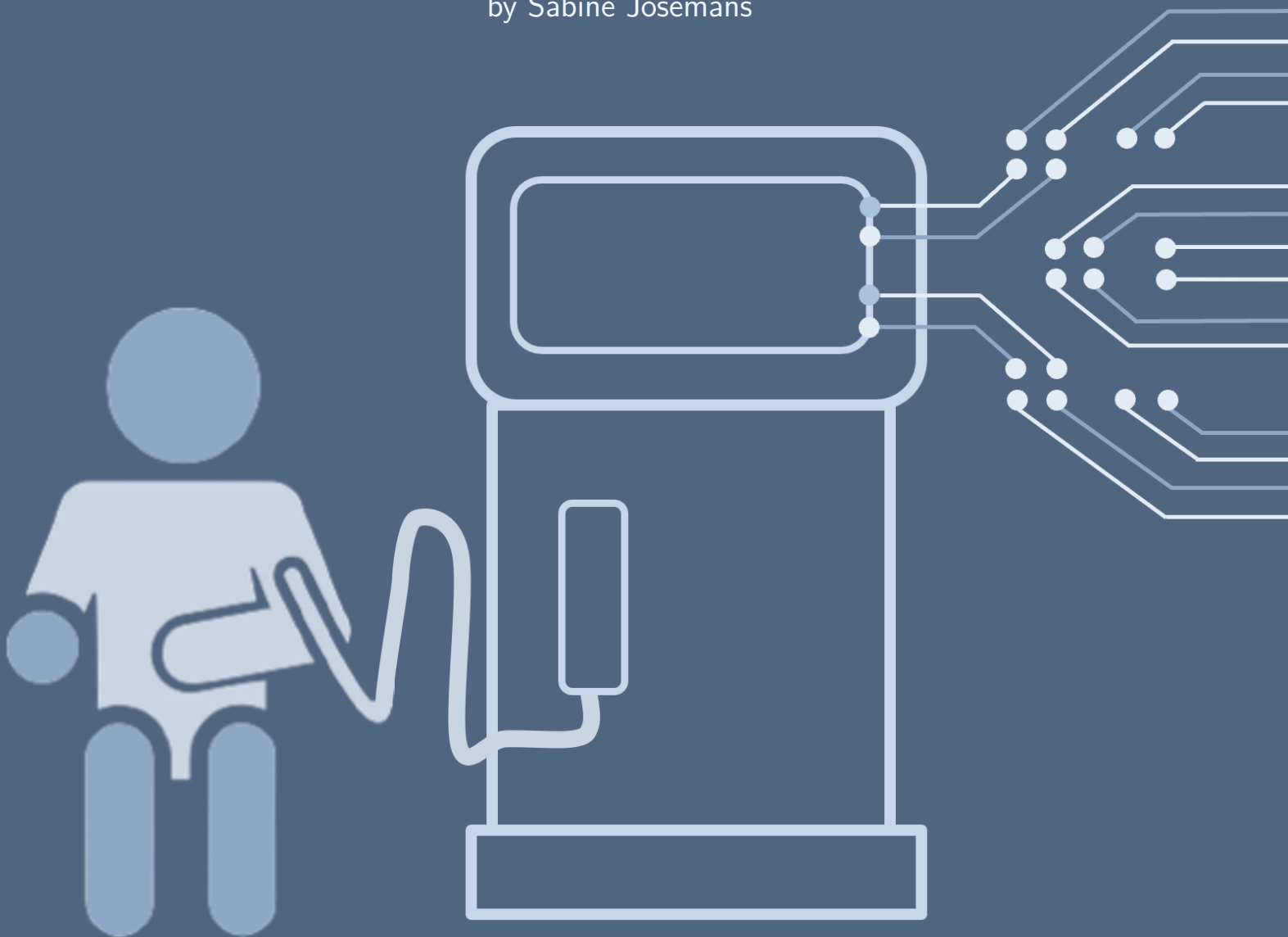
Machine learning for real-time prediction of intradialytic hypotension

Technical Medicine master thesis

by Sabine Josemans



Universiteit Leiden

TU Delft

Erasmus University Rotterdam

Machine learning for real-time prediction of intradialytic hypotension

Sabine Josemans

Student number: 4535022

25 Sep 2023

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University - Delft University of Technology - Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)
Department of Nephrology, LUMC
10 Oct 2022 - 9 Oct 2023

