Prediction of VA ECMO weaning success

Supervised Machine Learning models for prediction of weaning success in Veno-arterial Extracorporeal Membrane Oxygenation based on continuous hemodynamic parameters

Catharina Jansen

Master Thesis Technical Medicine

17 Oktober 2024

Educational institution
University of Twente
Science and Technology Faculty (TNW)
Master Technical Medicine - Medical Sensing and Stimulation

Clinical institution
University Medical Centre Utrecht
Department Vitale Functies
Intensive Care Unit

Supervisors

Chair and medical supervisor
Prof. Dr. D.W. Donker
Cardiovascular and Respiratory Physiology, University of Twente, Enschede, The Netherlands
Intensive Care Centre, University Medical Centre Utrecht, Utrecht, The Netherlands

Medical supervisor
Drs. J.A.J.M. Hermens
Intensive Care Centre, University Medical Centre Utrecht, Utrecht, The Netherlands

Technical and Technical-medical supervisor
Dr. L.M. van Loon
Cardiovascular and Respiratory Physiology, University of Twente, Enschede, The Netherlands

Process supervisor
Drs. A.G. Lovink
University of Twente, Enschede, The Netherlands

External member graduation committee
Dr. ir. J.M. Rivera Arbelaez
University of Twente, Enschede, The Netherlands

Daily supervisor
ir. L.P. Veerhoek
Digital Health, University Medical Centre Utrecht, Utrecht, The Netherlands

Abstract

Veno-arterial extracorporeal membrane oxygenation (VA ECMO) provides temporary cardiac support during cardiogenic shock but is resource-intensive and associated with significant mortality. This study aims to develop a machine learning model to predict weaning success using continuous hemodynamic parameters early in the VA ECMO treatment. The parameters used to predict weaning success include heart rate, pulse pressure, mean arterial pressure, central venous pressure, vasoactive inotropic score and ECMO flow from the first three days of the ECMO run. Additionally, age, gender, and lactate levels were considered. Weaning success was defined by ICU mortality, VAD implementation, heart transplantation, and the need for a second weaning trial. A total of 108 ECMO runs were included. Features were extracted by calculating the slope, standard deviation, and mean for each day and for the full three-day period. Machine learning models, specifically random forest, KNN, logistic regression, gradient boosting, and support vector machine, were developed and compared based on RMSE, MAE, and R^2. KNN and random forest models showed the best results, with RMSE of 0.48 and 0.48, MAE of 0.45 and 0.42, and R^2 of 0.07 and -0.04, respectively. These results are insufficient for implementation, likely due to the non-predictive nature of the features used. The findings point to the need to reconsider the input data and model design to improve the prediction accuracy of VA ECMO weaning outcomes.

Table of Contents

List of Abbreviations

ABPd	Arterial blood pressure, diastolic
ABPs	Arterial blood pressure, systolic
ANOVA	Analysis of Variance
CVD	Central venous pressure
ECLS	Extracorporeal life support
GB	Gradient boosting
HR	Heart rate
ICU	Intensive care unit
KNN	K-nearest neighbours
LR	Logistic regression
MCS	Mechanical circulatory support
MAE	Mean absolute error
MAP	Mean arterial pressure
PP	Pulse pressure
RF	Random forest regressor
RMSE	Root mean squared error
SHAP	Shapley Additive Explainer
SFS	Sequential forward selection
SVR	Support vector regressor
VA ECMO	Veno-arterial extracorporeal membrane oxygenation
VAD	Ventricular assistive device
VIF	Variance inflation factor
VIS	Vasoactive inotropic score

1. Introduction

Cardiogenic shock is a severe medical condition where the heart cannot provide enough cardiac output to meet metabolic demands of the body. This results in end-organ hypoperfusion, often due to myocardial infarction, severe heart failure, or cardiomyopathy [1]. Cardiogenic shock accounts for 6-10% of intensive care unit (ICU) admissions and has a mortality of roughly 60% [1]. In cases where the condition cannot be managed with other therapies such as medication or revascularisation, veno-arterial extracorporeal membrane oxygenation (VA ECMO) can be considered. VA ECMO is a lifesaving technology that provides temporary cardiac and circulatory support in patients with cardiogenic shock [2]. VA ECMO ensures continuous organ perfusion to allow the patient time to stabilise and facilitate cardiac recovery, provide time for decision making or bridge time to heart transplantation or ventricular assistive device (VAD) [3]. During treatment of cardiogenic shock with VA ECMO, physicians have to find out if the heart can recover sufficiently for the patient to survive without mechanical support, see Figure 1.1. This will determine if the patient can be weaned, decannulated and potentially recover or if alternative treatments such as heart transplantation or ventricular assist devices (VADs) need to be considered. If these options are not viable, the focus will shift to removing the ECMO and transitioning to palliative care.

Despite the increasing use and understanding of the procedure, VA ECMO therapy remains linked to significant mortality rates and complications [2]. In the UMC Utrecht, the mortality of ECMO patients was 56% over the past twelve years and in 57% of the cases weaning was not possible [4]. The technique also requires considerable financial and human resources [5]. Therefore, physicians aim to differentiate as early as possible between patient groups who can recover without mechanical support and those who cannot. Early prognosis of the feasibility of weaning can ensure efficient use of ECMO equipment and related human resources, reduce the potential for severe complications, and identify patients unlikely to have successful weaning trials [6, 7]. This is a significant percentage of the VA ECMO patients. In the UMC Utrecht, 11% of the VA ECMO patients require either a VAD or heart transplantation [4]. 21% of the patients who were weaned did not survive until hospital discharge, even though they were deemed ready for weaning [4].

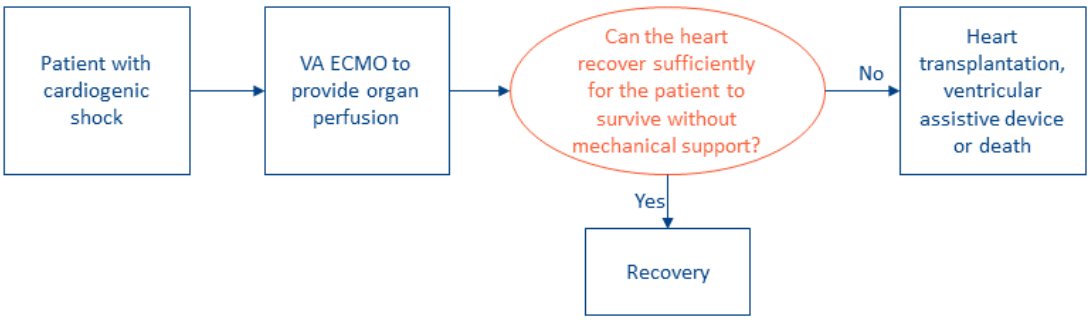


Figure 1.1: Flowchart of a simplified version of the clinical problem: After a patient with cardiogenic shock is placed on VA ECMO, the heart's potential for recovery is evaluated. If it can be concluded that sufficient myocardial recovery occurs, the patient can be weaned and recover. When the patient does not show potential for recovery of the native heart alternatives like heart transplantation (HTX) or a ventricular assist device (VAD) are considered. If these are not possible, the patient will be supported with palliative care.

Currently, readiness for weaning is clinically assessed by evaluating the available individual clinical, hemodynamic and echocardiographic parameters on a day-to-day basis, although accurate prediction of weaning success in the early stages of the VA ECMO treatment is not yet possible [8, 9]. However, exploration of prognostic parameters from the early stage of the ECMO treatment might be able to identify these patients requiring consideration of alternative solutions, such as heart transplantation or VAD therapy.

For early prediction of the cardiac status of the patient and thus early prediction of weaning success, machine learning can be a promising option. It has the potential to use multiple parameters and also incorporate the trends of clinical and hemodynamic parameters over time to predict weaning success. Therefore, it has the potential to help physicians determine for which patients the VA ECMO treatment can lead to recovery and for which patient the VA ECMO treatment cannot. Machine learning is the development of algorithms and models that enable computers to learn from data and make predictions or decisions without being explicitly programmed. It can generalise from examples and discover hidden patterns within large datasets that are not visible using normal statistics

Machine learning models have shown promising results in predicting hospital mortality in VA ECMO supported patients [10, 11, 12, 13]. However, none of these studies have distilled predictors for weaning success in the initial phase of the VA ECMO treatment. Additionally, these models are designed to be used for predicting survival prior to VA ECMO implantation, therefore these results cannot be extrapolated to patients already receiving VA ECMO support. Furthermore, these models do not consider the continuous interaction of the VA ECMO circuit with the native heart and circulation. In the ICU, continuous monitoring of hemodynamic status, respiratory status and organ perfusion provide a multitude of data [14]. These data hold substantial information regarding ongoing changes within the patients hemodynamics and are particularly suitable for machine learning models, as they may reveal intricate patterns underlying the hemodynamic responses.

To address the need for a tool that facilitates early prognostication in VA ECMO supported cardiogenic shock patients, this study aimed to develop a regression model to predict the probability of successful weaning. This information could subsequently be utilised in decision-making processes during the VA ECMO treatment. The objective of this study is to develop a model that predicts whether cardiac recovery during VA ECMO support will be sufficient to provide adequate circulation and oxygenation after VA ECMO support is weaned using continuous hemodynamic parameters.

2. Background

In the background section, more details will be provided about the clinical course of VA ECMO treatment and the machine learning models used in similar studies. Furthermore, the specific ML models to be employed will be discussed, along with the rationale for its selection.

2.1 Clinical course of VA ECMO

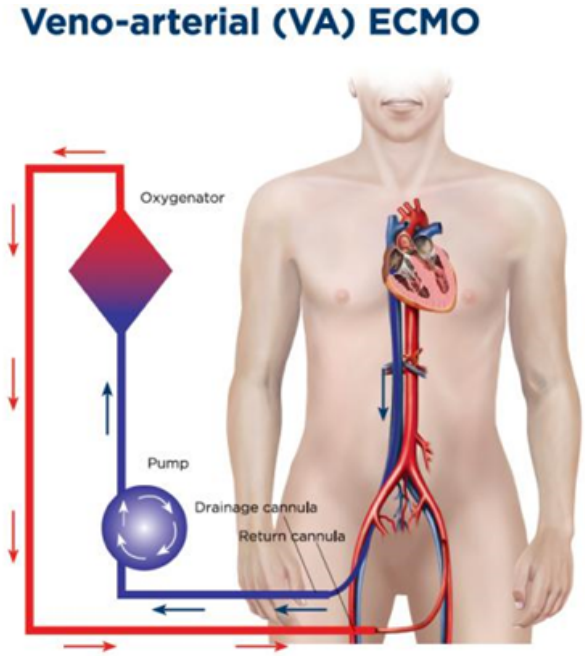


Figure 2.1: Schematic overview of venoarterial Extracorporeal Machine Oxygenation. In this configuration, blood is drained via a cannula in the v. femoralis, oxygenated and pumped back into the bloodstream in the a. femoralis. From "Veno-arterial ECMO" by Scott et al. [15].

ECMO can provide patients time to stabilise and facilitate organ recovery without relying on the heart to oxygenate the body. ECMO involves diverting blood from the patient's body to an external membrane oxygenator, which removes carbon dioxide and adds oxygen before returning the blood to the body, effectively bypassing the heart and lungs. Venoarterial-ECMO (VA ECMO) refers to a specific configuration in which one cannula, typically inserted into the femoral vein (v. femoralis), withdraws blood from the body and sends it through an artificial membrane oxygenator. The oxygenated blood is then pumped back into the patient's body through another cannula, typically inserted into the femoral artery (a. femoralis), as illustrated in Figure 2.1 [15]. The blood is pumped into the aorta under high pressure against the native flow of the heart. While life-saving, VA ECMO patients can suffer from a large amount of complications such as thrombotic and bleeding complications, infectious complications and left ventricular overloading [2]. These complications can result in death and permanent damage to the heart and other organs. Left ventricular overloading can occur when the left ventricle struggles to pump effectively against the elevated afterload resulting from the flow delivered by the ECMO

cannula. The dilation of the ventricle caused by the overloading can exacerbate heart failure [16].

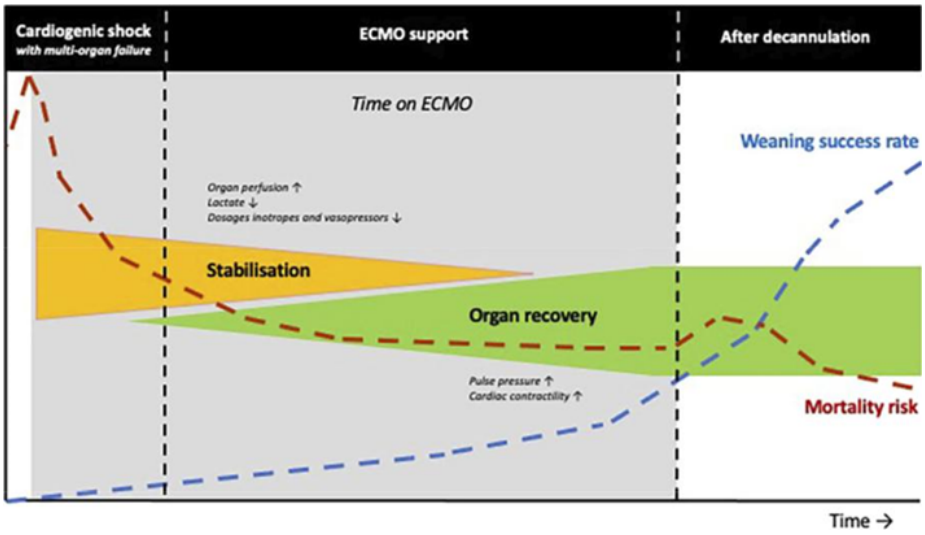


Figure 2.2: Clinical course of ECMO treatment. From 'Patient-centered weaning from venoarterial extracorporeal membrane oxygenation: "A practice-oriented narrative review of literature"' by Hermens et al. [9]

The clinical course of the VA ECMO treatment can be divided in a stabilisation and organ recovery phase, which overlap, see Figure 2.2. During the stabilisation, the organ perfusion is increased due to the ECMO support. Inadequate circulation caused by the cardiogenic shock results in an increased lactate. During ECMO support, the organ perfusion is increased and the lactate will decrease.

In the organ recovery phase, the goal is to provide organ perfusion to allow the patient to survive to facilitate time to make cardiac recovery possible. When cardiac recovery occurs, the myocardial contractility will increase over time during ECMO support. At the bedside, this can be assessed by a visual improvement in echocardiographic parameters and often an increase in pulse pressure, which is the difference between the systolic arterial pressure and diastolic arterial pressure [17]. If stabilisation in the organ perfusion phase is achieved and myocardial recovery is detected, reducing ECMO support is essential with ECMO extraction as an ultimate goal. Readiness for weaning is determined by evaluating the hemodynamic parameters, cardiac ultrasound and clinical presentation of the patient. In this study, weaning is defined as the process of reducing and eventually discontinuing ECMO support once a patient's heart or lungs have recovered enough to function independently.

First, VA EMCO flow will be reduced gradually. If ECMO flow is reduced to 2 L/min and the hemodynamic status of the patient remains stable, pulmonary function has improved, and end-organ failure is largely resolved, a weaning trial can be attempted. In a weaning trial the ECMO is reduced further to a minimal flow of 0.5 – 1 L/min. During a weaning trial, cardiac function is monitored by multiple hemodynamic, echocardiographic parameters. These are used to judge if the patient tolerates the weaning trial and thus if the weaning trial was successful [9]. After a successful weaning trial, decannulation can be scheduled. In case of a failed weaning trial, the ECMO flow will be increased to ensure adequate perfusion and weaning trial can be reattempted after at least another 48 hours. When weaning is not possible, heart transplantation, VAD or palliative care are the remaining options [3].

2.2 Literature review of machine learning models

Multiple studies have already attempted to use machine learning to predict outcome in ECMO patients. [10, 11, 12, 13, 18, 19]. Aim of these studies is both to provide support in deciding which patients are good ECMO candidates, and to provide support during the ECMO treatment. These studies are relevant to this research as they focus on predicting outcomes of VA ECMO. To identify promising machine learning models for this study and review previous work, the following sections elaborate on and compare these studies.

In the study of Stephens et al. [10] multiple models (logistic regression, gradient boosting, support vector machine, neural networks and adaboost ensemble) were trained to predict hospital mortality in VA ECMO patients to support physicians in decision making of the implementation of VA EMCO. The features that were used to predict mortality are lactate, blood pressure and breathing rate as well as information about the type of surgery done and intubation time six hours prior to implantation of the ECMO. They found lactate and age to be important features for mortality. The Neural Networks model performed best with a sensitivity of 82% and a precision of 78%.

Wang et al. [11] used multiple logistic regression to predict hospital mortality for patients whom had undergone a CABG operation and needed VA-EMCO for cardiogenic shock. They found age, presence of left main artery disease and lab values to be predictive of hospital mortality and achieved a AUC of 0.85. Ayers et al. [12] also developed a neural network model to predict survival to discharge based on lab values, but chose the lab values of the first 48 hours of the EMCO treatment. They achieved an overall accuracy of 82% and an AUC of 0.92. Braun et al. [13] used a conditional interference tree to predict survival to hospital discharge based on various lab values prior to VA ECMO implantation. They achieved an error rate of 35%, with an AUC of 0.71. The same group used a random forest approach to identify which features are predictive of in-hospital death and found urine output to be statistically significant [19]. Xue et al. [18] used Gradient Boosting to provide guidance in ECMO allocation 48 hours prior to implantation. Clinical values such as comorbidities, patient characteristics and medication use were used as features. The result was an AUC of 0.94.

These studies show the potential of predictive modelling to improve ECMO care by providing prognostic tools for physicians. However, none of these studies have used continuous hemodynamic parameters throughout ECMO support to train a model. These are the values weaning currently is based on and which gives the most direct available representation of the hemodynamic status of the patient. Also, only hospital mortality or ICU mortality was taken as outcome parameter. This includes patients who have received a heart transplant or VAD in the successful category, even though their heart has not recovered. Developing a model with continuous parameters and a newly defined criterion for weaning success would be an innovative approach.

2.3 Machine learning models

This study selected five machine learning models for their strong performance in prior studies [10, 20, 21] and suitability for this specific task; K-Nearest Neighbours (KNN) regression, Support Vector Regressor (SVR), Logistic Regression (LR), Random Forest Regressor (RF), and Gradient Boosting (GB). These models were chosen, because they perform well on smaller datasets, can predict probabilities, handle high-dimensional data,

and provide interpretable results regarding feature influence on the prediction, enabling physiological explanations of their predictions.

KNN is effective for small datasets, KNN predicts outcomes based on the majority class of the k nearest neighbours in the feature space [22]. SVR is ideal for high-dimensional data, finding a hyperplane that best separates classes while minimizing errors [22]. LR is a linear model for binary classification that estimates the probability given input belongs to a certain class based on its features. It is useful for interpreting confidence levels but limited to linear relationships [22]. RF is an ensemble method that builds multiple decision trees and averages their predictions [23]. It handles high-dimensional data well and provides feature importance, but may be less easy to interpret due to fact that it is an ensemble method constructed of multiple trees, so the effect of each feature on the outcome is harder to quantify [23]. GB is another ensemble method that sequentially improves decision trees, excelling in accuracy but requiring careful hyperparameter tuning [24]. It performs well even with small sample sizes, but has the same problem with interpretability as RF [21].

