar structure. The core structure of these functions involved defining the model, tuning its hyperparameters using `GridSearchCV()`, and then selecting the best estimators based on CV performances run with 5 folds. Within the grid search, 'accuracy' was used as the scoring method, since this method was known to be insensitive to class imbalance, which the dataset suffered from. Finally, the best estimators based on the grid search were selected and used to create the optimal model, defined as `best_model` within each function. Following this, the `predict_proba()` function from scikit-learn was called to return the probability estimates for all classes for each instance in the test set. This meant that the probability of each model predicting 0 and 1 were appended to create a 2D array, which was then sliced into a 1D array to only hold the probability of the model predicting 1.

To explain the implementation of each model in depth, starting with the LR model, which used a Sigmoid activation function by default to model the probabilities for binary outcomes, and during training, the cross-entropy loss function was used as a measure of the difference between the predicted probabilities and the actual labels. The hyperparameters that were tuned included 'C' (the inverse of regularization strength) and 'max_iter' (the maximum number of iterations for the solver to converge). Furthermore, the RF model did not use activation functions. Instead, it relied on the chosen Gini impurity criteria to measure the quality of splits in the decision trees, which helped the model determine the best splits to reduce impurity and error, thus improving its predictive performance. The hyperparameters tuned included 'max_depth' (the depth of the trees), 'max_features' (number of features to use for making the best splits within the trees), and 'n_estimators' (number of trees). For the GBM, it, too, did not use activation functions. It optimized a Sigmoid activation function (used as default) by sequentially adding trees that corrected the errors of the previous ones, thereby improving the model iteratively. The hyperparameters that were tuned were `n_estimators` (number of boosting stages (trees) to perform), `learning_rate` (controlled how much each tree contributed to the training), and `max_depth` (maximum depth of each tree). Finally, the MLP was set to use the cross-entropy, and the hyperparameters that were tuned were 'activation' for what activation function to use, choosing between 'relu' or 'logistic', 'alpha' (the strength of the L2 regularization term), and 'max_iter'. When all models were trained, they would be ready for performance evaluation.

2.5 Performance metrics

In this step, each model would be evaluated using nested grouped K-fold cross-validation (CV) as a leave-one-out strategy since many IDs had multiple rows associated with them. Using grouped CV ensured that all rows with the same ID were kept together in either the train or test split, hence, independence between the splits was maintained. The nested grouped K-fold CV was implemented in Python using the `GroupKFold` function from the `sklearn.model_selection` module and this Grouped CV was run for each of the four different AI models.

This method involved creating 5 outer folds, each with 5 inner folds while ensuring that all samples with the same ID were either in the training set or the test set for each fold. The evaluation function `evaluate_model` performed a nested CV, where an outer loop split the data into training and testing sets, and then an inner loop performed further splits on the training set for hyperparameter tuning. Finally, the performance metrics PR AUC and ROC AUC were calculated for the outer folds, and these were stored in a `results` dictionary, which would be used for model evaluation.

To elaborate on the performance metrics, firstly, PR AUC was computed by calculating precision and recall using the function `precision_recall_curve()` from `scikit-learn`, and then PR AUC could be calculated using the `pr_auc()` function. Precision is the proportion of positive predictions that were actually correct, and recall is proportion of actual positives that were identified correctly. Furthermore, ROC AUC was computed by calculating False Positive Rate (FPR) and True Positive Rate (TPR) via the `roc_curve()` function and then ROC AUC was found using the `roc_auc_score()` function. TPR is the ratio of correctly predicted positive observations compared to all the actual positives, and FPR is the ratio of incorrectly predicted positive observations compared to all the actual negatives.

ROC AUC and PR AUC were chosen as performance metrics, as they were known to be able to provide comprehensive insights into a model's ability to handle imbalanced datasets, which was a critical aspect of this project since the datasets had extremely few instances of positive values of both morbidity and mortality. In short, ROC AUC is a measure for the discrimination ability of a model, while PR AUC measures how well the model balances the trade-off between precision and recall. Moreover, this choice of performance metrics was also based on the fact that the EuroSCORE II model, too, used ROC AUC for evaluating its performance, and the article (26) used ROC AUC and PR AUC as well. In general, PR AUC is less sensitive to class imbalance, because it considers the performance over all thresholds, incorporating both TPR and FPR in a balanced manner. On the contrary, PR AUC is more sensitive to class imbalance, because it only focuses on the performance with respect to the positive class, where precision can be heavily influenced by the number of false positives, hence, making it a critical metric for understanding how well the models would perform in identifying the minority class without being overwhelmed by false positives (4). Therefore, it was hypothesized that the models would be biased towards predicting the majority class for both datasets (no morbidity and no mortality), leading to extremely low PR AUC values.

Following all of this, the ROC- and PR curves would be plotted across the five outer folds for all models. Graphs plotted for the ROC curves would have FPR as the x-axis and TPR as the y-axis, and graphs plotted for the PR curves would have recall as the x-axis and precision as the y-axis.

2.6 Model evaluation on extracted features

The final step that was done was to investigate how the different groups may influence the performance of the models. This investigation was done by calculating separate performance metrics. The groups chosen were age, split into people above and under age 60, and biological sex, as male or female. The influence of these groups were tested by letting the LR model, trained on the full training split of the pre-operative data, predict the values of each subgroup separately. Then, a nested grouped K-fold CV was used for calculating the previously mentioned metrics (ROC AUC and PR AUC). Returned would, therefore, be four measurements of both ROC- and PR AUC, one for each of the groups male, female, above and below the age 60, and the measurements would be used to create a Bar plot using `Matplotlib`, illustrating how well the model could distinguish between mortality outcomes across each group.