Supervisors
Prof. dr. Joris I. Rotmans
Dr. David M.J. Tax

Thesis committee members
Prof. dr. Joris I. Rotmans, LUMC, Department of Nephrology
Dr. David M.J. Tax, TU Delft, Pattern Recognition Laboratory
Dr. Karin G.F. Gerritsen, UMCU, Department of Nephrology

Preface and Acknowledgements

With this thesis, I conclude my study program in Technical Medicine and thereby my student time in Delft. When I first showed up at the Nephrology department in the LUMC, I was enthusiastically received and invited to dive into the world of technical innovations regarding fluid management of hemodialysis patients. While many different ideas emerged over time, the final choice of subject for this master thesis fell on the development of machine learning algorithms, with special interest in a recurrent neural network, to predict intradialytic hypotension and assess the feasibility of these predictions as a tool to assist in fluid management. In my opinion, this project aligns well with the scope of Technical Medicine, as it aims to integrate these technical models in clinical context.

Throughout the course of this project, I acquired valuable knowledge on launching a new project from scratch. This gave me the opportunity to gain experience in what it is like to be a researcher. What greatly supported me in this period, is that I was completely included in the research group of my supervisor, Joris Rotmans, to whom I want to express my sincere gratitude. Joris also introduced me at the dialysis department, where I had the opportunity to become familiar with the dialysis practices in the LUMC and, not to forget, the hemodialysis patients. Since hemodialysis patients are subjected to treatment several times a week, it is not hard to gain insight in the burden that hemodialysis has on their daily lives. This greatly motivated me to contribute towards enhancing the outcomes for this patient population.

Development of the machine learning models would not have been possible without the supervision of David Tax, my technical supervisor from the TU Delft. He put a lot of time in elaborating on important concepts regarding this topic. Designing a recurrent neural network was completely new to me, but I now feel I have acquired sufficient knowledge to continue working with this type of network in future research.

I am convinced that all the knowledge and experience I have gained during this project will aid me in my future work, where I hope to bring technical innovations in clinical practice to contribute to the improvement of patient outcomes.

Sabine Josemans
September 2023

Abstract

Introduction
In the Netherlands, approximately five thousand people are dependent on maintenance hemodialysis (HD) treatment. The one-year mortality of this patient population in 2021 was as high as 18%. A significant contributor to this high mortality and burden on HD patients is the high incidence of intradialytic hypotension (IDH), which can lead to hypoperfusion of vital organs. Despite advances in fluid management, including technical innovations such as relative blood volume (RBV) monitoring and bioimpedance measurements, the incidence of IDH remains unacceptably high. Therefore, the primary goal of this master thesis is to assess the feasibility of machine learning-based real-time prediction of IDH using patient data from HD sessions at the LUMC with the aim to enhance individualized fluid assessment. In this study, the performance of a recurrent neural network, known for its ability to retain time-varying information, will be compared to other machine learning models.

Methods
Patient demographics and intradialytic measurements, such as blood pressure and RBV measurements, were obtained from all HD patients of the LUMC with an age $\geq$18 years. The two definitions of IDH that were assessed were defined as IDH1, a drop of $\geq$20 mmHg compared to the pre-dialytic SBP and IDH2, is an absolute intradialytic SBP of $\leq$90 mmHg. Performance of a logistic regression, random forest and extreme gradient boosting (XGBoost) model were assessed and compared to that of a recurrent neural network (RNN). The models provided a prediction of IDH at every 30 minutes during HD treatment. Feature importances were determined for the logistic regression, random forest and XGBoost and two subanalyses were conducted to assess predictions made 60 minutes ahead instead of 30 minutes, and the effect of removal of blood pressure measurements on the prediction.

Results
37,025 HD sessions of 436 patients were included in the analysis of IDH1 and 43,722 sessions of 441 patient in the analysis of IDH2. The Area under the Curve (AUC) of the Receiver Operating Characteristics (ROC) curve for the logistic regression, random forest, XGBoost and RNN were 0.80, 0.81, 0.81 and 0.87 respectively for the prediction of IDH1 and 0.93, 0.84, 0.86 and 0.92 for IDH2. At the highest clinically acceptable false positive rate (FPR) of 0.1, the RNN demonstrated a precision, recall, and specificity of 0.56, 0.59, and 0.90, respectively, for IDH1, and 0.10, 0.76, and 0.90 for IDH2, and the logistic regression 0.63, 0.50 and 0.90 for IDH1 and 0.14, 0.80 and 0.90 for IDH2.