3. Method

3.1 Data collection

This study uses the database from a previous study by Meuwese et al. [4] as the initial cohort for extracting data to train the machine learning models. They collected information of patients who underwent ECMO treatment at the Intensive Care Unit of the University Medical Centre Utrecht between 2007 and 2022. This information includes data on demographics, comorbidities, EMCO indications, ECMO set-up, reason for discontinuation of the EMCO trial, complications and mortality over a 12-month follow-up period.

3.1.1 In- and exclusion criteria

From the database of Meuwese et al. [4], adult patients with a VA ECMO configuration were included. Exclusion criteria for the current study were a non-cardiac indication for ECMO, presence of mechanical circulatory support (MCS) prior to or during EMCO treatment and EMCO support duration <72 hours (Table 3.1). Patients with a non-cardiac indication for VA ECMO support have a different recovery pattern than patients who receive VA EMCO support for cardiogenic shock, therefore these patients were excluded from this study [14]. The presence of MCS prior to and during ECMO treatment would artificially affect hemodynamic parameters, thus prohibiting assessment of native cardiac function, so these patients were also excluded from analysis [25, 26]. The goal is to predict weaning success based on parameters measured during the first three days of ECMO support. Therefore, a dataset with the full first 72 hours was necessary to train the model and patients who were less than 72 hours on ECMO were excluded.

Inclusion	**Exclusion**
Age $\geq$ 18 years	Non-cardiac indication for VA ECMO
VA ECMO config.	Ventricular assistive devices present during ECMO implantation
	$<$ 72 hours on ECMO

Table 3.1: In- and exclusion criteria.

3.1.2 Input parameters

The data collected for this study included patient demographics and parameters to determine outcome from the database of Meuwese et al. [4]. The demographic data that was collected are age at the time of ECMO implantation and gender. Outcome parameters included ICU mortality, reason for ECMO discontinuation, and occurrence of a second ECMO trial, see appendix A.

The continuous hemodynamic parameters used to train the models are mean arterial pressure (MAP), pulse pressure (PP), lactate levels, heart rate, central venous pressure (CVD), VIS, and VA ECMO flow, as detailed in Table 3.2. The parameters were chosen as they reflect the hemodynamic status of the patient, specifically the functioning of the

heart during ECMO treatment. PP indicates the contribution of the native heart to the blood pressure, given that ECMO flow provides only non-pulsatile pressure. PP is a marker for cardiac dysfunction in VA ECMO, especially in cases of cardiogenic shock [14]. MAP and heart rate can reflect hemodynamic stability [8]. The amount of vasoactive and inotropic medication, represented by the vasoactive inotropic score (VIS) and VA ECMO flow indicate the level of support the patient requires [27]. The central venous pressure (CVD) can be used to assess preload and volume status. Additionally, lactate levels have been identified as important features in previous studies and were also be included in the model [10, 11].

Parameter	Unit	Frequency of measurement
Pulse pressure (ABPs-ABPd)	mmHg	/min
Mean arterial pressure	mmHg	/min
Heart rate	/min	/min
Central venous pressure	mmHg	/min
Vasoactive inotropic score (VIS)	-	/min
Flow from ECMO	L/min	/hour
Lactate	Mmol/L	/day

Table 3.2: Hemodynamic parameters with unit and frequency of recorded measurements.

Hemodynamic parameters for the included patients were retrieved separately from the electronic health record Metavision as these were not collected by Meuwese et al. PP was not directly available from Metavision and was calculated by subtracting the diastolic arterial blood pressure (ABPd) from the systolic arterial blood pressure (ABPs) [17]. The VIS was calculated using Equation 3.1 and was determined for every minute of the ECMO trial.

The parameters were collected during the first three days of ECMO support. The early phase, just after ECMO initiation, provides valuable insights into the patient's initial condition during shock and the rate of circulatory stabilisation and organ recovery. Utilising data from the first days of the VA ECMO treatment also enables the model to make predictions after the first days. The training period for the data is a trade-off between more extensive information on a patient from a longer period and the ability to make early predictions from a shorter period. Additionally, a longer time period reduces the number of patients available for inclusion. A three-day period was selected because it reveals significant differences in pulse pressures between successfully and unsuccessfully weaned patients and includes a sufficient number of patients [17, 25].

$$\begin{aligned} \text{VIS} \quad = \quad & \text{Dobutamine dose}(\mu g/kg/min) + 100 \cdot \text{Epinephrine dose}(\mu g/kg/min) \\ & + 10 \cdot \text{Milrinone dose}(\mu g/kg/min) + 10000 \cdot \text{Vasopressin dose}(units/kg/min) \\ & + 100 \cdot \text{Norepinephrine dose}(\mu g/kg/min) \quad [27] \end{aligned} \tag{3.1}$$

3.2 Weaning success

The model aimed to predict whether a the native heart function improves sufficiently during ECMO treatment to enable successful weaning. To develop and train a machine learning model, it is essential to define weaning success in order to categorise the input

data on an outcome. Within literature, various definitions for weaning success have been described including "alive without mechanical cardiac support for 48 hours, 30 days after ECMO extraction or at ICU discharge" [9]. Also, different survival metrics are used in similar studies to determine weaning success; 30-day mortality, in hospital mortality and ICU mortality are all used [9, 10, 11, 28].

In this study, weaning was considered successful when; the patient survived the ICU admission, without the need for renewed mechanical support (temporary or permanent) of a heart transplant within 30 days. ICU mortality was chosen with the intention to exclude all causes not directly related to ECMO weaning and extraction [7]. The first ECMO run was classified as a weaning failure if a second ECMO run was performed within 30 days [9]. The second run was classified based on the remaining conditions. The occurrence of a heart transplantation or VAD implantation was determined by looking at the reason for discontinuation of ECMO. These outcome parameters were recorded up until the last follow up of the study, which was at 12 months after initiation of ECMO support. The exact parameters and the corresponding options can be found in Appendix A.

3.3 Data preparation

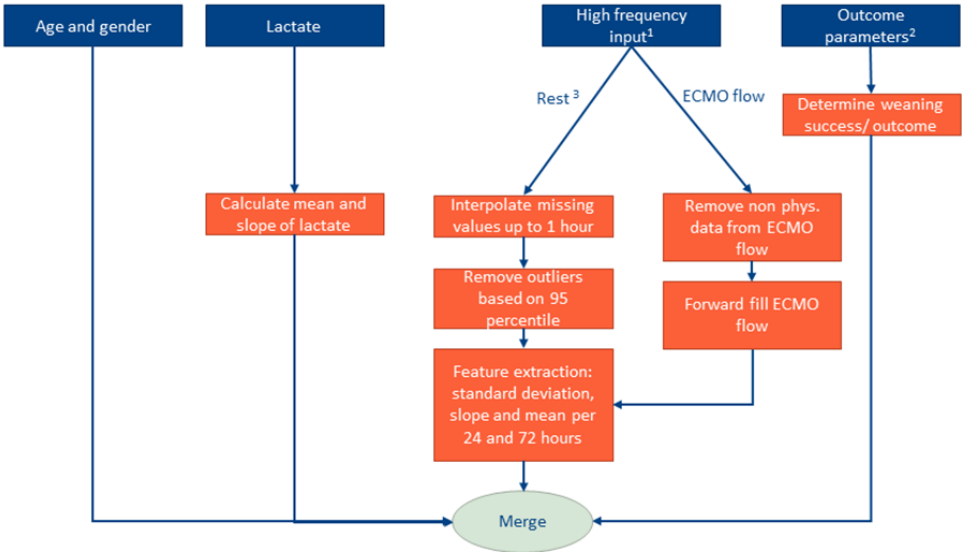


Figure 3.1: Flowchart of data preparation and feature selection. Parameters are divided in four groups based on the different data preparation methods. 1) High frequency input includes heart rate, PP, MAP, VIS, EMCO flow and central venous pressure. 2) Outcome parameters are the parameters used to determine weaning success. These include reason for discontinuation, cause of death and occurrence of a second weaning trial. 3) Rest: includes heart rate, PP, MAP, VIS and central venous pressure.

An overview of the data preparation and feature selection steps can be found in Figure 3.1. It shows the different methods that have been applied to each of the different parameter categories. The parameters are categorised in high frequency input (heart rate, PP, MAP, VIS, EMCO flow and central venous pressure), lactate, age and gender, and the outcome parameters on which the weaning success is based. These categories were created as the parameters in these categories are processed differently from each other, as the frequency and nature of the data requires a different method of data preparation.

3.3.1 Missing values

Missing values were replaced using linear interpolation. Linear interpolation estimates values between two data points by constructing a linear equation that connects these points [29]. Interpolation was performed on periods with missing data up to one hour. ECMO trials with more than four hours of missing data on one of the high frequency (/min) parameters were excluded, because this would cause serious error on the features as more than 15% of data on which the features are based would be missing. Age, gender, lactate and the outcome parameters did not have missing values.

3.3.2 Outliers

All hemodynamic parameters, except for lactate, were tested on having a normal distribution. A Shapiro Wilk test was performed on these six values and found a non-normal distribution [30]. This was also found based on visual inspection of Q-Q plots, see appendix B. Based on the Q-Q plots there seems to be a skewed distribution. For this type of distribution a non-parametric outlier removal based on an interquartile rang is a robust method and can also be used on future data without needing to adhere to strict assumptions [31]. Outliers of MAP, PP, CVD, HR and VIS were removed based on a quantile range. Datapoints above the 97.5 percentile and below the 2.5 percentile were removed, based on the entire included population per parameter.

3.3.3 Flow from ECMO

The parameter flow from the VA ECMO was not subjected to removing outliers. The flow is manually registered by nurses directly read from the machine and is therefore less likely to be subjected to noise caused by faulty equipment. Upon inspection of the data some values were non-physiological ($>$ 7L/min [15]) and seemed to be entries of the flow in rounds per minute instead of litre per minute. These datapoints were removed by replacing all values above 7 l/min with a NaN value. The flow from the ECMO was forward filled to /min instead of /hour, by assuming the flow stayed the same until a new value was entered.

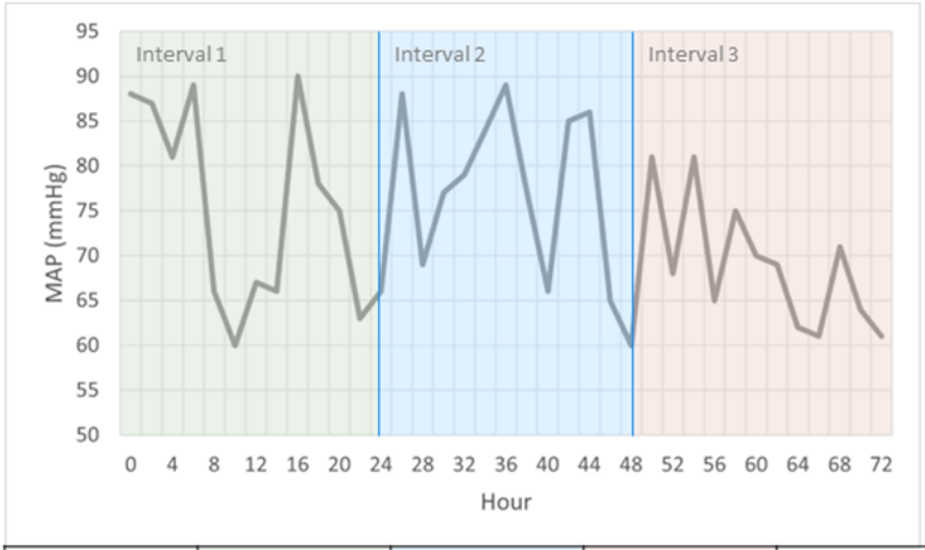
3.3.4 Normalising data

The values of the parameters where normalised to a value between 0 and 1, based on the values of the entire population. This is called min-max scaling [32]. The goal is to equalise the features and facilitate the identification of the best performance of the algorithm. The upper and lower boundary will also stay in place after implementing the model, when inputting the data on which to predict. Therefore, it must be assumed that the population used to train the model is representative of the general population of individuals on VA ECMO for future predictions.

3.4 Feature extraction

A feature extraction method was chosen to convert continuous data into interpretable features while preserving as much of the information on variability and trends of the original

data as possible. For this, a bag of features method was used to gather features from the high frequency hemodynamic parameters, see Figure 3.2 [33]. This method calculates the mean, standard deviation and slope of the data in a specified interval. Additionally, the bag of features method also provides the mean, standard deviation and slope of the entire original interval. The slope is calculated by first fitting a least squares regression line on the data of the specified time interval and then finding the coefficient of this line, this is then the slope. The method implemented based on a 24 hour interval and thus splitting the entire interval in three sub-intervals.



	Interval 1	Interval 2	Interval 3	Complete interval
Mean	75.1	77.1	69.0	73.8
Standard deviation	10.6	9.5	6.8	9.7
Slope	-0.70	-0.55	-0.67	-0.17

Figure 3.2: Visual representation of the bag of features method on example data of the MAP of one ECMO run. The 72 hour run is split in three intervals. For each of the pictured interval the mean, slope and standard deviation was calculated. This is also done for the complete 72 hours. The resulting features are shown in the table below the graph.

The bag of features method can show the change of a parameter between the intervals, by looking at difference in mean, as well as within the intervals by looking at the slope. The standard deviation for each time interval shows the amount of change occurring within each interval. Therefore, it is an effective method to capture the time series comprehensively while converting it into features that can be used to train the model.

Lactate was treated separately, because this value is recorded only once a day. The mean and slope from the available values were calculated as features. This gives a good representation of the trend as well as the absolute values of lactate. Age and gender were also included as features. This resulted in 12 features for every high frequency hemodynamic parameter, two for lactate and one for age and gender. Thus, 76 features were extracted in total.

3.5 Feature exploration

In order to test whether the features have a significant impact on the outcome, weaning success, a one way ANOVA test was performed on each of the features. A Bonferroni

correction was implemented on the one way ANOVA [34]. This correction is done, because of the increased risk of false positive results when doing multiple statistical tests.

To better understand the relationship between the features and the outcome, a correlation matrix was calculated based on the Pearson correlation coefficient. This calculates the linear correlation between the features and the weaning success on a scale from -1 to 1 with 1 signifying a high positive correlation [35]. This shows a measure of the relationship between features and outcome, positive or negative, and which features have a high co-linearity. The specific relationship of between some features was visualised using scatterplots.

3.6 Dealing with Multi-collinearity

Multi-collinearity occurs when two or more predictor variables in a multiple regression model are highly correlated. This can be problematic for several reasons [36]. Highly correlated variables provide redundant information about the outcome variable. It can become difficult to assess the individual effect of each feature on the outcome as multicollinearity means two variables hold the same information and therefore have a similar effect on the prediction. The primary issue arises from the chosen method of feature selection, which relies on assessing whether removing or altering features affects the model's performance. However, if two features essentially contain the same information, the performance will not differ significantly if one is removed. This can result in falsely low importance values for highly correlated features [37].

Several methods can be used to deal with multicollinearity, such as removing the highest correlated features, combining variables or using techniques that can account for multicollinearity [36, 34]. To keep the interpretability of the features, the first method was chosen.

The variance inflation factor (VIF) is a measure that most directly can quantify multicollinearity of features [36]. It measures how much of the variance is inflated due to multicollinearity. The multi-collinear features were removed by selecting the feature with the highest VIF. After the removal of that feature, the VIF was recalculated for the remaining features. This process was iterated until all VIF values were below 10, which is a threshold indicating high multi-collinearity [36]. A total of 37 features were removed using this technique. A list of the removed features can be found in Appendix G.

3.7 Model selection

The goal of the model is to predict the probability of a successful weaning. Therefore, regression was chosen. Regression models can be used to predict weaning success in a percentage , i.e. patient X is Y percent likely to wean successfully, in contrast to a classification model that can be used to predict the weaning success in a binary matter. Models such as logistic regression, random forest, gradient boosting, support vector regression and deep learning have been shown to have good results in comparable studies [10, 19, 28]. Another criterion for model selection is the need to interpret the used features, as this gives an opportunity to physiologically explain the model. While deep learning is widely adopted and has shown high accuracy, it is challenging to interpret the features as it prioritises for decision-making [38]. This lack of transparency makes it less preferable, as

the exploration of the importance of the included features is intended to be part of the results. Accounting for the points mentioned and which model had good results in similar studies [10, 11, 12, 13, 18, 21, 28], the following supervised models were chosen: k-nearest neighbours (KNN), logistic regression, support vector regression, gradient boosting and random forest regression [33]. All models were implemented using the scikit-learn library in Python (version 3.8.10) [39].

3.8 Feature selection

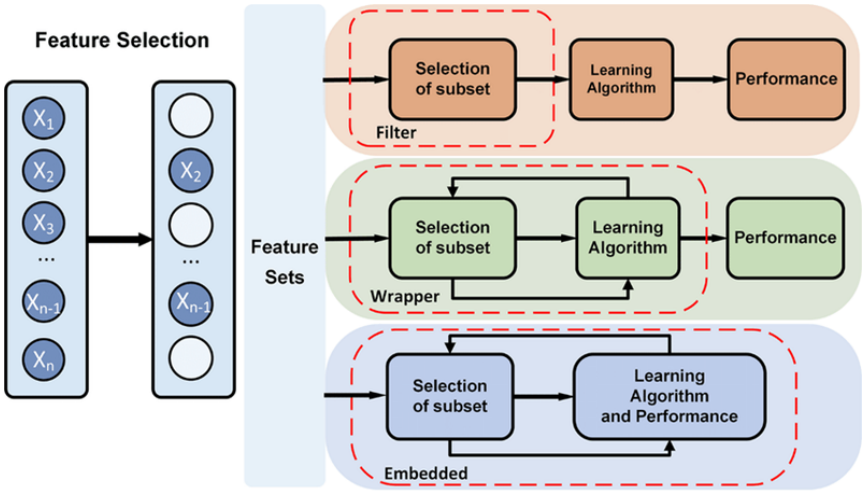


Figure 3.3: Visual representation of different feature selection methods. The filter method can include both univariate filtering and multivariate filtering. The main feature of this method is the selection prior to implementing the learning algorithm. The wrapper method selects features based on which subset of feature performs best. The embedded method uses an integrated feature importance to select the optimal subset. From "Computational Diagnostic Techniques for Electrocardiogram Signal Analysis" by Xie et al.[40].

Selecting the most relevant features is crucial when dealing with a large number of variables, as it can enhance model performance and reduce the risk of overfitting [15]. Feature selection reduces data dimensionality by identifying features relevant to predicting the outcome. Several methods exist for feature selection, categorised into filtering methods, wrapper methods, and embedded methods (see Figure 3.3). Filtering methods select a subset based on each feature's significance in relation to the outcome, but does not capture model-specific high-performing features. Wrapper methods evaluate various feature subsets based on model performance, iteratively refining the feature set to optimise performance. Embedded methods integrate feature selection within the model algorithm, using feature importance metrics to identify relevant features [41, 42].

Taking this into account, an embedded method (see Figure 3.4) was applied for the Random forest and Gradient Boost algorithms. This approach was selected for its ability to account for both algorithm performance and feature interactions. Random forest and Gradient Boosting have an integrated feature importance calculation based on the Gini importance [17]. This calculates how well a potential split is separating the samples of the two classes and is computed as the normalised total reduction of criterion due to that feature. This importance was calculated for each feature and visualised in a bar graph. After the feature importance was calculated, the RMSE was calculated for each number of included features, starting with one feature (the most important) and progressively

including additional features based on their importance. This way the highest performing amount of features and thus subset was calculated.