3 Results

Yielded results from each of the trained models are presented in this section.

3.1 Results from pre-operative data

The calculated values for ROC AUC and PR AUC for all of the AI models, LR, RF, GBM and MLP, trained on the pre-operative dataset for predicting mortality are shown in table 3.1 below.

	ROC AUC	PR AUC
LR	0.8140	0.2282
RF	0.8036	0.1974
GBM	0.7940	0.1716
MLP	0.7953	0.2471

Table 3.1: ROC AUC and PR AUC for each model trained on pre-operative data to predict the risk of mortality

From the table 3.1, it is evident that the LR model had the highest ROC AUC value of 0.8140 among all models. The LR model also had the second highest PR AUC value. The RF model had the second highest ROC AUC value, with a PR AUC value lower than both the LR and MLP models. The GBM had the lowest ROC AUC and PR AUC values of 0.7940 and 0.1716 respectively. Finally, the MLP model had a performance similar to the GBM in terms of ROC AUC but had the highest PR AUC value of 0.2471.

All models demonstrated decently high accuracy, with values ranging from 0.79 to 0.81. However, the PR AUC values were relatively low, between 0.17 and 0.25.

ROC AUC and PR AUC curves for each of the above mentioned models have been plotted to visualize the results. These plots show the PR curves and ROC AUC curves for each of the 5 outer folds, presented as different colors. For the PR curve, recall was plotted on the x-axis while precision was plotted on the y-axis. For the ROC AUC curve, the FPR was plotted on the x-axis, and the TPR was plotted on the y-axis.