Conclusion
The machine learning algorithms showed the ability to learn from HD data to make a prediction of IDH at different intradialytic timepoints. However, the current models do not meet the criteria for clinical implementation due to their limited performance metrics. Evaluation of the subanalyses revealed that the RNN had the ability to detect patterns that did not only rely on the blood pressure. Therefore, the prospect of real-time predictions of IDH appears promising with further refinement of the RNN and the expansion of the database by incorporating additional features and sessions.

List of Abbreviations

AUC	Area under the curve
BCM	Body composition monitor
BIA	Bioimpedance analysis
BMI	Body mass index
CKD	Chronic kidney disease
ECV	Extracellular volume
EHR	Electronic health record
ESKD	End stage kidney disease
FPR	False positive rate
GRU	Gated recurrent unit
HD	Hemodialysis
ICV	Intracellular volume
IDH	Intradialytic hypotension
RBV	Relative blood volume
RNN	Recurrent neural network
ROC	Receiver operating characteristic
SBP	Systolic blood pressure
SQL	Structured query language
TPR	True positive rate
UF	Ultrafiltration
UFR	Ultrafiltration rate
UFV	Ultrafiltration volume
XGBoost	Extreme gradient boosting
IDH1	Intradialytic hypotension defined as a drop of ≥ 20 mmHg compared to the pre-dialytic systolic blood pressure
IDH2	Intradialytic hypotension defined as an absolute intradialytic systolic blood pressure $\leq$90 mmHg

Contents

Introduction

Hemodialysis

In the Netherlands, approximately 18 thousand patients were receiving renal replacement therapy in the year 2021 [1]. Of these patients, five thousand were dependent on hemodialysis (HD), signifying that in this year, 299 out of 1 million individuals from the Dutch population were on maintenance HD. The one-year mortality of this patient population was as high as 18%, which is consistent with the European one-year mortality of HD patients that was reported to be 15.6% [1, 2]. Two major factors contributing to these high mortality rates arise from the challenges regarding fluid management of patients on HD. The first factor is the high occurrence of cardiovascular disease, with presentations such as heart failure and stroke due to atherosclerosis. This cardiovascular risk is elevated in end-stage kidney disease (ESKD) patients due to excessive fluid accumulation and the associated hypertension [3]. The second factor related to the high mortality rate and burden on HD patients is the high prevalence of dialysis-induced hypotension, which has a reported incidence of up to 40% [4–8].

Addressing the complexities of fluid management in HD patients challenging. Water in the human body is distributed over the intracelullar volume (ICV) and the extracellular volume (ECV), of which the latter consists of intravascular plasma and interstitial fluid [9, 10]. As 90% of the body sodium appears in the ECV, this volume is greatly determined by the total sodium content. In renal failure, sodium excretion is inadequate and the resulting increase in ECV leads to hypertension and fluid overload. When maintained for a longer period of time, it is known to cause pulmonary congestion, left ventricular hypertrophy and dilatation and is associated with a higher risk of mortality [11–14]. Consequently, in addition to removal of toxins from the blood, one of the main goals of HD treatment is the removal of excess ECV in order to achieve the adequate "dry-weight" of the patient. This process is known as ultrafiltration (UF) with dry weight referring to the ideal post-dialytic weight [15].

Fluid assessment of dialysis patients in the LUMC

Achieving optimal fluid management requires a multiparametric approach. In current practice regarding fluid management at the HD department in the LUMC, the estimation of dry weight and required ultrafiltration rates (UFRs) are based on the interdialytic weight gain, clinical signs, pre- and intradialytic blood pressure measurements, continuous relative blood volume (RBV) monitoring and measurements with the Body Composition Monitor (BCM) on indication (Figure 1).

The BCM provides a measure of the amount of overhydration as measured by bioimpedance. This monitor applies bioimpedance analysis (BIA) by applying a small current through skin electrodes and measuring the total body impedance. Impedance, being inversely related to body water volumes, serves as a metric for assessing both intracellular and extracellular fluid. To perform these calculations, the BCM integrates mathematical modeling and specific equations that consider variables such as body height [16]. BIA and has been proven to be very useful in following changes in ECV over time, but is less reliable as an individual measurement of ECV as a consequence of the many factors of influence on the measurements [17, 18].