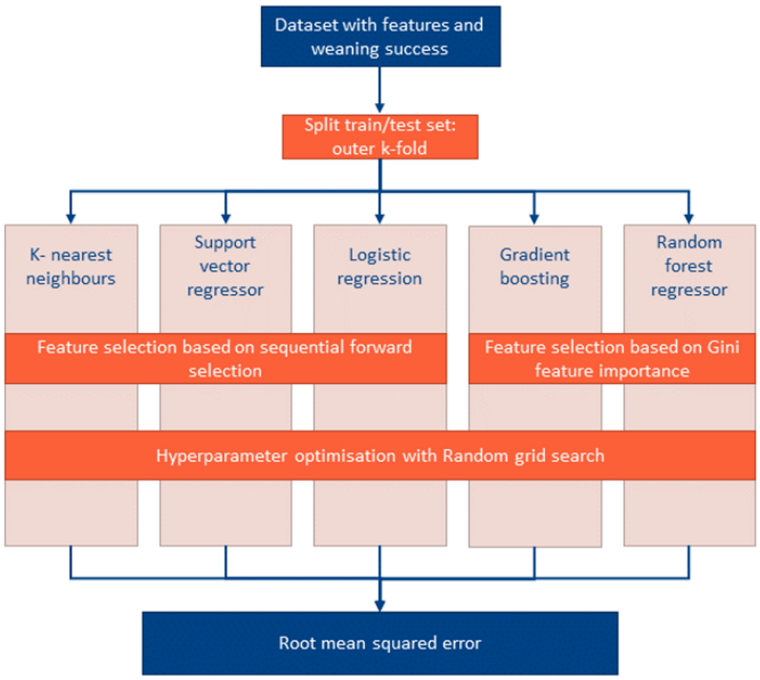


Figure 3.4: Flow chart of model training. With SFS feature selection for KNN, SVR and logistic regression and embedded feature selection in GB and RF. Outer k-fold with k=5 is used to determine the performance of each model.

To select the features for KNN, SVM and logistic regression, a wrapper method was implemented, specifically sequential forward selection (SFS) [43]. This wrapper was chosen, because it evaluates feature subsets based on model performance and considers feature interactions. Starting with an empty set, SFS iteratively adds features until the optimal subset is identified based on RMSE. Given the initial high feature-to-data point ratio (36 features to 109 data points), SFS was preferred over sequential backward selection, which begins with all features [44].

3.9 Model optimisation

3.9.1 Nested K-fold

To optimally train the model, a nested k-fold cross-validation method was implemented. This technique facilitates performance estimation of the final model and aids in hyperparameter optimisation. Cross-validation estimates a model's performance on unseen data by partitioning the dataset into k equal segments [41]. In this study, a 5-fold cross-validation was employed to ensure a 80/20 train test split and enough datapoints in the test group [22]. The dataset is divided into five equal parts, with the model trained on four parts and tested on the remaining one. This procedure is repeated five times, and the average performance across all test sets is computed as the final outcome.

Nested k-fold cross-validation involves both an inner and outer loop of cross-validation. Each of the k training segments is further subdivided into k parts, as depicted in Figure 3.5 [45]. The inner folds are utilised for hyperparameter tuning, producing outcomes more

robust to new data and mitigating the risk of overfitting during feature selection. This method ensures that the model performance and tuned parameters are not reliant on any specific data split, as all data points are ultimately used for evaluation.

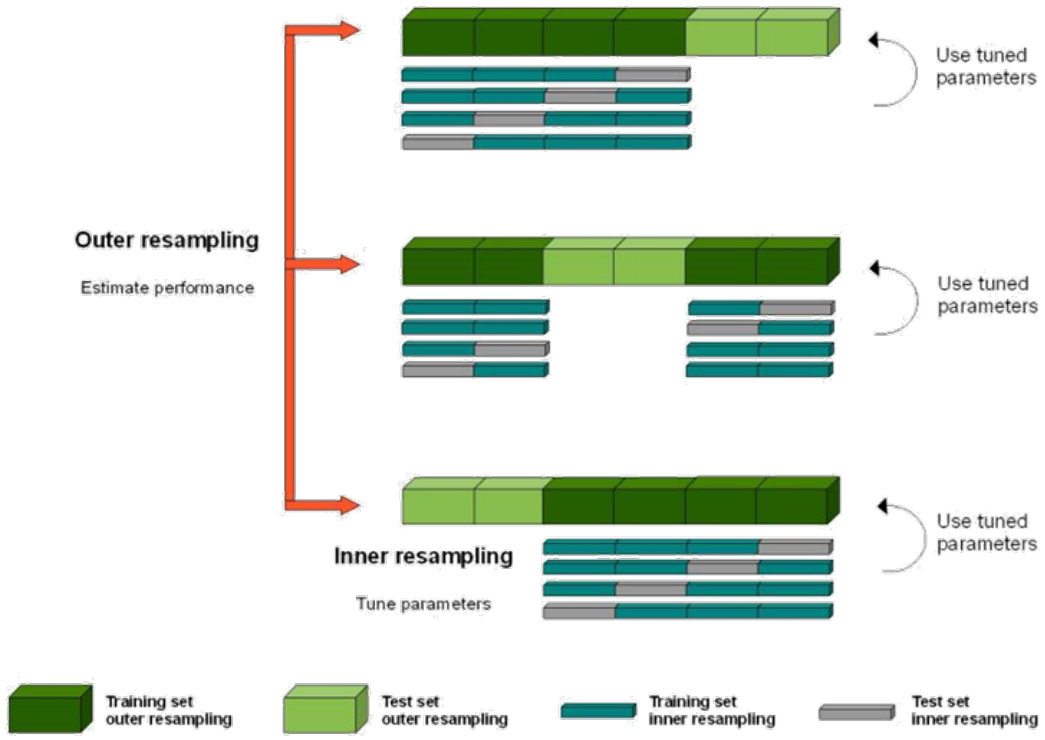


Figure 3.5: Illustration of nested K-fold, when k=3 on the outer fold and k=4 in the inner fold. From "Nested and Repeated Cross Validation for Classification Model with High-Dimensional Data" by Zhong et al. [45].

3.9.2 Hyperparameter tuning

The last step of model optimisation is finding the best hyperparameters for each model. Hyperparameters are 'settings' of the machine learning models. The optimal hyperparameters were found with a grid search. It works by making a grid of all the possible combinations of hyperparameters and testing each of the places in the grid until the optimal combination is found [46]. In order to reduce computation time a random search algorithm with 100 iterations was used for the RF and GB, which randomly tests combinations from the grid search. This was done using the RandomSearchCV function from the sklearn library [39].

For each model a different set of hyperparameters has to be tuned. Below a list of hyperparameters with a short description for each parameter and the chosen value [46]. The used grids and the choice of hyperparameter can be found in appendix C.

Random Forest Regressor

- **N-estimators**: Number of trees in the forest. *N = 1000*
- **Max depth**: Maximum depth of each tree. Deeper trees can capture more complex patterns but may overfit. *Max depth= None*
- **Min samples split**: Minimum number of samples required to split an internal node. This controls the complexity of the model. *Min split = 10*

- **Min samples leaf**: Minimum number of samples required to be at a leaf node. *Min leaf = 1*
- **Max features**: Number of features to consider when looking for the best split. Can be set to a fixed number, a fraction of the total number of features, or left unrestricted. *Max = None*

Support Vector Regressor

- **Kernel**: The function used to map data into a higher-dimensional space. Common kernels include linear, radial basis function (rbf), and polynomial (poly). *Chosen kernel = rbf*
- **C**: Regularization parameter that controls the trade-off between achieving a low error on training data and minimizing the norm of the coefficients. *C = 1000*
- **Gamma**: Kernel coefficient for 'rbf', 'poly', and 'sigmoid' kernels. Determines the influence of a single training example. "auto" uses 1/n_features. *Gamma = auto*
- **Epsilon**: Specifies the epsilon-tube within which no penalty is associated in the training loss function. *Epsilon = 0.4*

Gradient Boosting Regressor

- **N-estimators**: Number of boosting stages to be run. *N = 1000*
- **Learning rate**: Shrinks the contribution of each tree by this value. *Rate = 0.01*
- **Max depth**: Maximum depth of the individual regression estimators. Limits the number of nodes in the tree. *Max depth = 6*
- **Min samples split**: Minimum number of samples required to split an internal node. *Min split = 15*
- **Min samples leaf**: Minimum number of samples required to be at a leaf node. *Min leaf = 4*
- **Max features**: Number of features to consider when looking for the best split. *Max features = None*

K-Nearest Neighbours

- **N-neighbours**: Number of neighbours to use for n-neighbours queries. *N = 5*
- **Weights**: Weight function used in prediction. "Uniform" weights all points equally, while "distance" weights points by the inverse of their distance. *Weights = Distance*

Logistic Regression

- **C**: Inverse of regularization strength; must be a positive float. Smaller values specify stronger regularization. *C = 0.1*
- **Penalty**: Used to specify the norm used in the penalization. 'L1' leads to sparse models and can be useful for feature selection. 'L2' is the standard form of regularization. *Penalty = L1*

3.10 Model evaluation

The models were compared based on the root means squared error (RMSE), the mean absolute error (MAE) and the R^2. The RMSE and MAE both show an average error. However, due to the way they are calculated, see Equation 3.2 and 3.3, the outliers have a bigger impact on the RMSE than on the MAE [47]. The R^2 is a quantification of how well the datapoints are explained by the model and can be very valuable in assessing regression models [48]. A R^2 of 0 means the points are not explained by the model and a value of 1 means all the points are explained by the model. The formula for the R2 can be found in equation 3.4.

$$RMSE = \frac{\sqrt{\sum(y_{actualvalue} - \hat{y_{prediction}})^2}}{N} \tag{3.2}$$

$$MAE = \frac{\sum |(y_{actualvalue} - \hat{y_{prediction}}|)^2}{N} \tag{3.3}$$

$$R^2 = 1 - \frac{\sum(y_{actualvalue} - \hat{y_{prediction}})^2}{\sum(y_{actualvalue} - \overline{y_{meanvalue}})^2} \tag{3.4}$$

Based on these three metrics, the two best performing models were chosen and further compared. For these models, the ROC curve was calculated. Also, classification metrics were calculated based on different thresholds to determine what the result would be if decisions were based on different thresholds. The thresholds ranged from 0.3 to 0.7, as values outside this range did not show differences. The specific metrics that were calculated are sensitivity, specificity, accuracy and precision. For the best performing two models, SHAP (Shapley additive explainer) summary graphs were also created. SHAP works by decomposing the prediction of a model into the sum of each feature's impact and calculating values that represent the contribution to the outcome per feature, which can then be used to understand feature importance and explain the result [49]. This creates insight in whether high or low values of the features predict for weaning success or weaning failure.

4. Results

4.1 Patient Inclusion

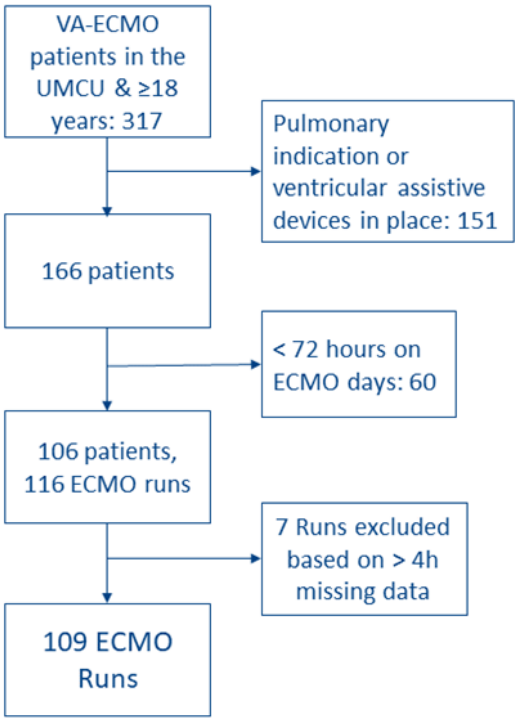


Figure 4.1: Flowchart of the amount included and excluded patients and the subsequent amount of EMCO runs in the final database.

Of the patients in the dataset of Meuwese et al. [4], 317 patients had ECMO treatment with the VA ECMO configuration. All patients were over the age of 18. 151 patients were excluded for having a pulmonary indication for ECMO or had a MCS in place prior to or during ECMO treatment. 60 patients were excluded based on an EMCO run duration of less than 72 hours. This resulted in 106 patients and 119 ECMO runs, see Figure 4.1. An ECMO trial was considered as a separate trial if there was more than 12 hours between weaning from the first trial and start of the second EMCO trial. In 66 EMO runs, there were parameters that had continuous missing values over a period one hour. seven ECMO runs had a continuous period of missing values of over four hours and were therefore excluded.

The characteristics of the 106 included patients are shown in Table 4.1. The distribution of gender as well as the age of the patients is in line with result from the full cohort of VA ECMO patients in the UMC Utrecht and seems to be representative on the VA ECMO population on this area [4].

	Successful weaning	Unsuccessful weaning
N	60	56
Gender	39 male, 21 female	32 male, 24 female
Age	Mean=54.0, std=14.0	Mean=55.8, std=11.4

Table 4.1: Patient characteristics of the successful and unsuccessful weaning group.

4.2 Missing data

For each high frequency parameter, the percentage of missing values before linear interpolation was calculated. These values can be found in Table 4.2. Heart rate, pulse pressure and mean arterial pressure all had around 1% of missing data in the total population. The CVD had a slightly higher percentage of missing data. The ECMO flow parameter had almost no missing data. It is important to consider that the percentage of missing values

of the ECMO flow were calculated before forward filling the data from a frequency of per hour to per minute. The VIS had the highest number of missing values.

Parameter	% of missing values in all included ECMO runs
Pulse pressure	1.0
Mean arterial pressure	0.9
Central venous pressure	1.9
Heart rate	0.8
ECMO flow	0.2
Vasoctive inotropic score	5.3

Table 4.2: Percentage of missing values per high frequency parameter in all included ECMO runs of the full 72 hours.

4.3 Feature exploration

The t-test between the two outcome groups, successful weaning and unsuccessful weaning, found five features to have a significant difference between the groups based on a $p < 0.05$. These were the standard deviation of the complete PP (p=0.00), standard deviation of the MAP on day three (p=0.05) the mean PP on day two (p=0.05) and day three (p=0.01) and the complete time (p=0.01). However, in this case a high change at a type 1 error, or false positive, is present due to the multiple statistical tests and therefore these results were unreliable. To resolve this, a Bonferroni correction was done. The t-test with Bonferroni correction found that none of the features had a significant impact on the outcome.

The linear correlation between all the features was calculated using the Pearson correlation coefficient. A correlation of 1.0 signifies a very strong positive correlation, -1.0 a very strong negative correlation. A correlation value of 0.0 means there is no linear relationship. The full correlation matrix can be found in Appendix D.1. The features with a coefficient greater than 0.7, usually referred to as having a strong correlation [35], are inspected further and are shown separately in Appendix D.2. Notably, the mean PP, MAP, CVD, flow from ECMO and VIS between the three days in the same feature show a strong correlation. The scatter plots of the mean PP on day two and three and the scatterplot of HR on day two and three, see Figure 4.2 and 4.3, concur.

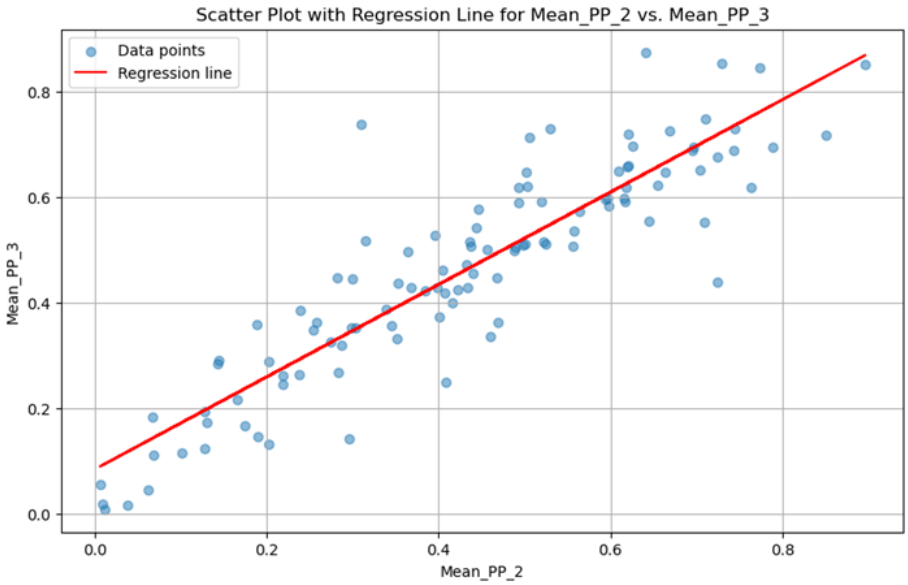


Figure 4.2: Scatter plot with regression line for the mean pulse pressure on day two and day three.

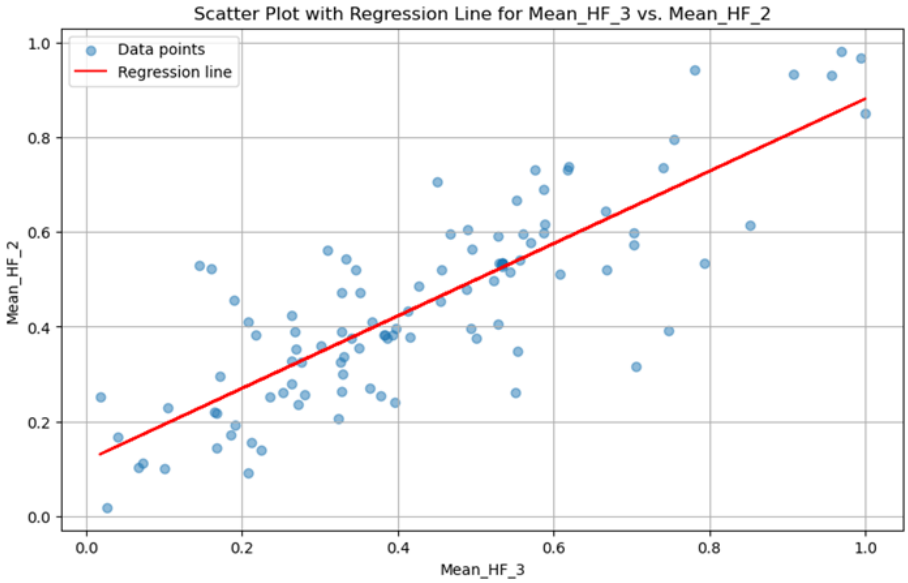


Figure 4.3: Scatter plot with regression line for the mean heart rate on day two and day three.

4.4 Feature selection

The features were selected separately for every model. This created five subsets of features with a different amount of selected features. An overview of the selected features for each model can be found in Appendix D.2. Each model had a different amount of features (RF:26, GB:11, LR:9, SVR:22, KNN:19). The amount of features were chosen based on the best RMSE for each amount of features in the subset. The graphs showing each RMSE for each number of features can also be found in Appendix F. The graphs for the features importance can also be found in Appendix F.

Noticeably, the feature subsets for the SVR and KNN models are similar, with 14 features selected by both in nearly the same order. Additionally, the four most important features are the same across SVR, KNN, and LR. These are the mean of the lactate of the

full three days, the patient gender and the standard deviation of the VIS on day two and three. In contrast, the RF, GB and LR models do not show the same amount of overlap, having no overlap in the top 10 features with any of the four other models.