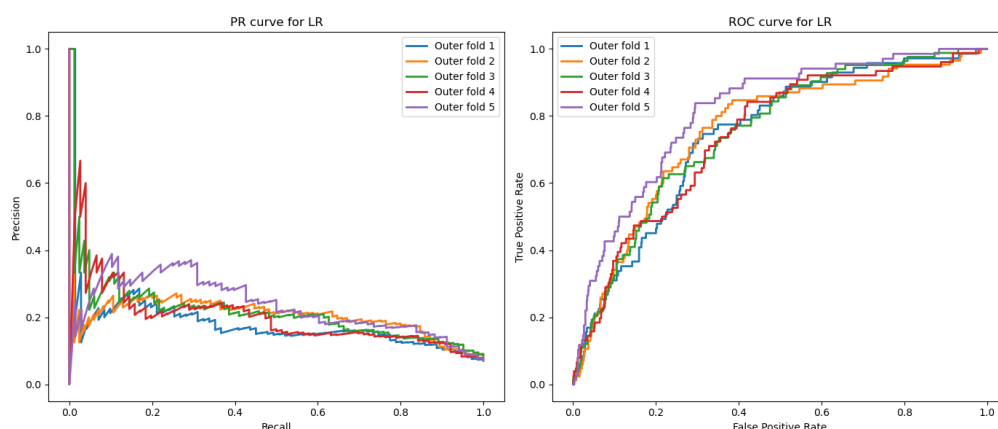


Figure 3.1: PR AUC curve and ROC AUC curve graphs for the LR model

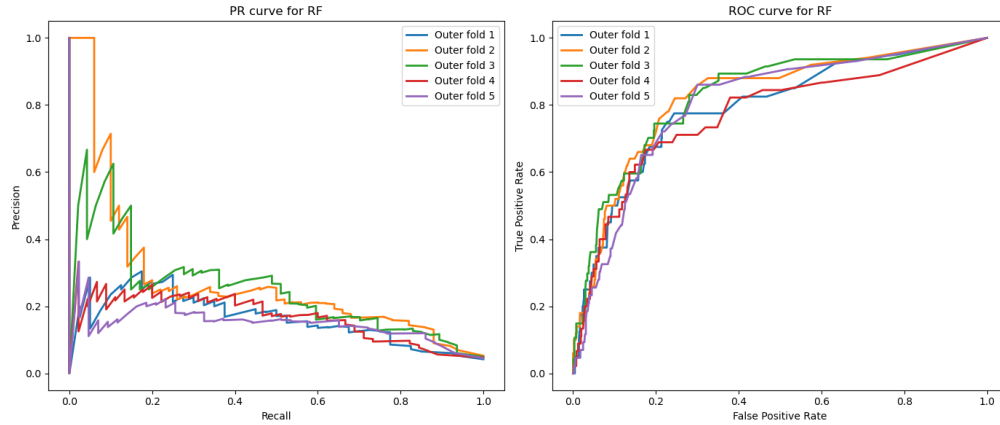


Figure 3.2: PR AUC curve and ROC AUC curve graphs for the RF model

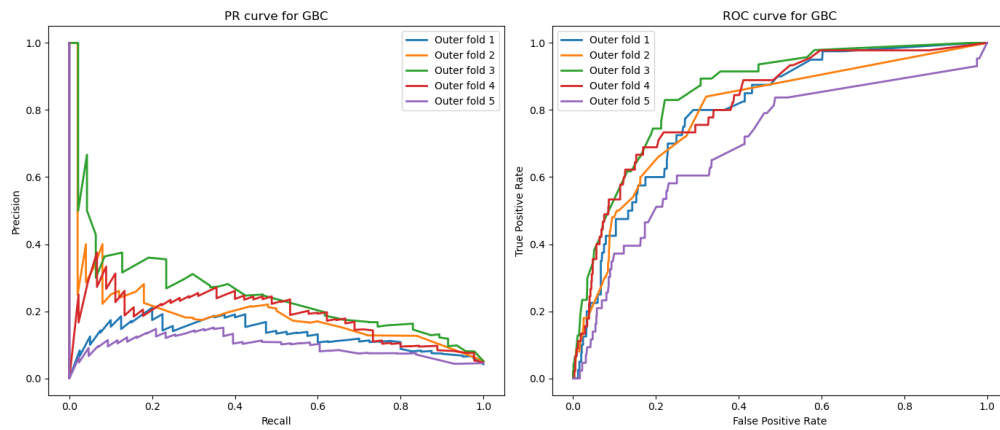


Figure 3.3: PR AUC curve and ROC AUC curve graphs for the GBM

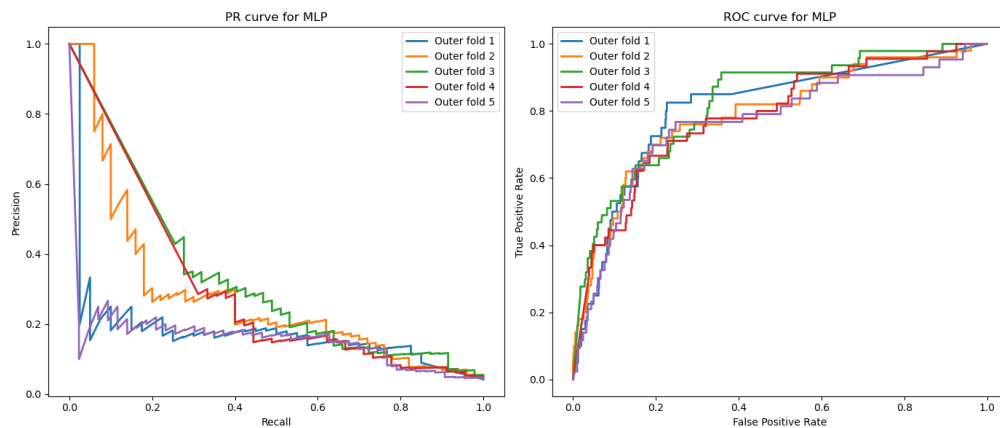


Figure 3.4: PR AUC curve and ROC AUC curve graphs for the MLP model

The PR AUC curves for each model demonstrate variations across different folds, with each fold represented by a different color in the plots. Additionally, all models exhibit a decrease in precision as recall increases, which is reflected in the shape of the PR AUC curves. Similarly, the ROC AUC curves for each model show variations across different folds, with each fold shown in different colors in the plots. The ROC AUC curves for all models generally bow towards the top-left corner of the plot.

These observations are consistent for all models (LR, RF, GBM and MLP) trained on both pre-operative and post-operative datasets, illustrating the performance metrics across different cross-validation folds and highlighting the variability in model performance due to the different splits in the data.

3.2 Models trained on post-operative data

Table 3.2 below presents the calculated ROC AUC and PR AUC values of the models trained on the post-operative dataset.

	ROC AUC	PR AUC
LR	0.9312	0.2736
RF	0.9213	0.2897
GBM	0.8992	0.2428
MLP	0.9132	0.2097

Table 3.2: PR AUC and ROC AUC for each model trained on post-operative data to predict the risk of morbidity (kidney failure)

It is evident from Table 3.2 that the LR model had the highest ROC AUC value, which was 0.9312, and it had the second highest PR AUC value. The RF model had the second highest ROC AUC value, and for this time, when all models were trained on the post-operative dataset, the RF had the highest PR AUC value. The GBM had the lowest ROC AUC value of 0.8992, but its PR AUC was still below the LR and RF but higher than the MLP. The MLP had a ROC AUC higher than the MLP, and it had the lowest PR AUC value of 0.2097. Compared to before, all models have demonstrated higher accuracy in general, as the ROC AUC values ranged from 0.8992 to 0.9312, and the PR AUC was between 0.2097 and 0.2736.

Like in the section above, the ROC AUC and PR AUC curves for each of the trained models have been plotted below.

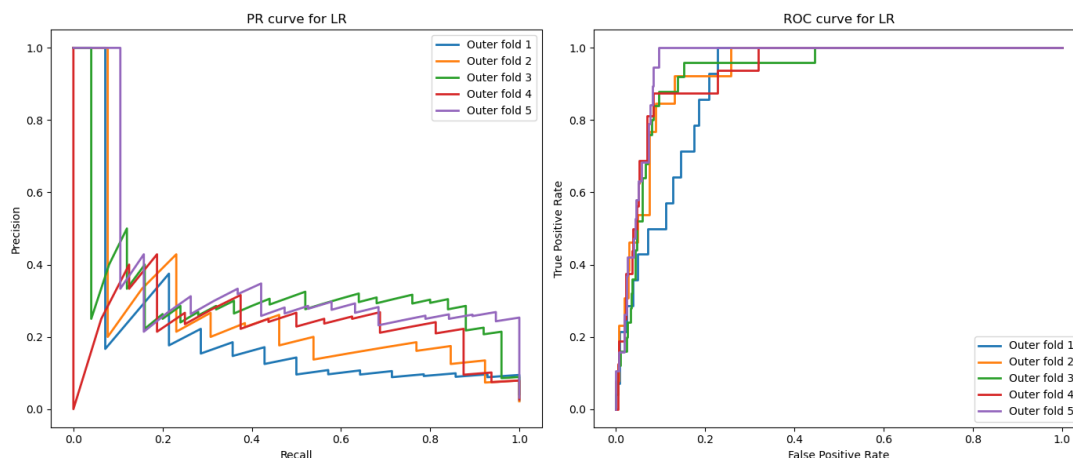


Figure 3.5: PR AUC curve and ROC AUC curve graphs for the LR model

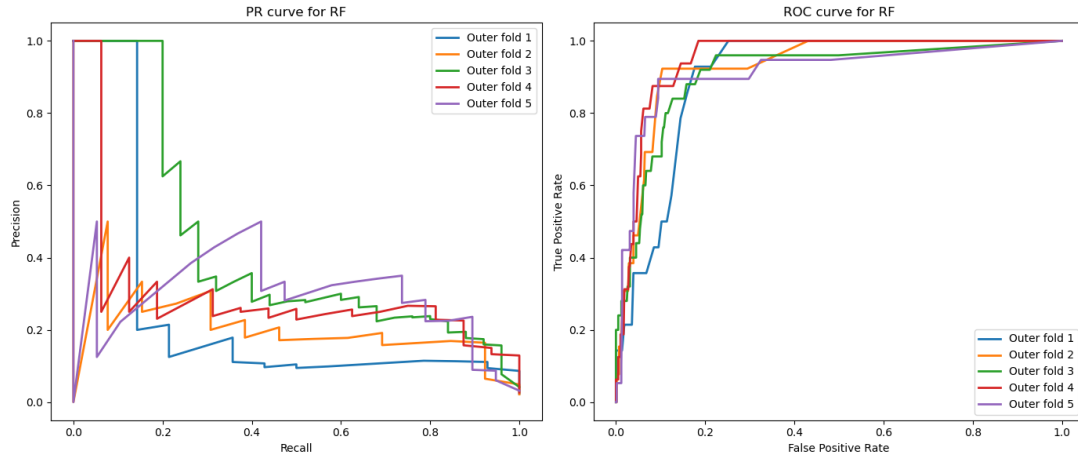


Figure 3.6: PR AUCcurve and ROC AUC curve graphs for the RF model

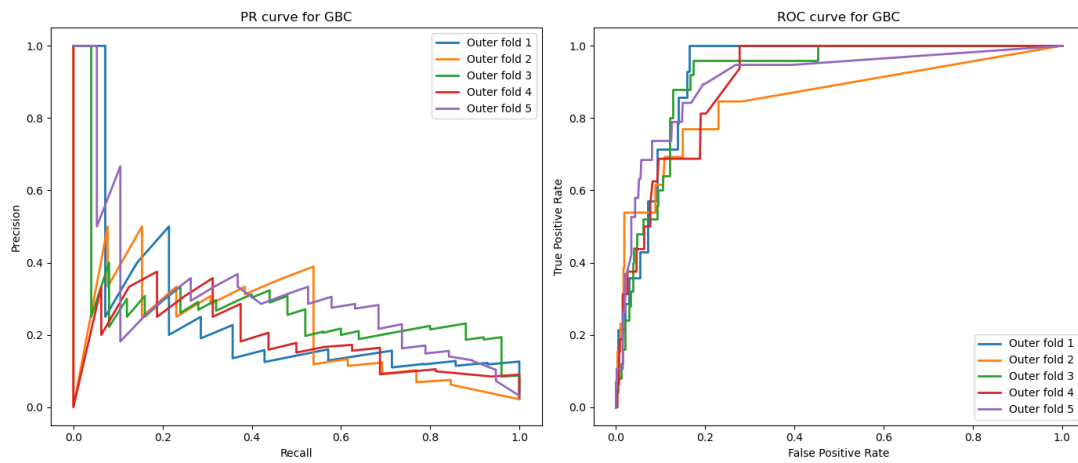


Figure 3.7: PR AUC curve and ROC AUC curve graphs for the GBM

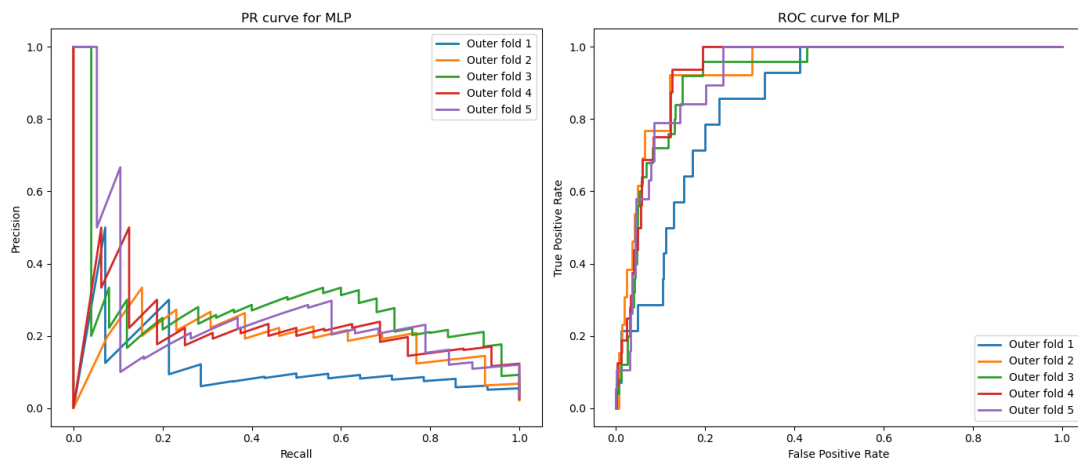


Figure 3.8: PR AUC curve and ROC AUC curve graphs for the MLP model

Both the ROC AUC and PR AUC curves had a similar form as the ROC AUC and PR AUC curves of all models trained on the pre-operative dataset, as plotted in the section above.