4.5 Performance of the models

Metric/ Model	Support vector regressor	K-Nearest Neighbours	Random forest	Gradient boosting	Logistic regression
RMSE	0.49	0.48	0.48	0.55	0.51
R^2	0.01	0.07	-0.04	-0.19	-0.08
MAE	0.48	0.45	0.42	0.51	0.49

Table 4.3: Performance of the five models based on root mean squared error (RMSE), R^2 and mean absolute error (MAE).

Table 4.3 shows the performance of the five models based on RMSE, R^2 and MAE. The K-nearest neighbours regressor and the random forest regressor have the best performance based on RMSE and MAE. SVR, KNN and RF all have a value below 0.5. The difference between MAE and RMSE also differs between the models with the RF model yielding the largest difference, 0.06, and the SVR yielding the smallest difference, 0.01.

Looking at the R^2, the support vector regressor and the K-nearest neighbours regressor are the only models to have a positive value. A negative R^2 means that the model fits the validation data worse than predicting the mean of the data for all data points. This indicates the model does not explain the relationship between the features and the outcome.

4.5.1 Random forest regressor

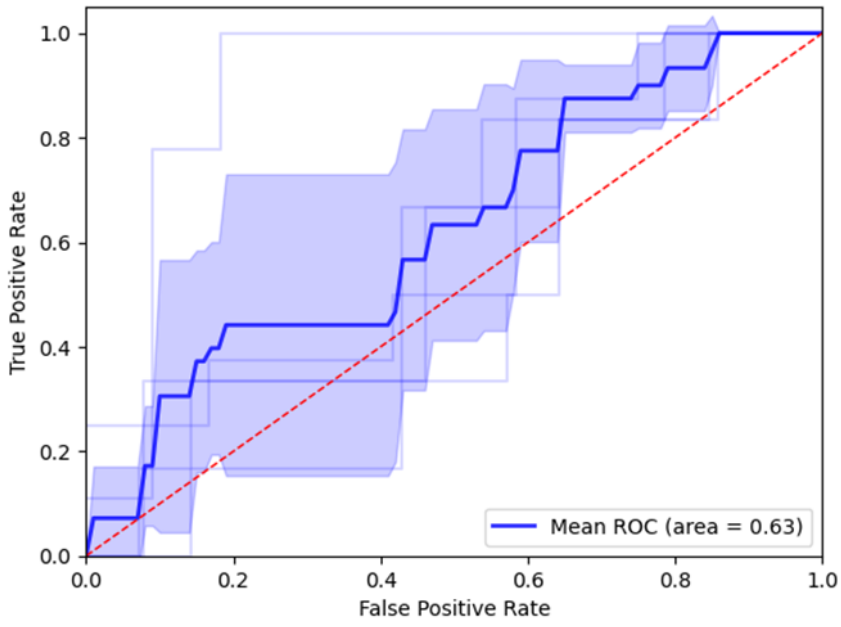


Figure 4.4: ROC- curve of the Random Forest Regressor model. The bold blue line represents the mean ROC of the 5 folds of the cross validation. The light blue lines each represent the ROC of one fold. The blue coloured area represents the confidence interval.

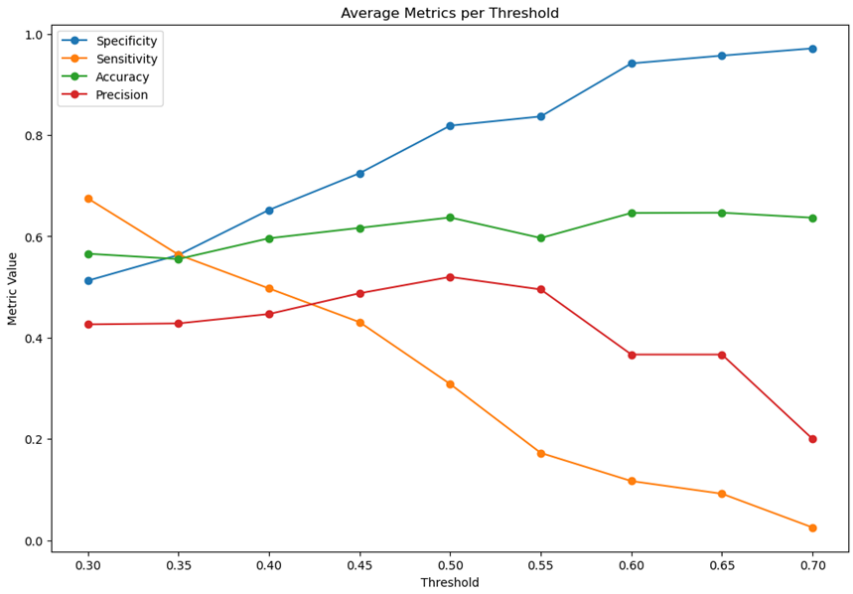


Figure 4.5: Classification metrics (specificity, sensitivity, accuracy and precision) for different thresholds between 0.3 and 0.7 of the predicted value from the RF regression model across the five outer folds.

Among the five models, the two best performers based on RMSE and MAE, are RF and KNN. The performance of these models are analysed further. To assess the implications of using probabilities as a decision-making tool, classification metrics are also evaluated. Figure 4.4 shows the ROC-curve of the Random forest model. The area under the curve of the average of the five outer folds is 0.63. The ROC-curve for each fold (pictured as the light blue lines) shows there is a high variation between each fold.

The classification metrics for thresholds between 0.3 and 0.7 can be found in Figure 4.5. Based on these lines, a cut-off around 0.35 gives a more balanced outcome between the four classification metrics. The specificity is higher than the sensitivity on nearly all thresholds.

Figure 4.6 shows the SHAP values for each feature. Each dot shows how the datapoint had an effect on the model. A positive SHAP value means the point was predictive for weaning success and a negative value meaning the point was predictive for an unsuccessful weaning. The colour of the point reveals whether the value was high or low. In this case it shows that a high value of the slope of the MAP on day three was used to predict for a successful weaning. The slope of the ECMO flow of the complete run, the slope of the VIS of the complete run and the slope of the VIS on day one predicted successful weaning in the RF model if they had a higher value. A higher positive slope of the pulse pressure of the complete run, on day one and day three was also used to predict a successful weaning. Lastly, a low standard deviation of VISis used to predict weaning success. The other values have a less distinct impact on the model output.

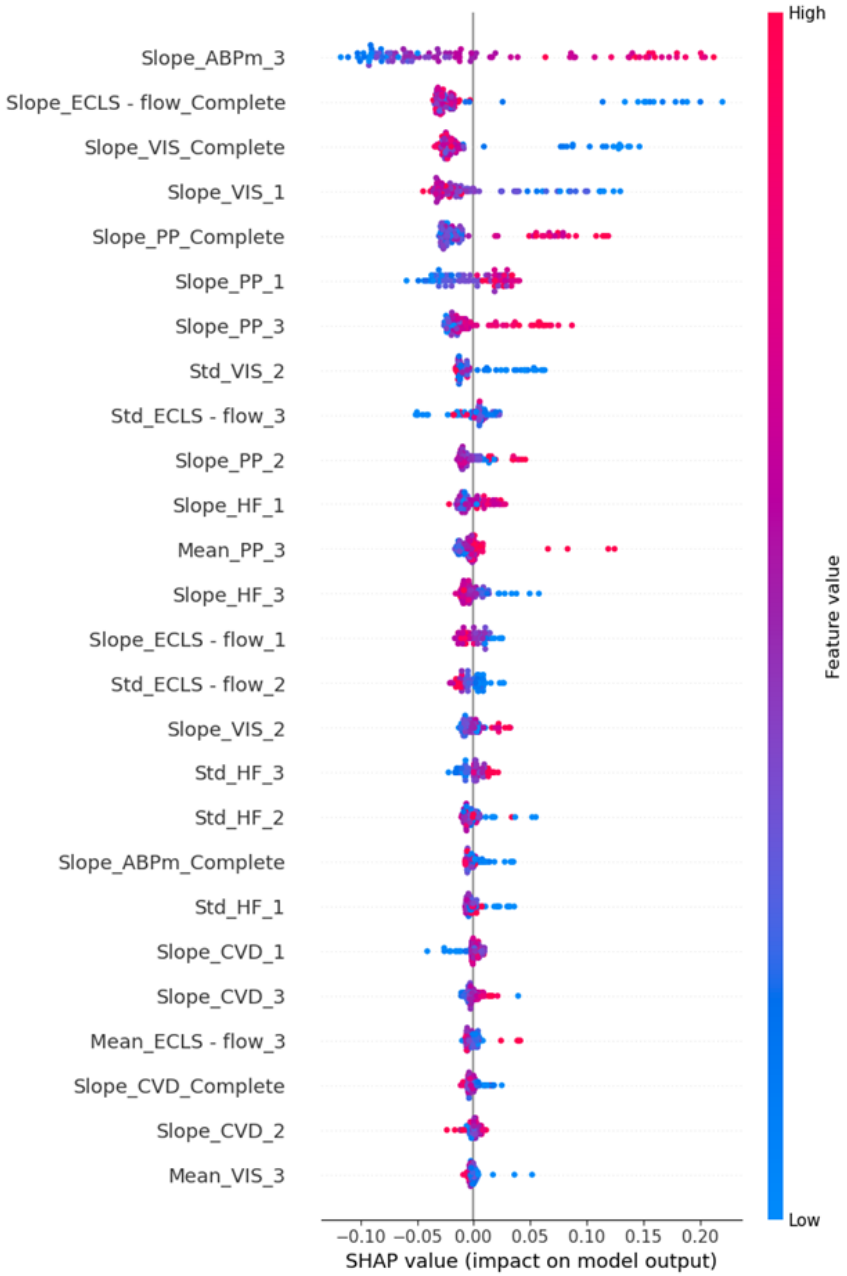


Figure 4.6: Shapley Additive Explainer (SHAP) value for each of the prediction per feature for the Random Forest model. It shows the impact on the model output with a positive value relating to weaning success and a negative SHAP value to weaning failure. The colour of the dots refers to the actual value of the feature being high (red) or low (blue). The features are named based on the calculation (slope, mean or std (standard deviation)) and the abbreviation of the parameter. The number behind the feature name corresponds with the day of ECMO treatment, 'complete' refers to the full 72 hours as interval.

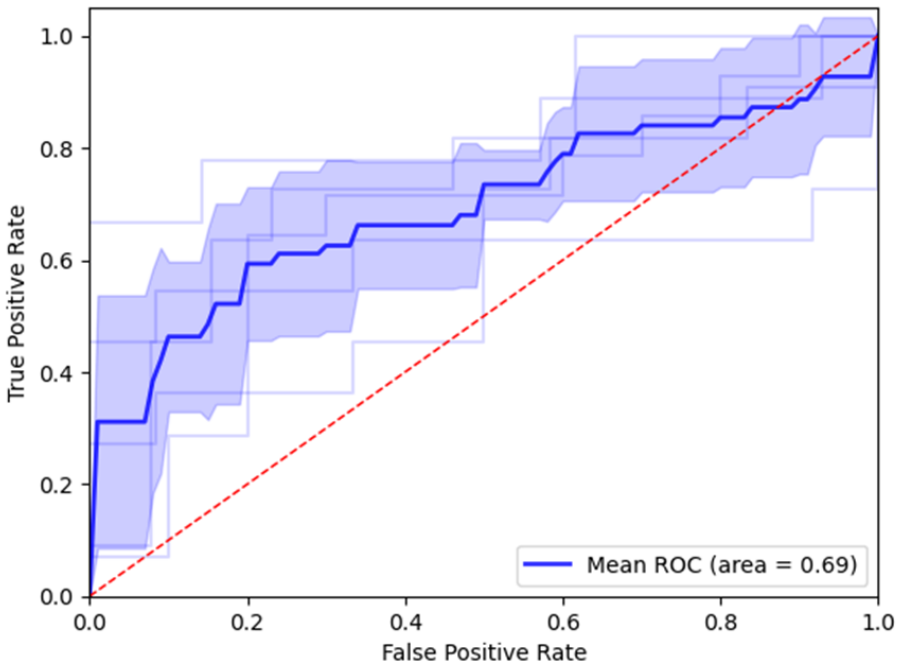


Figure 4.7: ROC- curve of the K-Nearest Neighbours model. The bold blue line represents the mean ROC of the 5 folds of the cross validation. The light blue lines each represent the ROC of one fold. The blue coloured area represents the confidence interval.

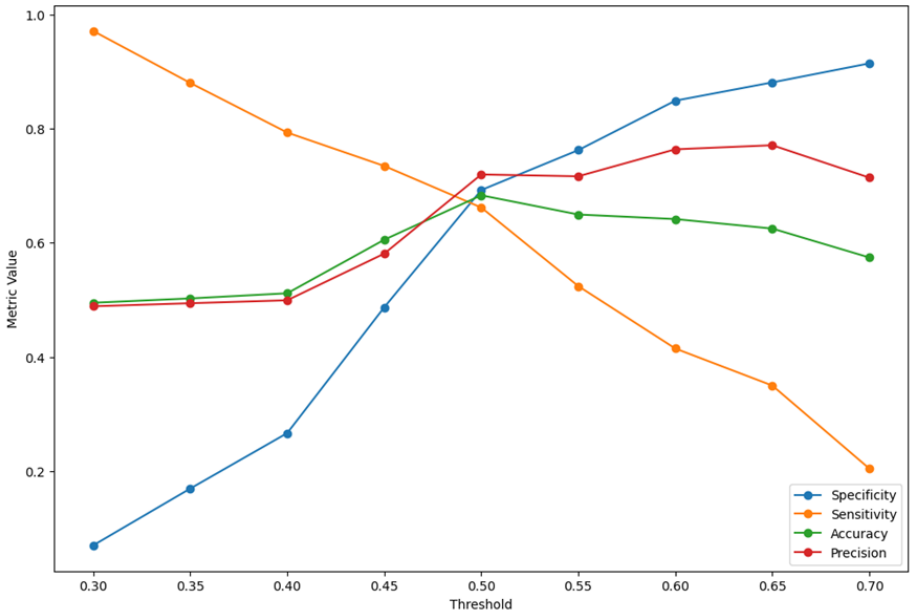


Figure 4.8: Classification metrics (specificity, sensitivity, accuracy and precision) for different thresholds between 0.3 and 0.7 of the predicted value from the KNN regression model across the five outer folds.

4.5.2 K-Nearest Neighbours regressor

In Figure 4.7, the ROC-curve of the KNN model can be found. The area under the curve of the average of the five folds is 0.69. The ROC-curve for each outer fold, the light blue lines, show less variation between the folds than the RF model. The main attribution of the area under the curve value is due to the bottom left quadrant of the ROC curve. This area indicates that for low false positive rates, the true positive rate remains high, which means the model can achieve both high specificity and a sensitivity greater than 0.5.

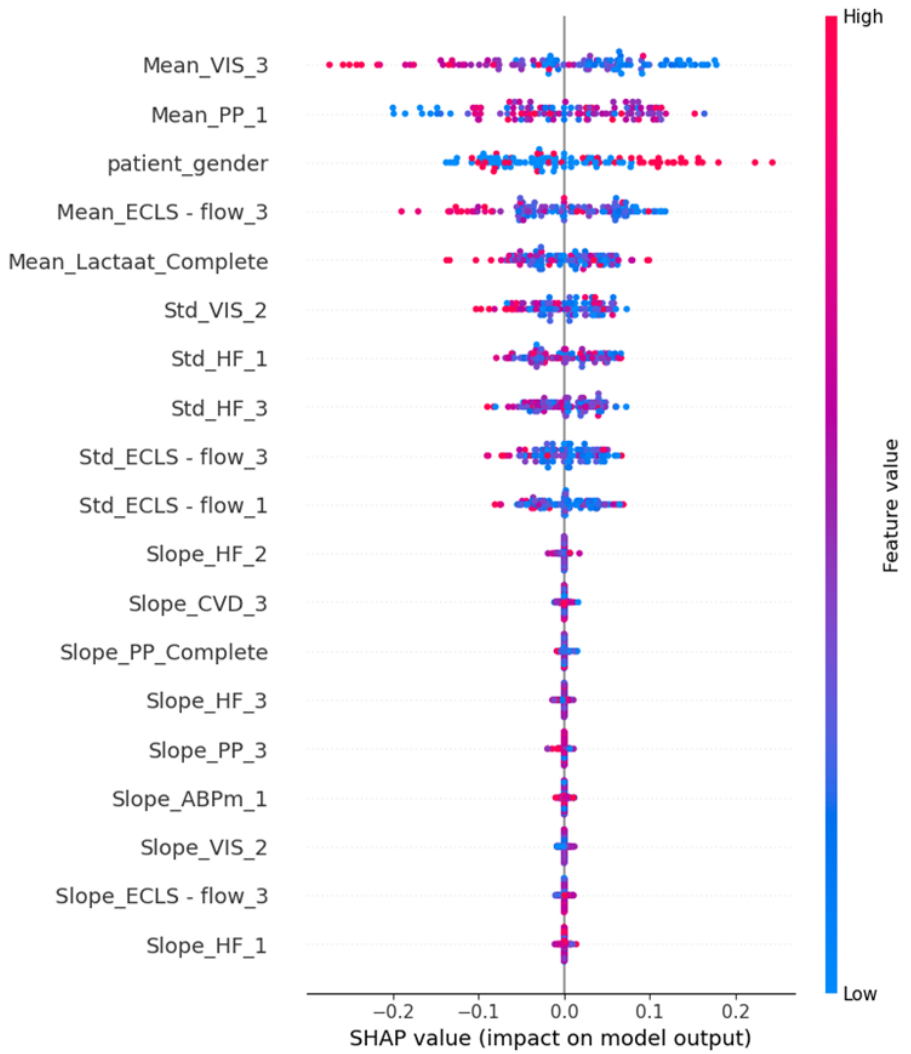


Figure 4.9: Shapley Additive Explainer (SHAP) value for each of the prediction per feature for the K-nearest neighbours model, showing the impact on the model output with a positive value relating to weaning success and a negative SHAP value to weaning failure. The colour of the dots refers to the actual value of the feature being high (red) or low (blue). The features are named based on the calculation (slope, mean or std (standard deviation)) and the abbreviation of the parameter. The number behind the feature name corresponds with the day of ECMO treatment, 'complete' refers to the full 72 hours as interval.

The classification metrics, graphed in Figure 4.8, show that at a threshold of 0.5 all the metrics come to nearly the same value. In this figure, we can also see that the sensitivity tends to achieve higher values then the specificity, albeit less clear than in the Random Forest model. The precision and accuracy do not show as much change as the accuracy and precision between the thresholds. They show an increase from 0.5 to 0.7 between a threshold of 0.4 and 0.5 and remain steady outside those thresholds.

The SHAP value for each prediction are graphed in Figure 4.9. The mean of the VIS on day three shows the highest impact on the model output with a low VIS being predictive of weaning success. A distinction can been seen between low values (blue dots) having a positive impact and high values (red dots) having a negative impact. A low flow from the ECMO on day three also has a positive impact on the output. The mean of the pulse pressure also has a high impact on the model output, but has a more mixed distribution of the low and high values. Although, higher values have a positive impact on the model output. The bottom nine selected features show very little impact on the model output and do not show grouping of low or high value having a distinct impact on the prediction.