3.3 Models trained on subsets from different groups

The Figure 3.9 below shows that it was evident that the average ROC AUC value of the LR model was the different across each of the four groups (men, women and above- and below age 60). The values for the men and women groups were seen to be 0.8130 and 0.78660 respectively, and for people aged above- and below 60, the values were 0.79160 and 0.77540 respectively.

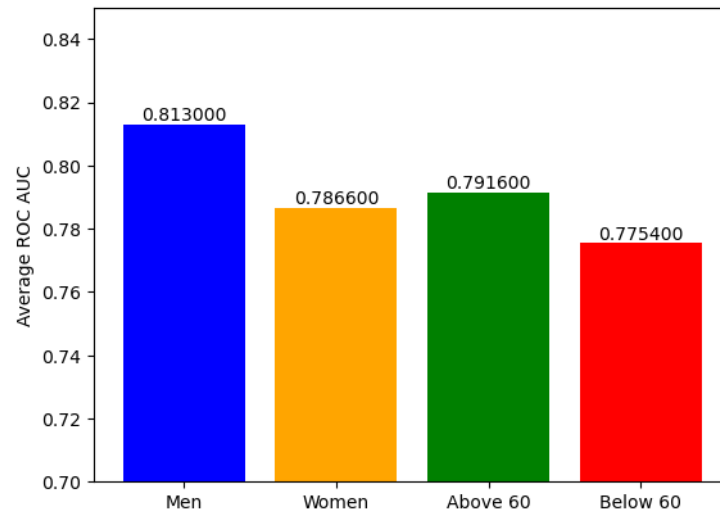


Figure 3.9: Bar plot of the average ROC AUC performance of the LR on the different groups

These values showed that the performance of women was slightly lower than the performance of men, which had the best performance. Furthermore, the PR values for the LR model across each groups are shown in Figure 3.10.

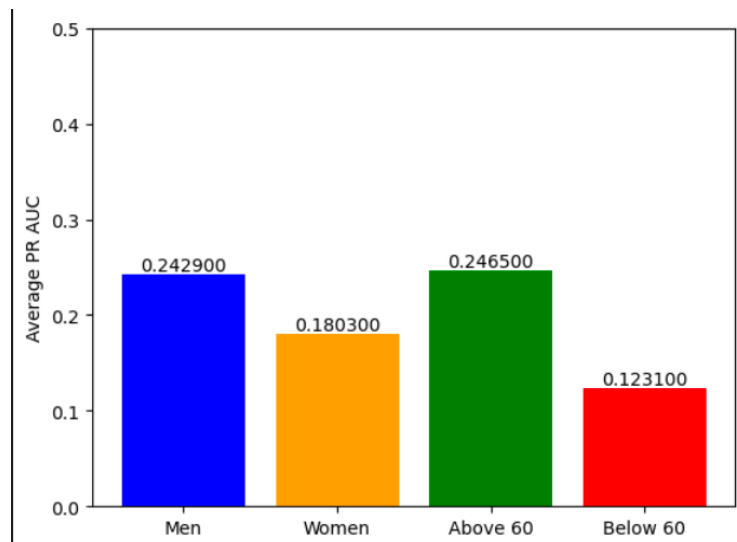


Figure 3.10: Bar plot of the average PR AUC performance of the LR on the different groups

The values for the men were seen to be 0.24290 and the values for the women 0.18030 respectively, and for people aged above- and below 60, the values were 0.24650 and 0.12310 respectively. From this, it was shown that the performance of people above 60 was slightly better than the performance of men, and that the performance of people above 60 was best.

4 Discussion

4.0.1 Interpretation of results

In this section, the results will be interpreted and discussed, and the ethics behind the models will be covered. The global goals of the UN regarding this field of research will also be discussed. Furthermore, perspectives of future research will be formulated.

From the Results section(3), it was evident that all models had high ROC AUC values. However, high ROC AUC values can be quite misleading due to the class imbalance in the dataset (including both the pre and post-operative datasets). When most entries in the dataset were negative labels (no mortality nor morbidity), the model might learn to classify most entries as negative, thereby, bringing the FPR down and the FNR up, due to the large number of true negatives. Hence, the ROC AUC values are not the only metric to consider, when deeming a model successful, as this metric does not consider class imbalance. Due to the highly imbalanced nature of the target variables, the model has very poor individual performance and PR AUC (18).

The reason for the poor output of the PR AUC, is that this plot is very sensitive to class imbalance because of the nature of precision being indifferent to true negative values and very sensitive to false positives.

The ROC curve loses some of its value when working with a large unbalanced dataset, due to the fact that the false positive rate is a ratio between false positives and total true negatives. This means that when the number of total true negatives is very large, a large amount of false positives will not influence the plot significantly.