5. Discussion

This study attempted to develop a model that predicts whether cardiac recovery will be sufficient to provide adequate circulation and oxygenation after VA ECMO support is discontinued. The five models developed for this purpose did not perform sufficient to provide meaningful predictions based on the model performance. Based on these findings, it cannot be concluded that using hemodynamic parameters in the initial stage of VA ECMO treatment can predict whether a patient can be weaned from the VA EMCO successfully. The poor results suggest the used features did not possess enough predictive power.

5.1 Interpretation of results

5.1.1 Regression metrics of the five models

The performance metrics reveal that none of the five models can predict weaning success in VA ECMO with usable accuracy. All models have an average RMSE exceeding 0.45. This indicates poor predictive power since the average error is nearly 0.5, the midpoint of a binary outcome scale. The R^2 values are also weak, with the best performing model scoring just 0.09 [47]. This suggests the models explain very little of the variance, leaving most of the outcome variance unexplained. Moreover, three out of the five models have negative R^2 scores. This suggests that the model explains the variance between the feature and outcome groups even less effectively than if the mean value were used as the prediction for all data points [46].

The poor regression metrics across all models suggest that the input data may have been insufficient. While a poorly designed model could account for these outcomes, the fact that all models performed similarly poor suggests the features do not have enough predictive power. This can be explained by different factors. The continuous hemodynamic parameters have a lot of potential to tell a lot about the cardiac status of the patient, however the method of feature extraction could explain the low predictive power. Using only the first three days of data might reduce predictive value. Particularly in patients who are on ECMO for extended periods and are eventually successfully weaned [50]. In such cases, initial parameters may indicate slower or less cardiac recovery, which could be misleading, as these patients might show good myocardial recovery after the three-day interval. Secondly, the chosen interval of 24 hours for the bag of features method means changes in the hemodynamic parameters on a smaller time scale will be missed. Also, during the day, the parameters will most likely change due to daily activities. That will ensure similar patterns of change in both outcome groups and reduce distinction between groups.

5.1.2 Classification metrics of KNN and RF

The KNN and Random Forest models were the best performing models. For these models, classification metrics were calculated to provide further insight in the results. The classification metrics reflect the consequences of using the probability to categorise the patients

and give better insight in the usability of the model as a prognostic tool for physicians. The best area under the ROC-curve was 0.69. This is comparable to the SAVE prediction by Schmidt et al. [51], which predicts survival in VA ECMO patients based on medical history, pulse pressure, and lower serum bicarbonate prior to implementation. However, the AUC of the external validation of SAVE was better (0.90) and was therefore determined to be accurate enough for implementation. The AUC in our models is also comparable to the survival prediction of VA ECMO with the conditional inference trees by Braun et al. [13]. They found an AUC of 0.71. However, the AUC is significantly lower compared to Ayers' AUC of 0.92 [12] and Wang's AUC of 0.85 [11].

The discrepancy between AUC and R^2 is noteworthy. For the KNN model, the AUC was 0.69 and the R^2 0.04. This suggests that while the model can distinguish between categories, it has poor calibration [52]. This poor calibration could be explained by the binary nature of the outcome parameter, which does not reflect the actual situation. For instance, both a patient showing just enough cardiac recovery and another who achieves excellent recovery are recorded as 1 if both are weaned successfully. Additionally, it is important to note that some predicted values exceed 1, further increasing the error and subsequently the R^2. The classification metrics of KNN show a higher sensitivity across the thresholds and a higher specificity in the RF model. In a model that predicts the likelihood of successful weaning from VA ECMO, high specificity should be prioritised. This is because the consequences of incorrectly predicting that a patient will fail to wean, a false positive, can significantly impact subsequent treatment decisions, such considering the implementation of ventricular assist devices (VAD), heart transplantation (HTx), or even transitioning to palliative care. On the other hand, if a patient is predicted to have a high chance of weaning successfully, VA ECMO support should be maintained until echocardiographic, clinical and hemodynamic parameters meet the criteria to start a weaning trial. However, in this model the classification metrics are not good enough to base decisions on, as a specificity and sensitivity of over 80% are necessary to consider a model as a prognostic tool [53].

5.1.3 Influence of features on prediction in KNN and RF

SHAP explainers were used to understand the effects of the features on the predicted outcome. The SHAP graphs of the random forest model show that a positive slope of the MAP and PP were used in the Random Forest model to predict weaning success. This indicates that an increase in MAP and PP at the start of the ECMO run result in a higher weaning success prediction. This aligns with existing literate relating higher PP and MAP with better cardiac recovery [9, 14]. A bigger slope of the PP was suspected to be indicative of better cardiac function, because it shows an improvement in pressure provided from the heart [25]. The fact that the slope of the PP had a high features importance in RF, LR and GB concurs with this fact.

In the SHAP graphs, it was also found that a low (or negative) slope of the VIS and ECMO flow were used to predict weaning success in the RF model. VIS and ECMO flow both are a type of hemodynamic support [16, 27]. So, a lower setting of flow on the ECMO and less vasoactive and inotropic medication means that the heart needs less support [14]. However, these are factors controlled by physicians. Thus, the reduction in VIS and ECMO indicates that the physicians observed an improvement in the hemodynamic status, leading to a decrease of hemodynamic support. This would mean these values are more likely a representation of the opinion of the physicians on the clinical status of the patient, then predictive values of their own.

The SHAP graphs of the KNN model show different features having a higher impact on the model output. The mean of the VIS on day 3 shows the highest impact on the model output with a low VIS being predictive of weaning success. This could be explained by the same reasoning that these are type hemodynamic support by physicians. A low VIS would mean the physicians found that the patient was stable at a lower dose of vasoactive and inotropic medication. Similar to the fact that a low mean flow from the ECMO on day 3 also had a positive impact on the output. It is interesting that in the KNN model, the mean had a higher impact on the weaning success prediction, but in the RF model it was the slope. The mean of the pulse pressure also has a high impact on the model output, but has a more mixed distribution of the low and high values. Although, higher values have a positive impact on the model output. This can be explained by the fact that a high pulse pressure correlates with a higher cardiac output from the heart and thus less cardiac dysfunction [17].

However, the physiological significance of the features is closely tied to the model performance. Given the models' poor performance, it is not possible to definitively determine which features might hold physiological importance in the prediction of weaning success.

5.1.4 Feature exploration

Statistical tests and a correlation matrix were calculated to better understand the features. The statistical tests show that none of the features can significantly explain the outcome result by themselves. This was expected as none of the features have been found to accurately predict weaning success on their own [54]. The choice for machine learning was made to use the interactions between features to predict outcomes by analysing patterns across multiple variables.

The correlation matrix shows high correlations (a correlation coefficient larger than 0.7 [35]) between the mean of the same feature over the days. This was the case for VIS, MAP, ECMO flow, PP, HF and CVD. The high correlation between the heart rate on day 1 and the heart rate of the patient on day 2 is most likely caused by the resting heart rate of the patient having the biggest influence on the mean value. Especially if the patient remains stable or has a pacemaker. This analogy can also explain the high correlation of MAP and CVD between the different days. Also, these values are highly influenced by the settings of the ECMO flow. A high correlation between the mean VIS and ECMO flow on different days indicates that the settings have not been changed much during the 72 hours.

5.2 Limitations

5.2.1 Data inclusion

A major limitation in machine learning is often the size of the dataset. A larger dataset improves machine learning models by enabling better generalisation, reducing bias, and allowing more complex patterns to be captured [22]. Only 109 ECMO trials were included in this study. A future possibility for increasing the dataset is to include data from other hospitals. Expanding the initial cohort could open up the possibility of including a longer ECMO treatment interval as training data, as the current cohort was limited by the number of patients on ECMO for more than 72 hours.

The dataset was retrospectively collected which leads to potential bias. Over the 16 year period of observation, there have been technological and procedural changes in ECMO treatment that may affect the chances of successful weaning.

5.2.2 Data preprocessing

Some limitations in the method also arise from the preprocessing of the data. Missing data was interpolated up to one hour and ECMO trials with more than four continuous hours of data were excluded. However, this has left 55 ECMO runs with missing data between one and four continuous hours. In these runs, the extracted features are influenced as missing data causes single values to have more effect on the mean and standard deviation. However, it was decided to not interpolate for these larger gaps as that might add more uncertainty. In those four hours a lot of changes could have happened and linear interpolation would not have reflected that accurately [55].

The flow from the VA ECMO was forward filled from a frequency of once per hour to once per minute with the assumption that the flow stayed the same up until a change was marked in the system. This might have resulted in faulty values for two reasons. One is that the input of VA ECMO is done manually and human errors could have been made. Due to the forward filling of the data, one wrong input might create an error for a longer period of time and will therefore cause a faulty feature. Unfortunately, it is impossible to detect the occurrence of the manual input of faulty data as long as it is within a physiological range. Another is that the flow from the EMCO device is the result of the amount of rounds per minute the device provides. The resulting flow is not a one-on-one result from this, but is also influenced by the pressure present in the aorta. This makes the feature not fully independent from other parameters such as the MAP.

The non-parametric outlier detection is based on a quantile. This is based on the assumption that values away from the dense centre are outliers, because faulty sensors mostly give extremely high or low values [31]. However, in biological data extreme physiological values can occur, especially in severely ill patients in the ICU. Visual inspection of the raw data does conclude that there are faulty measurements in the data and that outlier removal was necessary. However, the filtering of extreme data, might have removed hemodynamic data of critical situations and therefore reduced the possibility to predict based of these values.

5.2.3 Weaning success outcome

A limitation of this study is the fact that the outcome is not definitively measurable. Determining when a patient has been successfully weaned from ECMO is not governed by a clear rule [9]. Mortality outcomes, including 30-day, hospital, and ICU mortality rates, are commonly used as key measures [7, 11, 19]. However, mortality also includes deaths without a cardiac reason and is therefore not a perfect outcome. The used criteria for weaning success includes patients who have passed away shortly after weaning from the ECMO due to non-cardiac reasons. For these patients it is not known whether the heart showed myocardial recovery or not which might cloud the training data.

To explore the implications of the chosen criteria for weaning success, an analysis of patient characteristics was conducted. We aimed to identify the amount of patients that cannot be placed in successful or unsuccessful weaning with reasonable certainty. Patients who were alive at the 12-month follow-up without having undergone heart transplantation

or received a ventricular assist device (VAD) were categorised as weaning successes. This criterion applied to 38 patients. Conversely, patients who died in the ICU due to circulatory issues (n=35), underwent heart transplantation (n=2), or received a VAD (n=23) were categorised as being in the unsuccessful weaning group, indicating insufficient cardiac recovery.

However, five patients who died in the ICU from non-circulatory causes were included in the unsuccessful weaning group. In these cases, it is unclear whether the heart did not recover sufficiently or if a different issue was the cause of death. In reality, the cause of death is often more complex and interrelated with heart function [56]. A patient with marginal heart function might survive under stable conditions but could succumb if an infection or other complication arises.

Nine patients who died within the 12-month follow-up but survived the ICU were grouped with the successful weaning cases. However, it is arguable that some did not achieve full recovery, as hospital or 30-day mortality could be considered alternative criteria [9]. These criteria pose challenges similar to those discussed above, where the precise cause of death might not solely be due to cardiac function. In practice, the cause of death is often multifactorial and correlated with cardiac function.

In summary, while the criteria for defining weaning success aim to provide clear distinction, the complexity and multifactorial nature of patient outcomes present challenges in definitively categorizing weaning success. From this exploration, it can be concluded that for 14 patients the categorisation for weaning success is unclear. This is enough to have an impact on the performance of the models.

5.3 Future possibilities

5.3.1 Improvement of feature extraction

A wide range of approaches to extract features from multivariate time series data are available. In this study, a bag of features method was used. This method was chosen, because its capability to show trends of the parameters over time, both within the chosen time intervals and between them. With the current method, changes in the hemodynamic parameters on a smaller time scale can be missed, because the features are extracted based on an interval of 24 hours. Capturing smaller changes could be achieved by making the intervals smaller. However, this would result in an increased amount of features which would exceed the amount of datapoints in the current dataset. In that case prediction becomes unreliable with our current dataset size [44]. Thus, this method is only possible on a larger dataset or by using feature selection prior to the model implementation. Other feature extraction methods might give other insights in the state of the patient. An interesting possibility might be to interpret the lines of the continuous parameters as images and predict based on image recognition. This would eliminate the need to convert the data to features, but makes the explanation of predictions unclear.

Adding other parameters to the input could provide more information on the cardiac status of the patient and thus increase predictive power of the models. Echocardiographic parameters give a good insight in the cardiac status of the patient [14]. However, availability of echocardiographic data poses a problem. These measurements are not done routinely and were not recorded in the used database [4]. So, to include these parameters a new database would have to be build. Lastly, urine output was found to be predictive for in hospital death in ECMO patients and might also be predictive of weaning success [19].

5.3.2 Weaning parameter outcome

Instead of grouping the patient in successful and unsuccessful weaning, it might be interesting to predict the outcome based on a value that quantifies heart function. This could, for example, be an echocardiographic parameter such as ejection fraction. No other studies have attempted to predict cardiac recovery based on a numerical value. This would however require a cardiac ultrasound shortly after stopping the EMCO treatment.

5.3.3 Heterogeneity of population

The dataset used for this study only included patients with a cardiac indication for VA ECMO. However, this cardiac indication can encompass many diseases, which means the population is quite heterogeneous in terms of pathology [57]. Different diseases have different patterns of recovery and might therefore make it harder for a model to predict outcome based on these differing recovery patterns. A more homogeneous population, for example using only patients with myocardial infarcts, might result in better performance of the models. A downside to this is that the model will only be applicable to the chosen patient population.

5.3.4 Deep learning

Using deep learning instead of the used supervised methods might also be a possibility to achieve better performance. Stephens et al. found that a deep learning model outperformed models like Random Forest and Gradient Boosting significantly in predicting VA ECMO mortality [10]. Possibly due to the fact that deep learning can capture more complex patterns and does not need the amount of feature engineering as done in this study [38]. The ability to predict based on raw temporal hemodynamic data ensures that important data is not filtered out, which did happen in the method of this study. However, deep learning does not provide a clear influence of the features on the prediction. Machine learning models cannot be used as the only decision making tool in the clinical treatment of a patient and will only be used as an additional predictor. Therefore, it seems important to provide a justification for the prediction of a machine learning model to increase usability of the model for physicians. It would be beneficial to provide an explanation, such as: "For instance, this patient is unlikely to wean successfully because values X and Y are high." In deep learning, these justifications will be harder to provide as features are less interpretable and will therefore be harder to link to the physiological state of the patient.

6. Conclusion

In conclusion, the high RMSE and MAE, and low R^2 values show all five models struggled to provide meaningful predictions for weaning success in VA ECMO. The KNN and Random Forest models showed better classification metrics suggesting they can differentiate between success and failure but with limited accuracy. The models developed in this study perform inadequately to support decision-making during VA ECMO treatment. Overall, the findings underscore the need to rethink the input data and model design to enhance prediction accuracy in weaning success for VA ECMO patients. The use of continuous hemodynamic parameters did not provide good enough predictive power. Using a different feature extraction method to the one used in this study or rethinking the time intervals could improve predictive power. Also, adding alternative parameters that reflect cardiac status, such as echocardiographic parameters, could improve performance of machine learning models in future studies. Lastly, the definition of weaning success makes perfect categorisation of patients not possible. This definition also needs to be evaluated.

References

[1] Thiele, Holger ; Ohman, E M. ; Desch, Steffen ; Eitel, Ingo ; Waha, Suzanne de: Management of cardiogenic shock. In: *European Heart Journal* 36 (2015), 5, Nr. 20, 1223–1230. http://dx.doi.org/10.1093/eurheartj/ehv051. – DOI 10.1093/eurheartj/ehv051. – ISSN 0195–668X

[2] Ouweneel, Dagmar M. ; Schotborgh, Jasper V. ; Limpens, Jacqueline ; Sjauw, Krischan D. ; Engström, A E. ; Lagrand, Wim K. ; Cherpanath, Thomas G V. ; Driessen, Antoine H G. ; Mol, Bas A J M. ; Henriques, José P S.: Extracorporeal life support during cardiac arrest and cardiogenic shock: a systematic review and meta-analysis. In: *Intensive care medicine* 42 (2016), 12, Nr. 12, S. 1922–1934. http://dx.doi.org/10.1007/s00134-016-4536-8. – DOI 10.1007/s00134–016–4536–8. – ISSN 1432–1238 (Electronic)

[3] Darryl, Abrams ; Alain, Combes ; Daniel, Brodie: Extracorporeal Membrane Oxygenation in Cardiopulmonary Disease in Adults. In: *Journal of the American College of Cardiology* 63 (2014), 7, Nr. 25_Part_A, 2769–2778. http://dx.doi.org/10.1016/j.jacc.2014.03.046. – DOI 10.1016/j.jacc.2014.03.046

[4] Meuwese, C L. ; Hermens, J A. ; Haan, M de ; Braithwaite, S A. ; Ramjankhan, F ; Buijsrogge, M P. ; Jonge, N de ; Kirkels, J H. ; Jong, M de ; Pasma, W ; Vromen-Wijsman, J L P. ; Kraaijeveld, A O. ; Waal, E E. ; Torn, E ; Platenkamp, M ; Heijden, J J. d. ; Cremer, O L. ; Dijk, D van ; Donker, D W.: Twelve years of circulatory extracorporeal life support at the University Medical Centre Utrecht. In: *Netherlands Heart Journal* 29 (2021), Nr. 7, 394–401. http://dx.doi.org/10.1007/s12471-021-01552-z. – DOI 10.1007/s12471–021–01552–z. – ISSN 1876–6250

[5] Oude Lansink-Hartgring, Annemieke ; Hengel, Berber van d. ; Bij, Wim van d. ; Erasmus, Michiel E. ; Mariani, Massimo A. ; Rienstra, Michiel ; Cernak, Vladimir ; Vermeulen, Karin M. ; Bergh, Walter M. d. ; Group, on behalf of the Dutch Extracorporeal Life Support S.: Hospital Costs Of Extracorporeal Life Support Therapy. In: *Critical Care Medicine* 44 (2016), Nr. 4. https://journals.lww.com/ccmjournal/fulltext/2016/04000/hospital_costs_of_extracorporeal_life_support.9.aspx. – ISSN 0090–3493

[6] Pladet, Lara ; Luijken, Kim ; Fresiello, Libera ; Miranda, Dinis Dos R. ; Hermens, Jeannine A. ; Smeden, Maarten v. ; Cremer, Olaf ; Donker, Dirk W. ; Meuwese, Christiaan L.: Clinical decision support for ExtraCorporeal Membrane Oxygenation: Will we fly by wire? In: *Perfusion* 38 (2023), 4, Nr. 1_suppl, 68–81. http://dx.doi.org/10.1177/02676591231163688. – DOI 10.1177/02676591231163688. – ISSN 0267–6591

[7] Cusanno, Axelle ; Aissaoui, Nadia ; Minville, Vincent ; Porterie, Jean ; Biendel, Caroline ; Volle, Kim ; Crognier, Laure ; Conil, Jean-Marie ; Delmas, Clément: Predictors of weaning failure in case of VA ECMO implantation. In: *Scientific Reports* 12 (2022), Nr. 1, 13842. http://dx.doi.org/10.1038/s41598-022-18105-y. – DOI 10.1038/s41598–022–18105–y. – ISSN 2045–2322

[8] Ortuno, Sofia ; Delmas, Clément ; Diehl, Jean-Luc ; Bailleul, Clotilde ; Lancelot, Aymeric ; Naili, Mahassen ; Cholley, Bernard ; Pirracchio, Romain ; Aissaoui, Nadia: *Weaning from veno-arterial extra-corporeal membrane oxygenation: which strategy to use?*

[9] Hermens, Jeannine A. ; Meuwese, Christiaan L. ; Szymanski, Mariusz K. ; Gianoli, Monica ; Dijk, Diederik van ; Donker, Dirk W.: Patient-centered weaning from venoarterial extracorporeal membrane oxygenation: "A practice-oriented narrative review of literature". In: *Perfusion* 38 (2023), 10, Nr. 7, S. 1349–1359. http://dx.doi.org/10.1177/02676591221115938. – DOI 10.1177/02676591221115938. – ISSN 1477–111X (Electronic)

[10] Stephens, A F. ; Šeman, M ; Diehl, A ; Pilcher, D ; Barbaro, R P. ; Brodie, D ; Pellegrino, V ; Kaye, D M. ; Gregory, S D. ; Hodgson, C: ECMO PAL: using deep neural networks for survival prediction in venoarterial extracorporeal membrane oxygenation. In: *Intensive Care Medicine* 49 (2023), Nr. 9, 1090–1099. http://dx.doi.org/10.1007/s00134-023-07157-x. – DOI 10.1007/s00134–023–07157–x

[11] Wang, Liangshan ; Yang, Feng ; Wang, Xiaomeng ; Xie, Haixiu ; Fan, Eddy ; Ogino, Mark ; Brodie, Daniel ; Wang, Hong ; Hou, Xiaotong: Predicting mortality in patients undergoing VA-ECMO after coronary artery bypass grafting: the REMEMBER score. In: *Critical Care* 23 (2019), Nr. 1, 11. http://dx.doi.org/10.1186/s13054-019-2307-y. – DOI 10.1186/s13054–019–2307–y. – ISSN 1364–8535

[12] Ayers, B ; Wood, K ; Gosev, I ; Prasad, S: Predicting Survival After Extracorporeal Membrane Oxygenation by Using Machine Learning. In: *Annals of Thoracic Surgery* Bd. 110, 2020, 1193–1200

[13] Braun, Julia ; Sahli, Sebastian D. ; Spahn, Donat R. ; Röder, Daniel ; Neb, Holger ; Lotz, Gösta ; Aser, Raed ; Wilhelm, Markus J. ; Kaserer, Alexander: Predicting Survival for Veno-Arterial ECMO Using Conditional Inference Trees—A Multicenter Study. In: *Journal of Clinical Medicine* 12 (2023), Nr. 19. http://dx.doi.org/10.3390/jcm12196243. – DOI 10.3390/jcm12196243. – ISSN 2077–0383

[14] Fresiello, Libera ; Hermens, Jeannine A J. ; Pladet, Lara ; Meuwese, Christiaan L. ; Donker, Dirk W.: The physiology of venoarterial extracorporeal membrane oxygenation - A comprehensive clinical perspective. In: *Perfusion* 39 (2024), Nr. 1_suppl, 5S-12S. http://dx.doi.org/10.1177/02676591241237639. – DOI 10.1177/02676591241237639

[15] Scott, Sean ; McCanney, Peter ; Morgan, Steve: *Veno-Arterial ECMO*. 2021

[16] Extracorporeal Life Support Organization: ELSO Guidelines for Cardiopulmonary Extracorporeal Life Support. In: *Ann Arbor, MI, USA* Version 1. (2017). www.elso.org

[17] Homan, Travis D. ; Bordes, Stephen J. ; Cichowski, Erica: *Physiology, Pulse Pressure.* StatPearls Publishing, Treasure Island (FL), 2023 http://europepmc.org/abstract/MED/29494015http://europepmc.org/books/NBK482408https://www.ncbi.nlm.nih.gov/books/NBK482408

[18] Xue, B ; Shah, N ; Yang, H ; Kannampallil, T ; Payne, P R O. ; Lu, C ; Said, A S.: Multi-horizon predictive models for guiding extracorporeal resource allocation in critically ill COVID-19 patients. In: *Journal of the American Medical Informatics Association* 30 (2023), Nr. 4, 656–667. http://dx.doi.org/10.1093/jamia/ocac256. – DOI 10.1093/jamia/ocac256

[19] Braun, Julia ; Sahli, Sebastian D. ; Spahn, Donat R. ; Röder, Daniel ; Neb, Holger ; Lotz, Gösta ; Aser, Raed ; Wilhelm, Markus J. ; Kaserer, Alexander ; Morisson, Louis ; Duceau, Baptiste ; Do Rego, Hermann ; Lancelot, Aymeric ; Hariri, Geoffroy ; Charfeddine, Ahmed ; Laferrière-Langlois, Pascal ; Richebé, Philippe ; Lebreton, Guillaume ; Provenchère, Sophie ; Bouglé, Adrien ; Chang, Sheng-Nan ; Hu, Nian-Ze ; Wu, Jo-Hsuan ; Cheng, Hsun-Mao ; Caffrey, James L. ; Yu, Hsi-Yu ; Chen, Yih-Sharng ; Hsu, Jiun ; Lin, Jou-Wei: Urine output as one of the most important features in differentiating in-hospital death among patients receiving extracorporeal membrane oxygenation: a random forest approach. In: *European Journal of Medical Research* 28 (2023), Nr. 1, 347. http://dx.doi.org/10.1186/s40001-023-01294-1. – DOI 10.1186/s40001–023–01294–1. – ISSN 2077–0383

[20] Xue, B ; Said, A S. ; Xu, Z ; Liu, H ; Shah, N ; Yang, H ; Payne, P ; Lu, C: Assisting Clinical Decisions for Scarcely Available Treatment via Disentangled Latent Representation. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2023, 5360–5371

[21] Wang, K ; Yan, L Z. ; Li, W Z. ; Jiang, C ; Wang, N N. ; Zheng, Q ; Dong, N G. ; Shi, J W.: Comparison of Four Machine Learning Techniques for Prediction of Intensive Care Unit Length of Stay in Heart Transplantation Patients. In: *Frontiers in Cardiovascular Medicine* 9 (2022). http://dx.doi.org/10.3389/fcvm.2022.863642. – DOI 10.3389/fcvm.2022.863642

[22] Bishop, Christopher M.: *Pattern recognition and machine learning.* New York : Springer, [2006] ©2006 https://search.library.wisc.edu/catalog/9910032530902121

[23] Fawagreh, Khaled ; Gaber, Mohamed M. ; Elyan, Eyad: Random forests: from early developments to recent advancements. In: *Systems Science & Control Engineering* 2 (2014), 12, Nr. 1, 602–609. http://dx.doi.org/10.1080/21642583.2014.956265. – DOI 10.1080/21642583.2014.956265. – ISSN null

[24] Natekin, Alexey ; Knoll, Alois: Gradient boosting machines, a tutorial. In: *Frontiers in Neurorobotics* 7 (2013). https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2013.00021. – ISSN 1662–5218

[25] Azampanah, Kimmia: *The clinical value of pulse pressure in patients undergoing Veno-Arterial Extracorporeal Membrane Oxygenation: a single-centre retrospective cohort study*, University of Twente, Diss., 2023

[26] Gandhi, Kruti D. ; Moras, Errol C. ; Niroula, Shailesh ; Lopez, Persio D. ; Aggarwal, Devika ; Bhatia, Kirtipal ; Balboul, Yoni ; Daibes, Joseph ; Correa, Ashish ; Dominguez, Abel C. ; Birati, Edo Y. ; Baran, David A. ; Serrao, Gregory ; Mahmood, Kiran ; Vallabhajosyula, Saraschandra ; Fox, Arieh: Left Ventricular Unloading With Impella Versus IABP in Patients With VA-ECMO: A Systematic Review and Meta-Analysis. In: *The American Journal of Cardiology* 208 (2023), 53–59. http://dx.doi.org/https://doi.org/10.1016/j.amjcard.2023.09.023. – DOI https://doi.org/10.1016/j.amjcard.2023.09.023. – ISSN 0002–9149

[27] Sandrio, Stany ; Krebs, Joerg ; Leonardy, Eva ; Thiel, Manfred ; Schoettler, Jochen J.: Vasoactive Inotropic Score as a Prognostic Factor during (Cardio-) Respiratory ECMO. In: *Journal of Clinical Medicine* 11 (2022), Nr. 9. http://dx.doi.org/10.3390/jcm11092390. – DOI 10.3390/jcm11092390. – ISSN 2077–0383

[28] Fuller, Joshua ; Abramov, Alexey ; Mullin, Dana ; Beck, James ; Lemaitre, Philippe ; Azizi, Elham: A Deep Learning Framework for Predicting Patient Decannulation on Extracorporeal Membrane Oxygenation Devices: Development and Model Analysis Study. In: *JMIR Biomed Eng* 9 (2024), 2, e48497. http://dx.doi.org/10.2196/48497. – DOI 10.2196/48497. – ISSN 2561–3278

[29] Terry, W R. ; Lee, J B. ; Kumar, A: Time series analysis in acid rain modeling: Evaluation of filling missing values by linear interpolation. In: *Atmospheric Environment (1967)* 20 (1986), Nr. 10, 1941–1943. http://dx.doi.org/https://doi.org/10.1016/0004-6981(86)90335-5. – DOI https://doi.org/10.1016/0004–6981(86)90335–5. – ISSN 0004–6981

[30] Shapiro, S S. ; Wilk, M B. ; Chen, H J.: A Comparative Study of Various Tests for Normality. In: *Journal of the American Statistical Association* 63 (1968), 12, Nr. 324, 1343–1372. http://dx.doi.org/10.1080/01621459.1968.10480932. – DOI 10.1080/01621459.1968.10480932. – ISSN 0162–1459

[31] Su, Xiaogang ; Tsai, Chih-Ling: Outlier detection. In: *WIREs Data Mining and Knowledge Discovery* 1 (2011), 5, Nr. 3, 261–268. http://dx.doi.org/https://doi.org/10.1002/widm.19. – DOI https://doi.org/10.1002/widm.19. – ISSN 1942–4787

[32] Ahsan, Md M. ; Mahmud, M A P. ; Saha, Pritom K. ; Gupta, Kishor D. ; Siddique, Zahed: Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. In: *Technologies* 9 (2021), Nr. 3. http://dx.doi.org/10.3390/technologies9030052. – DOI 10.3390/technologies9030052. – ISSN 2227–7080

[33] Castán-Lascorz, M A. ; Jiménez-Herrera, P ; Troncoso, A ; Asencio-Cortés, G: A new hybrid method for predicting univariate and multivariate time series based on pattern forecasting. In: *Information Sciences* 586 (2022), 611–627. http://dx.doi.org/https://doi.org/10.1016/j.ins.2021.12.001. – DOI https://doi.org/10.1016/j.ins.2021.12.001. – ISSN 0020–0255

[34] García-Pérez, Miguel A.: Use and misuse of corrections for multiple testing. In: *Methods in Psychology* 8 (2023), 100120. http://dx.doi.org/https://doi.org/10.1016/j.metip.2023.100120. – DOI https://doi.org/10.1016/j.metip.2023.100120. – ISSN 2590–2601

[35] Schober, Patrick ; Boer, Christa ; Schwarte, Lothar: Correlation Coefficients: Appropriate Use and Interpretation. In: *Anesthesia & Analgesia* 126 (2018), 2, S. 1. http://dx.doi.org/10.1213/ANE.0000000000002864. – DOI 10.1213/ANE.0000000000002864

[36] Paul, Ranjit K.: Multicollinearity: Causes, effects and remedies. In: *IASRI* 1 (2006), Nr. 1, S. 58–65

[37] Bi, Jian: A eview of statistical methods for determination of relative importance of correlated predictors and identification of drivers of consumer liking. In: *Journal of Sensory Studies* 27 (2012), 4, Nr. 2, 87–101. http://dx.doi.org/https://doi.org/10.1111/j.1745-459X.2012.00370.x. – DOI https://doi.org/10.1111/j.1745–459X.2012.00370.x. – ISSN 0887–8250

[38] Janiesch, Christian ; Zschech, Patrick ; Heinrich, Kai: Machine learning and deep learning. In: *Electronic Markets* 31 (2021), Nr. 3, 685–695. http://dx.doi.org/10.1007/s12525-021-00475-2. – DOI 10.1007/s12525–021–00475–2. – ISSN 1422–8890

[39] Pedregosa, Fabian ; Varoquaux, Gaël ; Gramfort, Alexandre ; Michel, Vincent ; Thirion, Bertrand ; Grisel, Olivier ; Blondel, Mathieu ; Prettenhofer, Peter ; Weiss, Ron ; Dubourg, Vincent ; VanderPlas, Jake ; Passos, Alexandre ; Cournapeau, David ; Brucher, Matthieu ; Perrot, Matthieu ; Duchesnay, Edouard: Scikit-learn: Machine Learning in Python. In: *CoRR* abs/1201.0490 (2012). http://arxiv.org/abs/1201.0490

[40] Xie, Liping ; Li, Zilong ; Zhou, Yihan ; He, Yiliu ; Zhu, Jiaxin: Computational Diagnostic Techniques for Electrocardiogram Signal Analysis. In: *Sensors* (2020), 11

[41] Pudjihartono, Nicholas ; Fadason, Tayaza ; Kempa-Liehr, Andreas W. ; O'Sullivan, Justin M.: *A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction*. https://www.frontiersin.org/articles/10.3389/fbinf.2022.927312. Version: 2022

[42] Jović, A ; Brkić, K ; Bogunović, N: A review of feature selection methods with applications. In: *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2015, S. 1200–1205

[43] Khaire, Utkarsh M. ; Dhanalakshmi, R: Stability of feature selection algorithm: A review. In: *Journal of King Saud University - Computer and Information Sciences* 34 (2022), Nr. 4, 1060–1073. http://dx.doi.org/https://doi.org/10.1016/j.jksuci.2019.06.012. – DOI https://doi.org/10.1016/j.jksuci.2019.06.012. – ISSN 1319–1578

[44] Altman, Naomi ; Krzywinski, Martin: The curse(s) of dimensionality. In: *Nature Methods* 15 (2018), Nr. 6, 399–400. http://dx.doi.org/10.1038/s41592-018-0019-x. – DOI 10.1038/s41592–018–0019–x. – ISSN 1548–7105

[45] zhong, Yi ; He, Jianghua ; Chalise, Prabhakar: Nested and Repeated Cross Validation for Classification Model With High-Dimensional Data. In: *Revista Colombiana de Estadistica* 43 (2020), 6, S. 103–125. http://dx.doi.org/10.15446/rce.v43n1.80000. – DOI 10.15446/rce.v43n1.80000

[46] SijiGeorgeC, G ; B.Sumathi: Grid Search Tuning of Hyperparameters in Random Forest Classifier for Customer Feedback Sentiment Prediction. In: *International Journal of Advanced Computer Science and Applications* 11 (2020). https://api.semanticscholar.org/CorpusID:222433720

[47] Tatachar, Abhishek V.: Comparative assessment of regression models based on model evaluation metrics. In: *International Journal of Innovative Technology and Exploring Engineering* 8 (2021), Nr. 9, S. 853–860

[48] Chicco, Davide ; Warrens, Matthijs J. ; Jurman, Giuseppe: The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. In: *Peerj computer science* 7 (2021), S. e623. – ISSN 2376–5992

[49] Lundberg, Scott M. ; Lee, Su-In: A unified approach to interpreting model predictions. In: *CoRR* abs/1705.07874 (2017). http://arxiv.org/abs/1705.07874

[50] Vuthoori, Ravi ; Heaney, Cassandra ; Lima, Brian ; Knisel, Alexis ; Miller, Ed ; Kennedy, Kevin ; Majure, David ; Stevens, Gerin ; Bocchieri, Karl ; Cassiere, Hugh ; Fernandez, Harold ; Maybaum, Simon: Assessment of cardiac recovery in patients supported with venoarterial extracorporeal membrane oxygenation. In: *ESC Heart Failure* 9 (2022), 8, Nr. 4, 2272–2278. http://dx.doi.org/https://doi.org/10.1002/ehf2.13892. – DOI https://doi.org/10.1002/ehf2.13892. – ISSN 2055–5822

[51] Schmidt, Matthieu ; Burrell, Aidan ; Roberts, Lloyd ; Bailey, Michael ; Sheldrake, Jayne ; Rycus, Peter ; Hodgson, Carol ; Scheinkestel, Carlos ; Cooper, D ; Thiagarajan, Ravi ; Brodie, Daniel ; Pellegrino, Vince ; Pilcher, David: Predicting survival after ECMO for refractory cardiogenic shock: The survival after veno-arterial-ECMO (SAVE)-score. In: *European heart journal* 36 (2015), 6. http://dx.doi.org/10.1093/eurheartj/ehv194. – DOI 10.1093/eurheartj/ehv194

[52] Steyerberg, Ewout W. ; Vickers, Andrew J. ; Cook, Nancy R. ; Gerds, Thomas ; Gonen, Mithat ; Obuchowski, Nancy ; Pencina, Michael J. ; Kattan, Michael W.: Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. In: *Epidemiology* 21 (2010), Nr. 1. https://journals.lww.com/epidem/fulltext/2010/01000/assessing_the_performance_of_prediction_models__a.22.aspx. – ISSN 1044–3983

[53] Power, Michael ; Fell, Greg ; Wright, Michael: Principles for high-quality, high-value testing. In: *Evidence Based Medicine* 18 (2013), 2, Nr. 1, 5. http://dx.doi.org/10.1136/eb-2012-100645. – DOI 10.1136/eb–2012–100645

[54] Ostadal, Petr ; Rokyta, Richard ; Karasek, Jiri ; Kruger, Andreas ; Vondrakova, Dagmar ; Janotka, Marek ; Naar, Jan ; Smalcova, Jana ; Hubatova, Marketa ; Hromadka, Milan ; Volovar, Stefan ; Seyfrydova, Miroslava ; Jarkovsky, Jiri ; Svoboda, Michal ; Linhart, Ales ; Belohlavek, Jan ; Investigators, for the ECMO-CS: Extracorporeal Membrane Oxygenation in the Therapy of Cardiogenic Shock: Results of the ECMO-CS Randomized Clinical Trial. In: *Circulation* 147 (2023), 2, Nr. 6, 454–464. http://dx.doi.org/10.1161/CIRCULATIONAHA.122.062949. – DOI 10.1161/CIRCULATIONAHA.122.062949

[55] Lepot, Mathieu ; Aubin, Jean-Baptiste ; Clemens, François H L R.: Interpolation in Time Series: An Introductive Overview of Existing Methods, Their Performance

Criteria and Uncertainty Assessment. In: *Water* 9 (2017), Nr. 10. http://dx.doi.org/10.3390/w9100796. – DOI 10.3390/w9100796. – ISSN 2073–4441

[56] Jeong, Joo H. ; Kook, Hyungdon ; Lee, Seung H. ; Joo, Hyung J. ; Park, Jae H. ; Hong, Soon J. ; Kim, Mi-Na ; Park, Seong-Mi ; Jung, Jae S. ; Yang, Jeong H. ; Gwon, Hyeon-Cheol ; Ahn, Chul-Min ; Jang, Woo J. ; Kim, Hyun-Joong ; Bae, Jang-Whan ; Kwon, Sung U. ; Lee, Wang S. ; Jeong, Jin-Ok ; Park, Sang-Don ; Lim, Seong-Hoon ; Yu, Cheol W.: Predictors of in-hospital mortality after successful weaning of venoarterial extracorporeal membrane oxygenation in cardiogenic shock. In: *Scientific Reports* 13 (2023), Nr. 1, 17529. http://dx.doi.org/10.1038/s41598-023-44679-2. – DOI 10.1038/s41598–023–44679–2. – ISSN 2045–2322

[57] Wilson-Smith, Ashley R. ; Bogdanova, Yulia ; Roydhouse, Stephanie ; Phan, Kevin ; Tian, David H. ; Yan, Tristan D. ; Loforte, Antonio: Outcomes of venoarterial extracorporeal membrane oxygenation for refractory cardiogenic shock: systematic review and meta-analysis. In: *Annals of cardiothoracic surgery* 8 (2019), 1, Nr. 1, S. 1–8. http://dx.doi.org/10.21037/acs.2018.11.09. – DOI 10.21037/acs.2018.11.09. – ISSN 2225–319X (Print)

APPENDICES

A. Outcome parameters

Outcome parameters	Option	Option meaning	Definition
ICU Mortality	0	No	The patient did not die during this ICU stay
	1	Yes	The patient did die during this ICU stay
Second ECMO run	0	No	There was no second run during this ICU stay
	1	Yes	There was a second run during this ICU stay
Reason for discontinutation	0	Recovery	The reason for ECLS discontinuation is expected cardiac and/or pulmonary recovery
	1	Palliation	The reason for ECLS discontinuation is to abstain (discontinuation of ECLS in case of medical futility/ poor prognosis or explicitly requested by patient and/or representatives).
	2	Complication	The reason for ECLS discontinuation is a complication (A complication of ECLS care required withdrawal of ECLS.
	3	VAD	The reason for ECLS discontinuation is transition to ventricular assist device (VAD) support (either HMIII, HW, total articial heart)
	4	Lung transplantation	The reason for ECLS discontinuation is lung transplantation.
	5	Heart transplantation	The reason for ECLS discontinuation is heart transplantation.
	6	Heart/lung	The reason for ECLS discontinuation is combined heart/lung transplantation transplantation.
	7	Death	The reason for ECLS discontinuation is sudden death of the patient whilst on ECLS.
	999	Unknown	The reason for ECLS discontinuation is not documented.