These remarks above are crucial, as the positive class is far more compelling and more important than the negative class. In projects like this, one would prefer that an ML falsely classify a patient positively, rather than falsely classifying the patient negatively, such that, potentially, one could avoid deaths or complications post-cardiac surgery. Generally, there could be many reasons to why the implemented models did not achieve the goal of this project, but the most significant reason is believed to be the limited size of positive target classes. The problem with the data used for the project is that, originally, it had a moderate size of about 8000 patients, however, not all patients (IDs) have clear and structured measurements. According to the doctors, when a patient undergoes surgery, the forceps are put on and taken off the aorta of the patient. Therefore, IDs that did not have recordings of the aortic forceps being put on or off were removed. Furthermore, as mentioned in the Methods section, all patients who did not have a *start* and *stop* timestamp for their surgery were also removed, as one would not have been able to collect pre-, peri or postoperative data from them. The result was that almost half of the IDs in the dataset were lost. In addition to this, numerous features across the dataset were not measured for all patients, resulting in an even greater loss, as missing values were removed in the data cleaning process. This led to choosing between more features with missing values or fewer features with less missing values, for this paper the ladder was chosen. It is also important to consider that many of the measurements that should have been included (they were recommended by the doctors, since they contain important information, for instance, QT interval) did not have enough entries to cover most patients in the population, hence, they were dropped. Furthermore, another important feature that the doctors mentioned was the KAG results of each patient, which included critical information about the patient's heart vessels. However, this dataset was deemed unusable due to its messy nature.

This displays another problem that needs to be balanced. Often what doctors deem important and descriptive does not always align with what is technologically feasible or possible. An effort was of course made to try and accommodate the doctors ideas but in the end many variables were dropped due to the quality of the dataset.

To further explain the encountered problems, the data is constructed such that data about each vessel location is recorded, however, the number of unique entries for vessel location is 186. This is because the column includes strings written by doctors, and, thus, some classes were similar, but written in different ways (for instance, *Ost LM* to *Ost LAD* and *LM* to *LAD* or *Midt LM* to *LAD*). Furthermore, when looking at the strings with the most appearances, *Midt LAD* appears as the most common class but with only 1670 recordings. There was no clear way to divide these classes, and it was completely unfeasible to create 186 new features for each patient (granted, many entries would still have to be deleted due to missing values in these columns). As a result, this feature was completely dropped, which is unfortunate, as it would have possibly provided valuable information for the AI models.

Another complication encountered in this study is that patients who have died during surgery were not considered, since it was unclear what measurements to consider when looking at this type of patient. If a patient has died during surgery, a large number of features that were chosen would, most likely, contain missing values, and, hence, be removed. Most, if not all of these patients were removed when filtering out all IDs that did not have a *stop* timestamp for their surgeries.

Furthermore, the small sizes of the final datasets were also a problem for the MLP. Here, a Neural Network (NN) can learn complex relationships and patterns, and thrive when datasets have high dimensional, provided that there are enough data entries. The small number of data entries at hand required the NN to be small and simple, with only one hidden layer that included 1 to 5 neurons found from hyperparameter tuning. This reason is that with large NNs, the number of calculations per data entry can quickly exceed the number of total data entries. This would result in an NN turning into a glorified LR, and despite considering these factors during model development, this still happened to the trained MLPs (which unexpectedly performed worse on the postoperative dataset, which is the dataset with the highest dimension). Hence, the sizes of both the pre and post-operative datasets caused the MLPs too much trouble in learning processes, leading to the models not performing as hypothesized.

Coming back to the previous points regarding the ROC and PR curves, a big problem with this dataset was the huge class imbalance. This has created challenges during this project. In retrospect, resampling techniques should have been experimented with, such that one may have randomly over-sampled the positive class, skewing the bias the other way. Moreover, models or methods that address class imbalance should have been considered, such as cost-sensitive learning, where one can change the cost function of an algorithm to penalize misclassification of the minority class more heavily (12), or algorithms where one can tune the hyperparameters such that classes have different weights (2). Mainly, the problem of this project was that class imbalance in the data was hugely underestimated. If this had been given more attention and not taken lightly, then perhaps, a moderately accurate and trustworthy model, close to the performance of the EUROScore II model, could have been achieved. The data size would still be a problem, however, oversampling would have probably mitigated this impact, even if it was only slightly.

Regarding the subject of model performance on different social groups of patients (the chosen were men, women, people aged above- and below 60), no extensive statistical analysis was done to conclude whether there is a significant difference in performance for the different groups. For instance, via the model's ROC- and PR AUC values, it is shown from the Bar plots (Figures 3.9 and 3.10) that the RL model performs better on patients above 60 compared to those below, but that cannot be concluded for certain without statistical analysis. One reason for dropping statistical analysis on the difference in model performance across the groups is simply lack of time, but the biggest reason is that this subject is not as interesting, and it would not grant much insight as long as the models are not performing optimally. This is the same reason why no statistical analysis was done to compare the different models. The models, simply, did not perform well enough for it to be interesting or meaningful to perform any comparisons between them.

4.1 The global goals of UN

With the issues regarding results and the complications that were encountered having been identified and discussed, it is also important to associate these with broader objectives. The findings are relevant to one of the global goals of the UN, namely *Good health and Well-being*, specifically sub-goal 3.4.1: *"Mortality due to cardiovascular disease, cancer, diabetes or COPD."* Therefore, by improving the AI models developed in this project, results may contribute to this sub-goal, as the aim is to ensure that people can live healthy lives and that their well-being is ensured. If the research done in this project were to contribute to making the goal of the UN more feasible, then the models must perform better, especially so that they have a low FNR. It is dangerous if a developed AI model misclassifies a patient as negative (not dying of the cardiac surgery), but in reality, they were a true positive (they would die). This would cause deaths that otherwise could have been avoided, since such a patient would be misled to undergo the operation, despite them being at extremely high risk of dying if having it done.

In addition to this, it is also important to balance a model's FPR, for the reason that a patient classified by the model could be lead to not undergo a required surgery to avoid a death that would not have happened, since the model made a misclassification, predicting the patient to be positive, when, in reality, they were a true negative.

Therefore, to obtain better performing models that could help achieve the *Good health and Well-being* goal of UN, the model needs to be trained on data that is more balanced. This can be achieved through methods such as resampling, as discussed earlier.

4.2 The ethics of AI usage

In addition to this, the Danish Medicines Agency (DMA), *lægemiddelstyrelsen*, discusses AI in the context of the Medical Device Regulation (MDR) laws. This discussion revolves around when AI-based software qualifies as a medical tool and how it should be used in that capacity. Based on the online article from DMA, AI based software is classified as a medical tool if it has an effect on, for instance, diagnosing, preventing, predicting, or treating diseases for an individual (8). The goal of this research was to create AI models that could predict mortality and organ failure risks, hence, by definition, these models would classify as a medical tool. However, if these AI tools were to be used in a medical setting, ethical questions would arise, such as who bears responsibility for AI when errors occur? Figure 4.1 from DMA is used to illustrate this problem. The figure shows that there must always be at least one responsible, either the manufacturer of the AI, the user, or the staff. In the case of the findings in this paper, it would be the staff's responsibility to make a decision, while considering the insight of the AI.

It is important to remember that a predictive model is just a tool like other measurements, meaning that if a decision is to be made using different tools, both the doctor and the tools should be mostly in agreement. Of course, the developers of the AI can be held accountable if a major flaw leads to several false predictions, but one must remember to consider the inevitable probability that the model prediction is wrong. A scenario where the developers should be held accountable is if the model greatly exceeds the expected probability of false predictions, leading to many more mistakes.

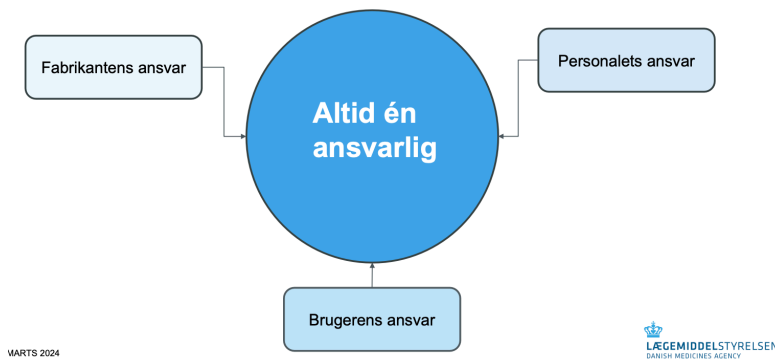


Figure 4.1: A diagram having "Always one responsible" as the center piece, with "The manufacturer", "The staff" and "The user" pointing towards it

4.3 Perspectives on future research

Given the values drawn from the ROC AUC and PR AUC curve of each model, it is noticeable that the prediction of mortality and organ failure has the potential to be more precise and useful. To obtain this, a much more balanced dataset is required.

If the mentioned ML and DL models can be deemed successful, then potentially more positive data points need to be collected. This will ultimately prove if it is possible for the models to consistently predict the correct outcome and keep the same ROC AUC. This is mostly a problem solved by time and more data being collected, it is also possible that during the preprocessing and preparation of the dataset, a significant amount of positive observations got filtered out. This would at least be worth investigating since each positive observation counts.

Furthermore, another shortcoming of this research paper is the lack of depth on the subject of different groups influencing the models. Simply looking at the effectiveness of the models on different groups is not enough to determine true patterns and possible correlations. It would be relevant to look at the proportion of positive class observations and a specific group. Conducting several statistical tests to determine if a significant difference could be found between the proportions of these groups would also be important. In addition to this, it would be relevant to investigate the influence that the variables may have on one another. It seems likely that the combination of diagnoses would have an influence on measurements like lung health or several laboratory results.

Similarly to the investigation of the groups, researching the different time categories and the influence of when the data on the outcome is collected would be highly informative. This could both improve the precision of the models along with being used as a real-time tool for determining the patients condition during the operation and afterward. This product could potentially help improve the surgeons' decision making during an individual's cardiac surgery and understanding of the process of it.

5 Conclusion

To conclude, the primary objective of this study was to assess the risk of mortality for patients who have undergone cardiac surgery, especially to investigate short-term mortality (set to be within 30 to 90 days) and long-term (within a year) post-surgery. Additionally, the aim was also to predict the risk of kidney failure, addressing non-mortal complications to try and predict the time spent in the ICU.

However, based on the numerous different problems regarding the data that were touched upon in the discussions section, the paper cannot conclude to what extent an ML or DL model can predict either short- or long-term mortality or morbidity of a patient who might undergo cardiac surgery. The AI models that were trained (LR, RF, GBM and MLP), simply, do not perform well when judged by the PR AUC, which is the unbiased metric when compared to ROC AUC, as PR AUC focuses on the minority class (ROC AUC does not prefer one class over the other, so it is better for balanced datasets). The imbalance of the data must have been a greater focus in the paper if meaningful models were to be achieved. Resampling methods should have been considered, and algorithms that allow for tuning of class weights should have been further investigated.