Table A.1: Outcome parameters with the options, the option meaning and the definition. These were used to determine the weaning success category of each ECMO run.

B. Q-Q plots

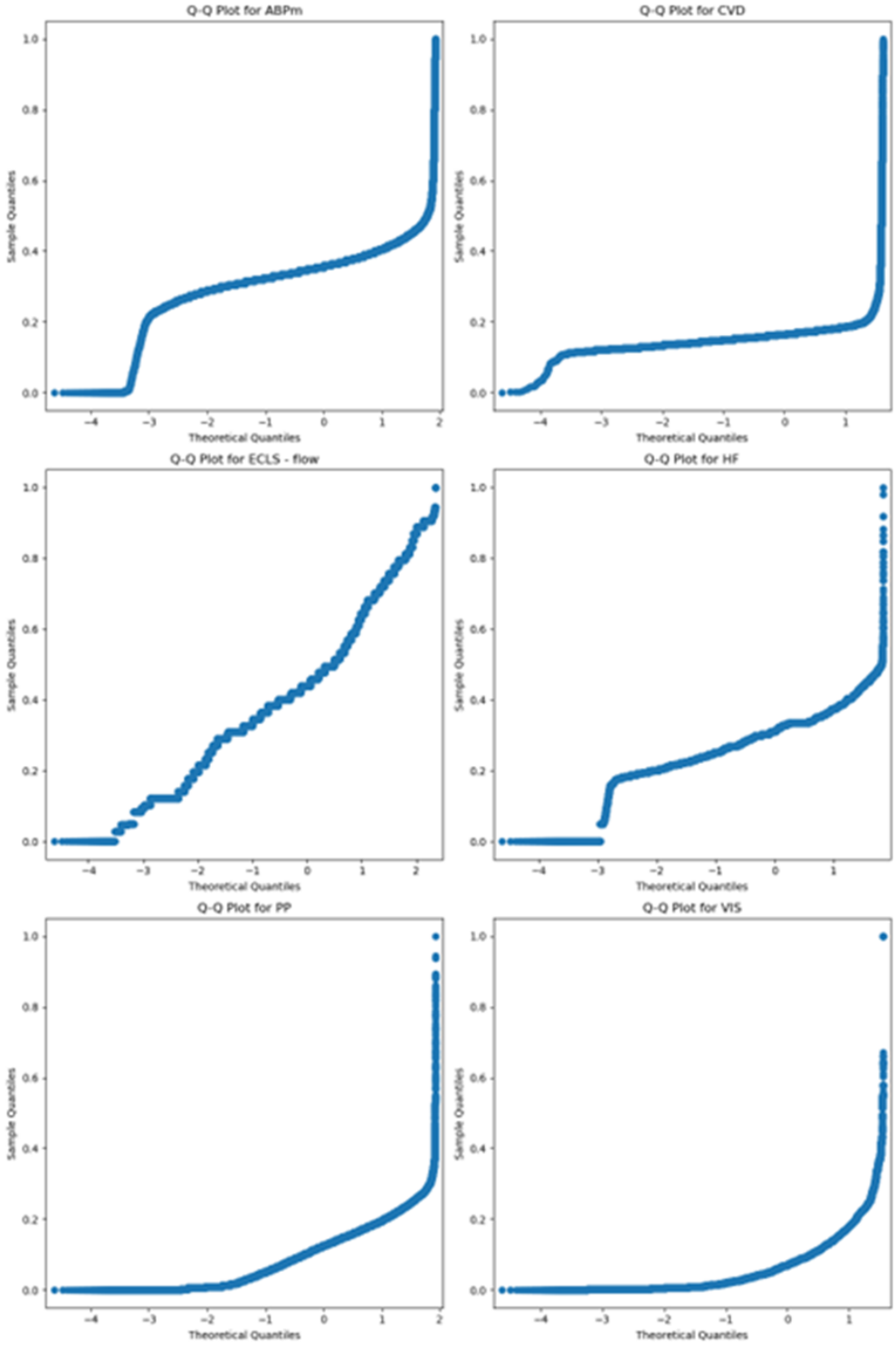


Figure B.1: Q-Q plots of the high frequency parameters (mean arterial pressure (ABPm), central venous pressure (CVD), ECLS flow, heart rate (HF), pulse pressure (PP) and vasoactive Inotropic score (VIS)) after normalising the data between 0 and 1.

C. Parameter grid used in Grid search

Hyperparameter	Values	Chosen value
N-estimators	Between 100 and 1500 with steps of 100	1000
Max depth	None, 10, 20, 30	None
Min samples split	2, 5, 10	10
Min samples leaf	1, 2, 4	1
Max features	Sqrt, log2, None	None

Table C.1: Hyperparameter grid for Random forest with chosen hyperparameters.

Hyperparameter	Values	Chosen value
Kernel	Linear, rbf, poly	rbf
C	1.e-03 1.e-02, 1.e-01, 1.e+00, 1.e+01, 1.e+02 1.e+03	100
Gamma	Scale, auto	auto
Epsilon	0.1, 0,2, 0.3, 0.4, 0.5	0.4

Table C.2: Hyperparameter grid for Support Vector Regressor with chosen hyperparameters.

Hyperparameter	Values	Chosen value
Solver	Liblinear	Liblinear
C	30 values between 10^{-4} and 10^{4} evenly spaced on log scale	0.1
Penalty	L1, l2	L1

Table C.3: Hyperparameter grid for Logistic regression with chosen hyperparameters.

Hyperparameter	Values	Chosen value
N-neighbours	3,5,10,15,20	5
Weights	Uniform, distance	Distance

Table C.4: Hyperparameter grid for K- nearest neighbours with chosen hyperparameters.

Hyperparameter	Values	Chosen value
N-estimators	Between 100 and 1500 with steps of 100	200
Learning rate	0.01, 0.05,0.1, 0.2, 0.3	0,01
Max depth	3, 6, 9, 12, 15, 18, 21, 24, 27, 30	6
Min samples split	2,5,10,15	15
Min samples leaf	1,2,4	4
Max features	Sqrt, log2, None	None

Table C.5: Hyperparameter grid for Gradient boosting with chosen hyperparameters.

D. Correlation matrix

D.1 Full correlation matrix

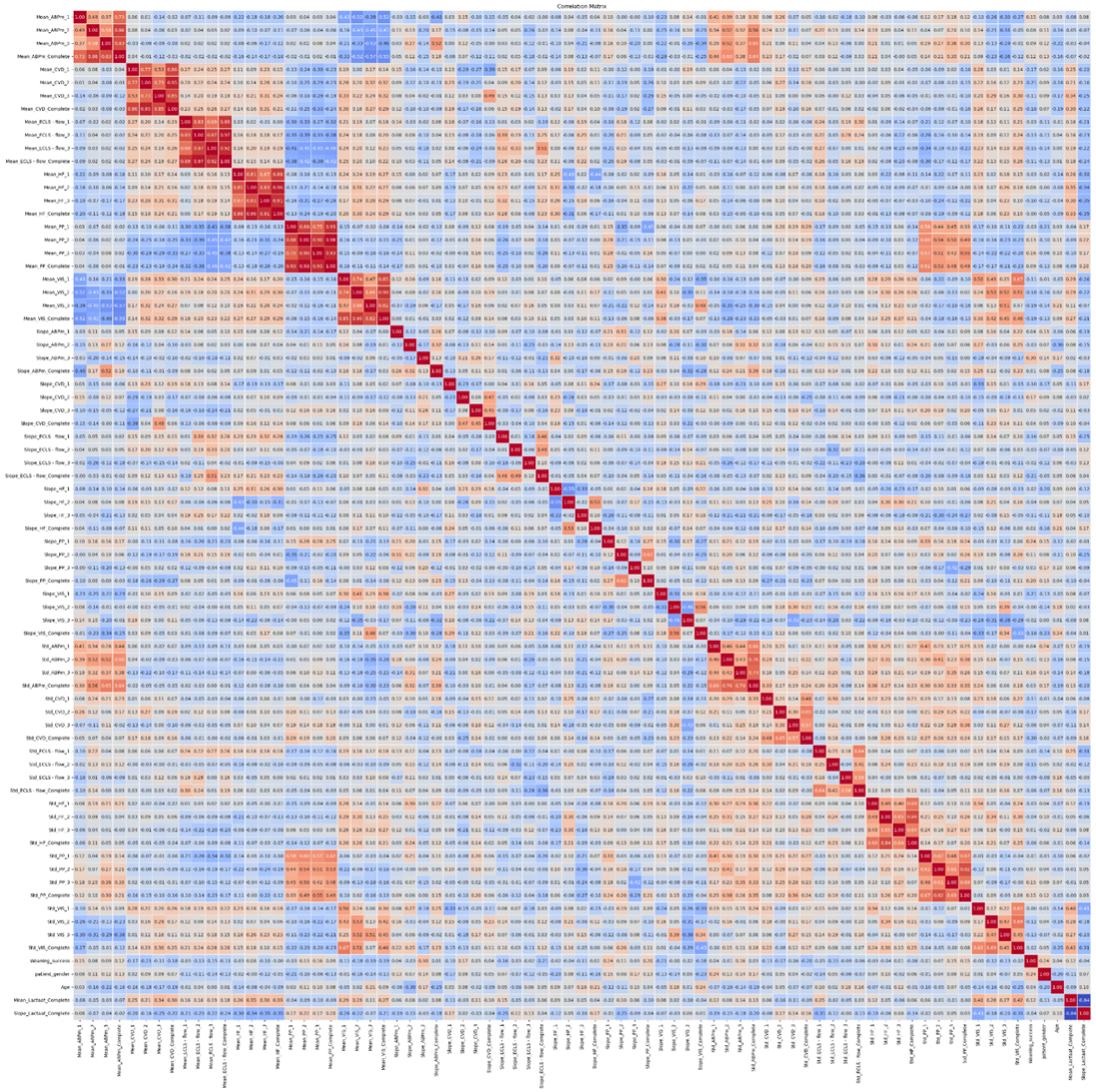


Figure D.1: Correlation matrix based on Pearson correlation. A correlation of 1.0 signifies a full positive correlation, a correlation of -1.0 a negative correlation. A correlation value of 0.0 means there was no correlation found between the parameters. Blue signifies a correlation below 0, and thus negative, red signifies a correlation above 0 and thus positive. The features are named based on the calculation (slope, mean or std (standard deviation)), the abbreviation of the parameter. The number behind the feature name corresponds with the day of ECMO treatment, 'complete' refers to the full 72 hours as interval.

D.2 Filtered correlation matrix

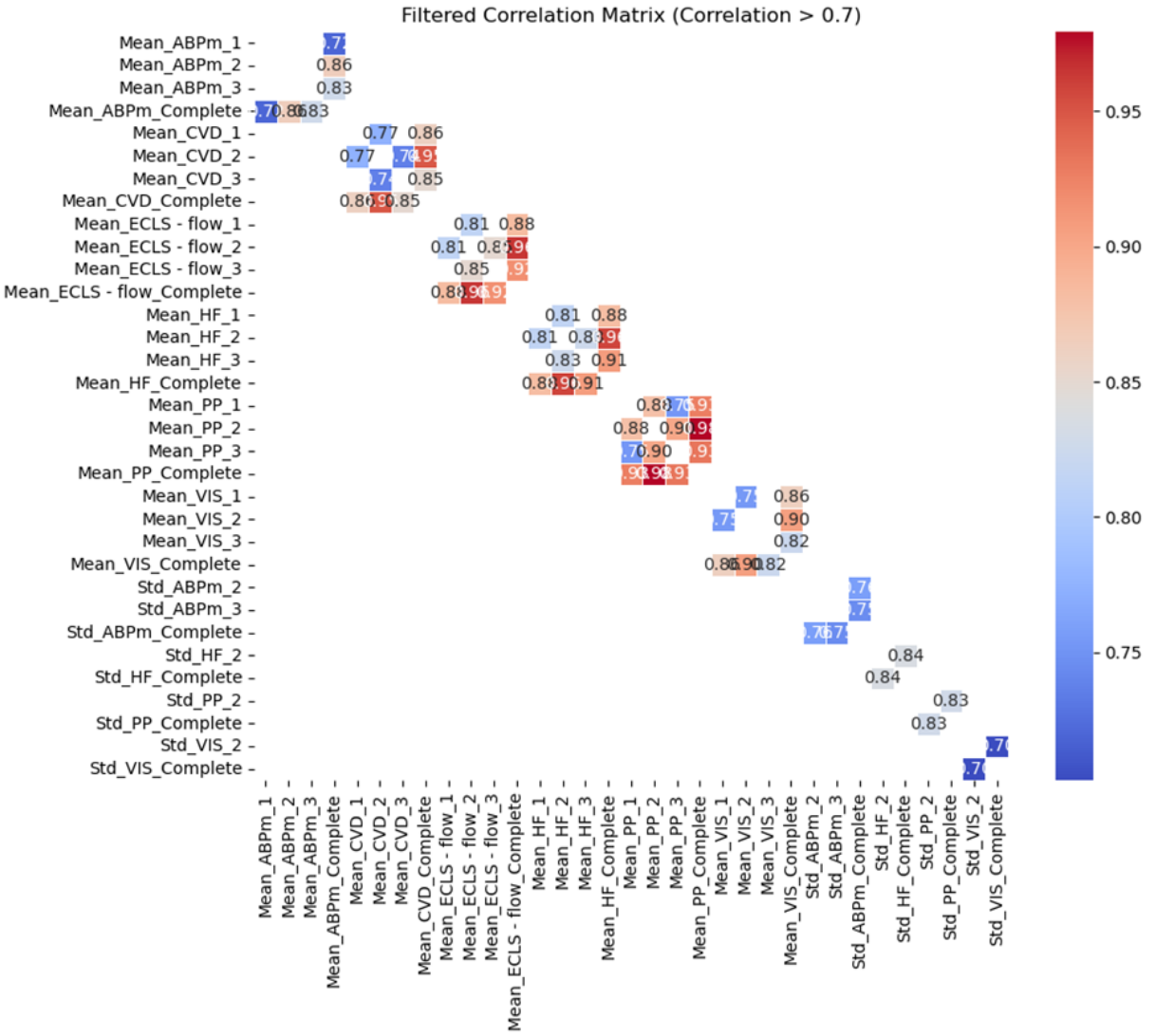


Figure D.2: Filtered correlation matrix based on a correlation higher than 0.7. No values for -0.7 were detected. A correlation of 1.0 signifies a full positive correlation, a correlation of -1.0 a negative correlation. A correlation value of 0.0 means there was no correlation found between the parameters. Blue signifies a correlation below 0, and thus negative, red signifies a correlation above 0 and thus positive. The features are named based on the calculation (slope, mean or std (standard deviation)), the abbreviation of the parameter. The number behind the feature name corresponds with the day of ECMO treatment, 'complete' refers to the full 72 hours as interval.

E. List of selected features

E.1 Overview of the included features per machine learning model

Support vector Regressor *22 features*	**K-Nearest Neighbours** *19 features*	**Random Forest** *26 features*	**Gradient Boosting** *11 features*	**Logistic Regression** *9 features*
Mean lactate complete	Mean lactate complete	Slope MAP 3	Slope PP Complete	Slope PP Complete
Patient gender	Patient gender	Slope ECLS flow complete	Std VIS 1	Slope HF 2
Std VIS 3	Std VIS 2	Slope VIS complete	Mean VIS 3	Slope CVD 3
Std VIS 2	Std HF 3	Slope PP Complete	Slope VIS 2	Slope VIS 1
Std ECLS flow 3	Std HF 1	Slope VIS 1	Std CVD 2	Mean PP 1
Std ECLS flow 1	Std ECLS flow 3	Slope PP 1	Std HF 2	Slope HF 1
Std CVD 1	Std ECLS flow 1	Slope PP 3	Std VIS 3	Std CVD 1
Slope VIS 1	Slope VIS 2	Mean PP 3	Mean PP 1	Mean HF 1
Slope PP Complete	Slope PP Complete	Slope PP 2	Slope MAP 1	Slope VIS 2
Slope PP 1	Slope PP 3	Std VIS 2	Slope HF 2	
Slope HF 3	Slope HF 3	Slope MAP Complete	Std HF 3	
Slope HF 2	Slope HF 2	Slope HF 1		
Slope HF 1	Slope HF 1	Std HF 2		
Slope ECLS flow 2	Slope ECLS flow 3	Slope CVD Complete		
Slope CVD Complete	Slope CVD 3	Slope HF 3		
Slope CVD 3	Slope MAP 1	Mean ECLS flow 3		
Slope CVD 2	Mean VIS 3	Slope ECLS flow 1		
Slope MAP 1	Mean PP 1	Slope CVD 3		
Mean VIS 3	Mean ECLS flow 3	Slope VIS 2		
Mean PP 1		Std ECLS flow 3		
Mean HF 1		Std HF 1		
Mean ECLS flow 3		Mean VIS 3		
		Slope CVD 2		
		Std ECLS flow 2		
		Slope CVD 1		
		Std HF 3		

Table E.1: Overview of the result of the feature selection with the amount of features selected per model and the selected features in order of importance. Most important features are at the top. The number behind the feature abbreviation refers to the day of the ECMO treatment. In case the feature name ends in 'complete' instead of a number, it is the feature for the full 72 hours.

F. Feature selection per algorithm

F.1 Support vector Regressor

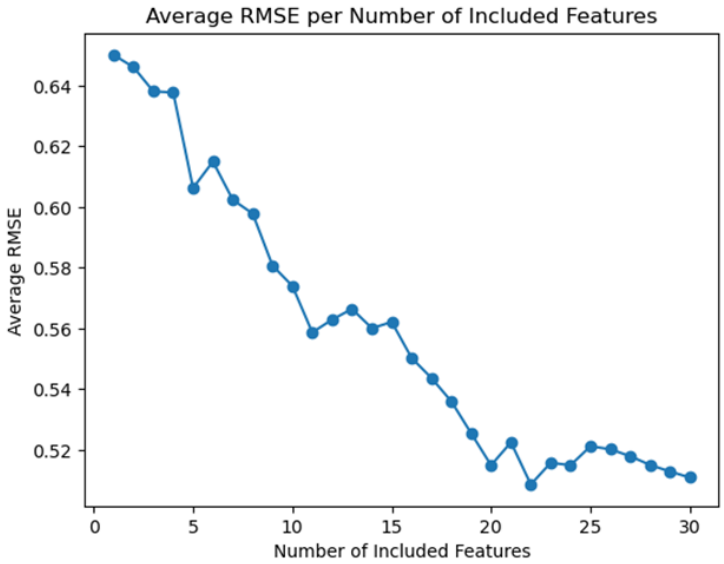


Figure F.1: Average Root Mean Squared Error of the five outer folds per number of included features in the SVR model.

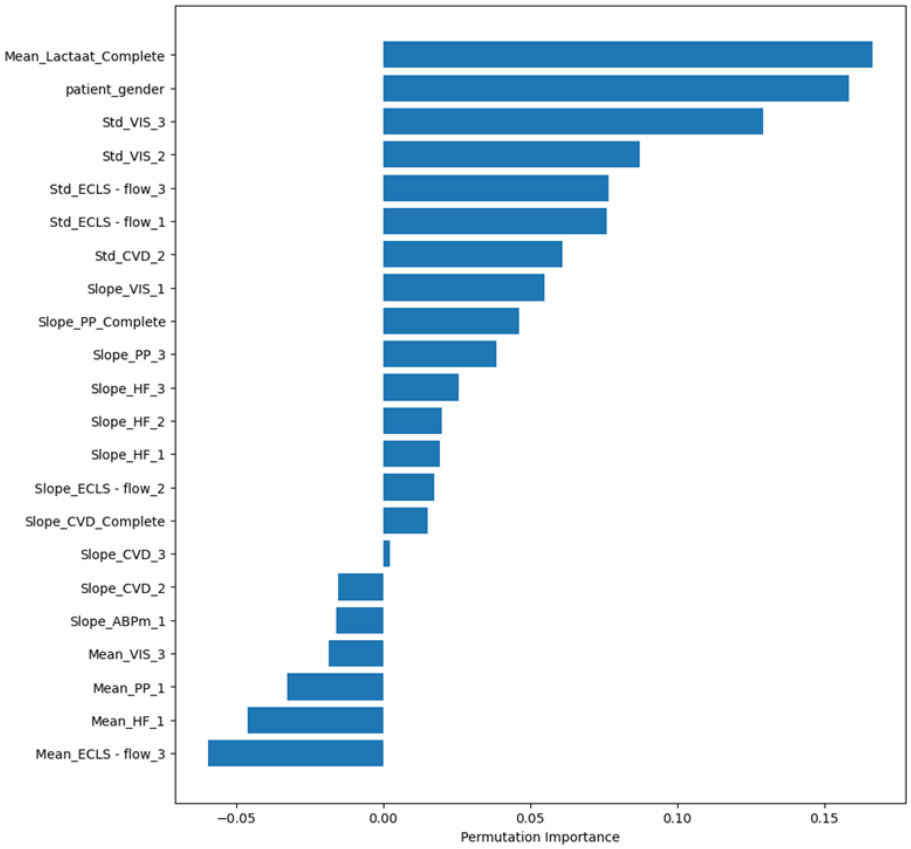


Figure F.2: Permutation importance of the included features of the SVR model. The features are named based on the calculation (slope, mean or std (standard deviation)) and the abbreviation of the parameter. The number behind the feature name corresponds with the day of ECMO treatment, 'complete' refers to the full 72 hours as interval.

F.2 Gradient Boosting

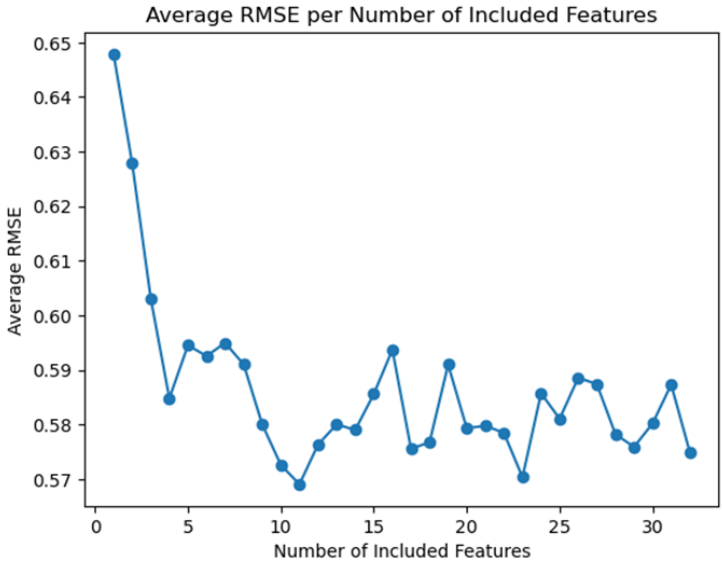


Figure F.3: Average Root Mean Squared Error of the 5 folds per number of included features GB.

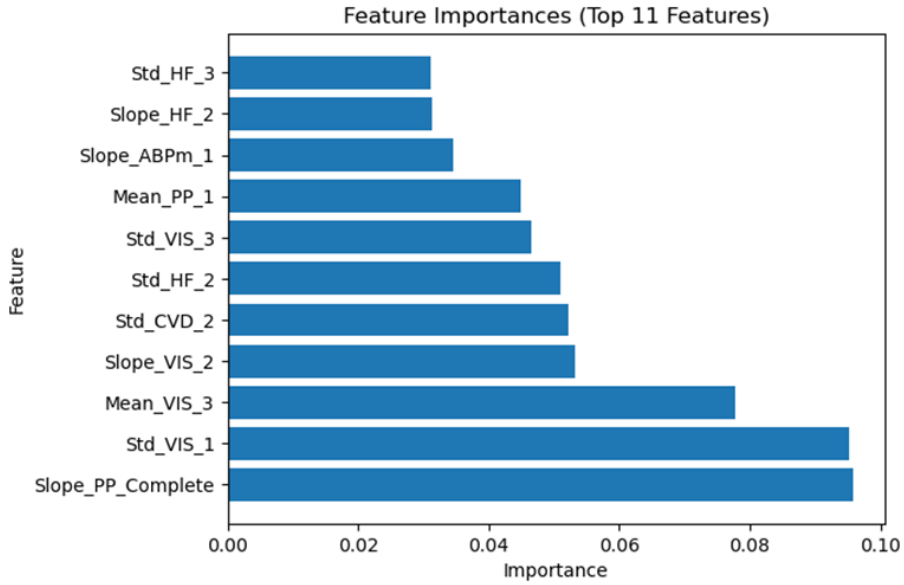


Figure F.4: Gini importance of the included features of the gradient boosting model. The features are named based on the calculation (slope, mean or std (standard deviation)) and the abbreviation of the parameter. The number behind the feature name corresponds with the day of ECMO treatment, 'complete' refers to the full 72 hours as interval.

F.3 Random Forest

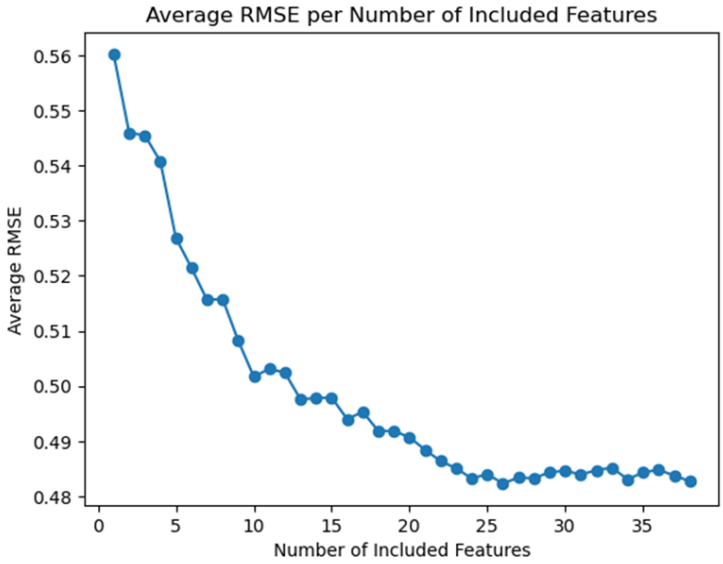


Figure F.5: Average Root Mean Squared Error of the 5 folds per number of included features RF.

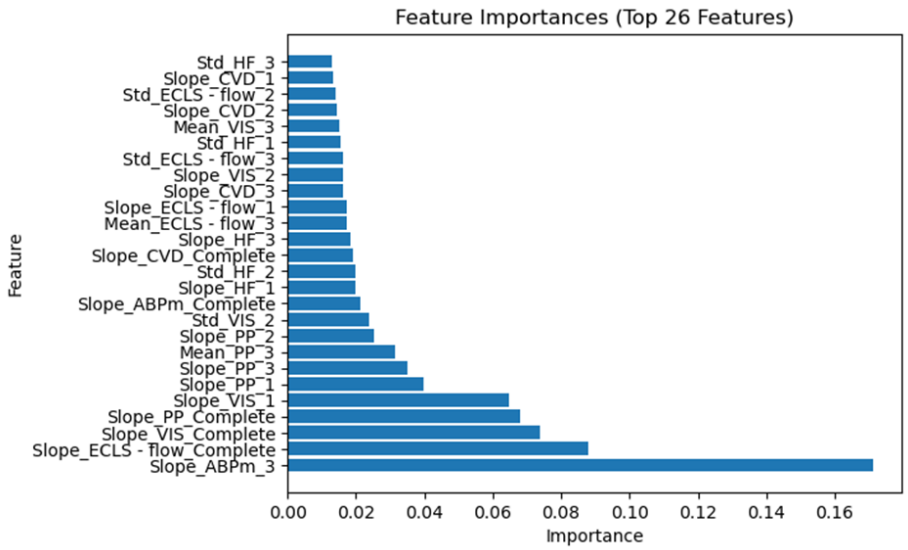


Figure F.6: Gini importance of the included features of the RF model. The features are named based on the calculation (slope, mean or std (standard deviation)) and the abbreviation of the parameter. The number behind the feature name corresponds with the day of ECMO treatment, 'complete' refers to the full 72 hours as interval.

F.4 K-Nearest Neighbours

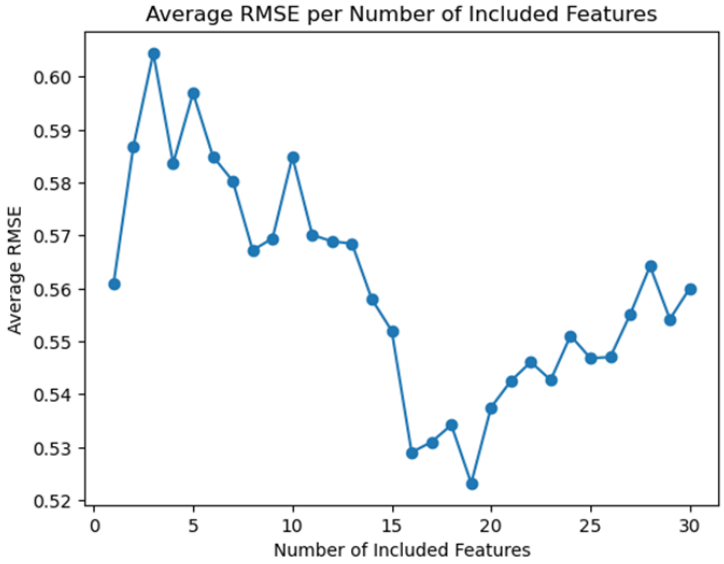


Figure F.7: Average Root Mean Squared Error of the 5 folds per number of included features KNN.

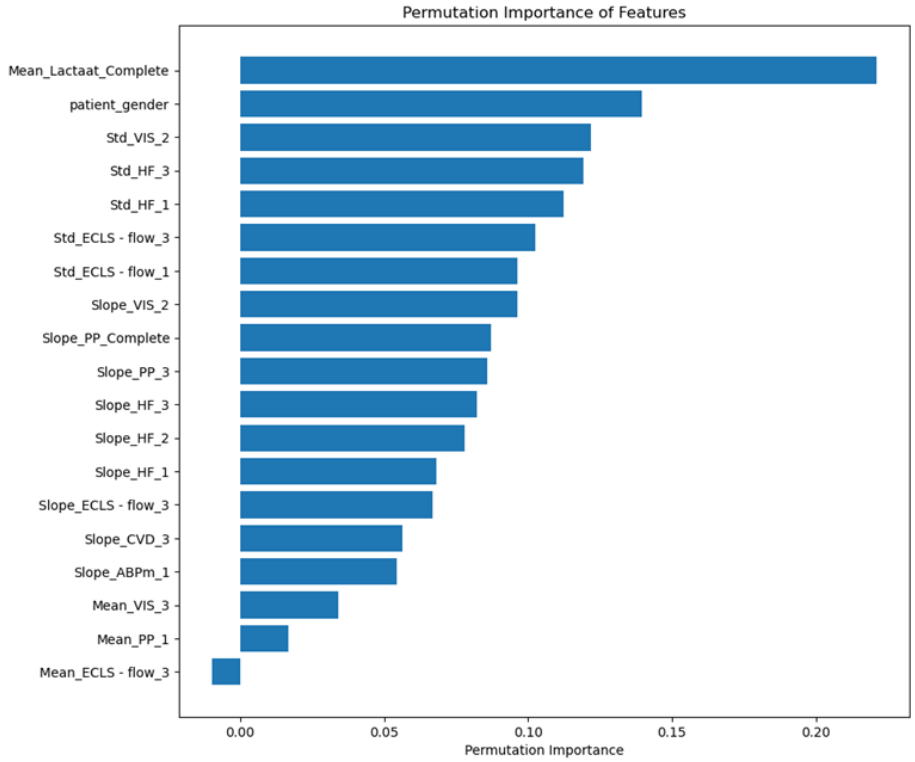


Figure F.8: Permutation importance of the included features of the KNN model. The features are named based on the calculation (slope, mean or std (standard deviation)) and the abbreviation of the parameter. The number behind the feature name corresponds with the day of ECMO treatment, 'complete' refers to the full 72 hours as interval.

F.5 Logistic regression

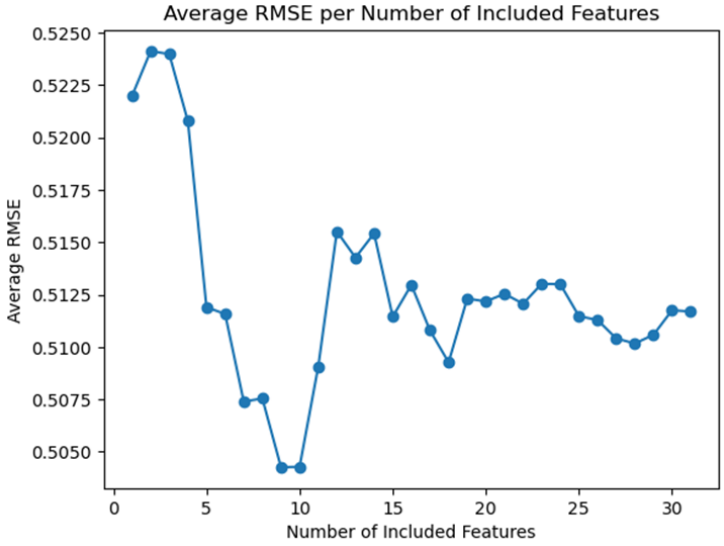


Figure F.9: Average Root Mean Squared Error of the 5 folds per number of included features of the logistic regression model.

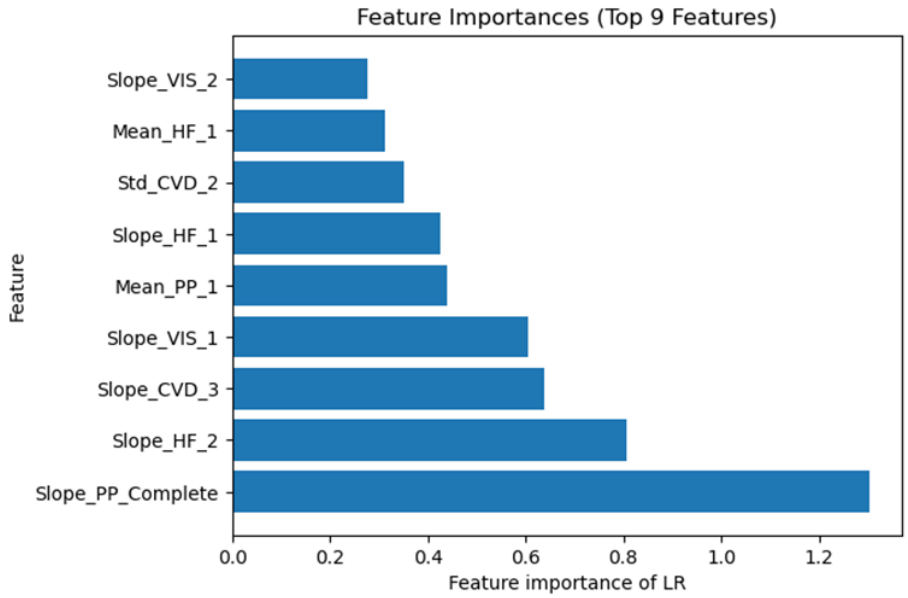


Figure F.10: Feature importance of the included features of the logistic regression model. The features are named based on the calculation (slope, mean or std (standard deviation)) and the abbreviation of the parameter. The number behind the feature name corresponds with the day of ECMO treatment, 'complete' refers to the full 72 hours as interval.

G. List of removed features using VIF

1. Mean ECLS - flow Complete
2. Mean PP Complete
3. Mean CVD Complete
4. Mean MAP Complete
5. Mean HF Complete
6. Std MAP Complete
7. Mean VIS Complete
8. Mean ECLS - flow 1
9. Mean PP 2
10. Mean CVD 3
11. Std PP Complete
12. Mean HF 2
13. Mean VIS 1
14. Mean ECLS - flow 2
15. Mean PP 1
16. Mean MAP 3
17. Std CVD Complete
18. Std MAP 3
19. Mean MAP 1
20. Std VIS Complete
21. Mean CVD 1
22. Std MAP 1
23. Std MAP 2
24. Mean HF 1
25. Std HF Complete
26. Std PP 2
27. Std ECLS - flow Complete
28. Mean VIS 2
29. Mean Lactaat Complete
30. Mean MAP 2
31. Std PP 1
32. Std PP 3
33. Mean CVD 2
34. Std CVD 3
35. Std CVD 1
36. Std CVD 2
37. Std VIS 1