Furthermore, the size of the data has been a great handicap as well, limiting the size of possible Neural Networks that could be trained as well as negatively impacting training in general due to the lack of positive entries. The small size of the dataset was the result of both a lackluster of positive classes and an extensive cleaning of the raw data, where many patients were removed from the data as their documentation was quite unclear. Furthermore, patients who have died under surgery were also not considered, as it was difficult to assess which measurements to use for this type of patient. As for the feature candidates such as *lab* results and *EKG* results, the ones that were recommended by doctors simply did not have enough recordings for all patients in the population, resulting in a decision to drop those from the dataset. Some of these features might have held great information for the models and would have resulted in a potentially higher performance, but having even less data entries could not be afforded. Other important features such as data recorded on the heart vessels, also had to be dropped due to the messy nature of the data labels. Extensive statistical analysis was not done on the different social groups of patients as this was deemed not meaningful as long as the models failed. That is also why the models were not compared against each other. For that same reason, it could not be concluded whether different attributes such as age or gender play a factor in how well ML and DL models perform.

A Appendix

A.1 Pictures from slide

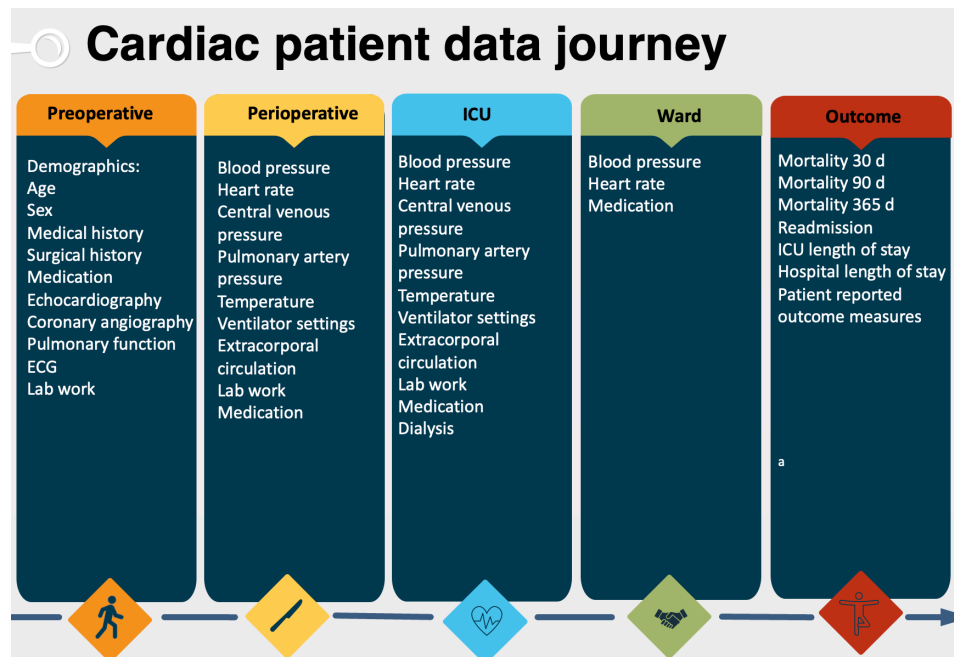


Figure A.1: Cardiac patient data journey (27)

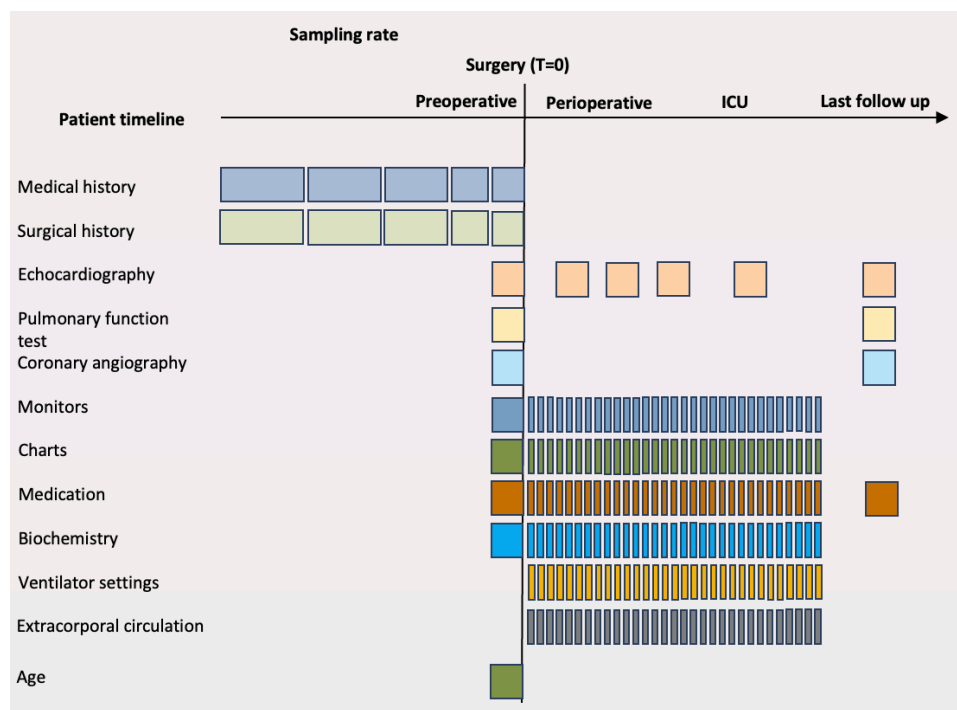


Figure A.2: Sampling rate (27)

Technical
University of
Denmark

DTU Compute Bygning 324
2800 Kgs. Lyngby
Tlf. 4525 1700

<https://www.dtu.dk/uddannelse/bachelor/uddannelsesretninger/kunstig-intelligens-og-data>