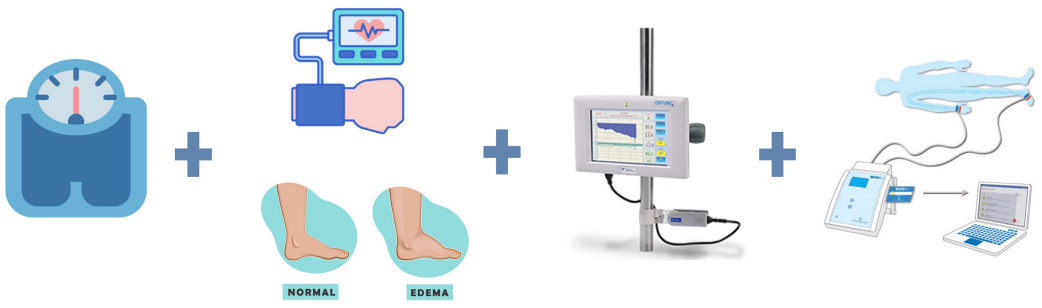


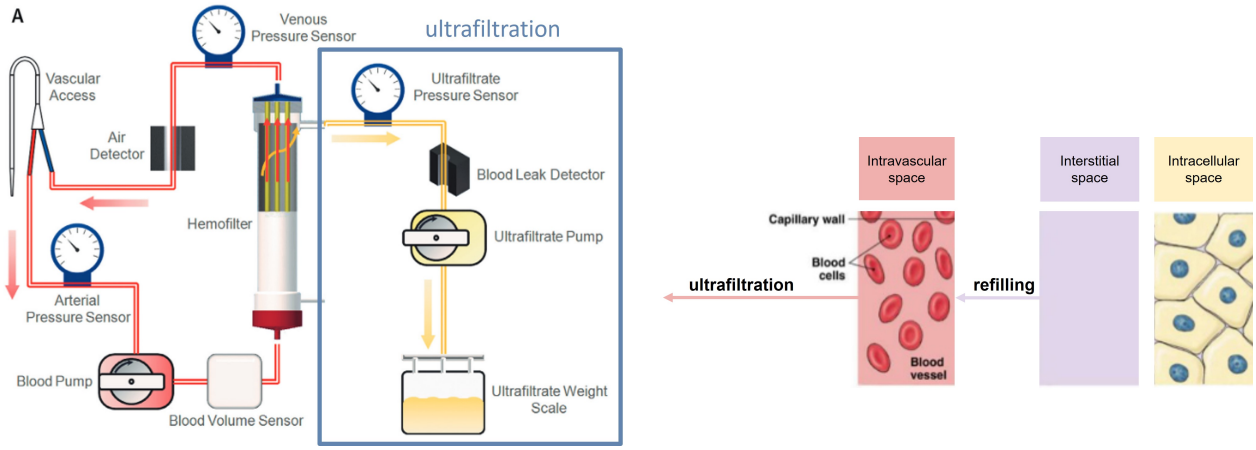
Figure 1: Current clinical practice for dry weight estimation. Includes from left to right: interdialytic weight gain, blood pressure measurements and clinical signs of edema, relative blood volume monitoring and bioimpedance analysis.

With RBV monitoring, the changes in blood volume are continuously measured during HD. Relative blood volume can be measured by use of several blood components, such as hemoglobin, hematocrit or plasma proteins and is calculated as a percentage by the following formula [19, 20]

$$RBV(\%) = ((C_0 - C_t) - 1) \times 100$$

where C_0 is the concentration of the blood compartment at the start of HD and C_t the concentration at any point during HD. Taking hematocrit as example, the monitor measures the blood levels by absorption of near-infrared light at the extracorporal circuit [21]. RBV monitoring is considered to be useful for determining safe margins for fluid removal and avoid IDH [22]. However, studies on RBV guided UF and IDH occurrence remain controversial [19].

At every individual HD session, the manually set UFR determines the amount of ECV that is extracted from the intravascular space. When water is removed from the intravascular space, the changes in hydrostatic and oncotic pressure gradients cause a shift from interstitial fluid to the vascular space (see Figure 2b for a schematic overview). Nevertheless, should the UFR exceed the vascular refilling rate, there will be a decline in cardiac output due to reduced venous return [23]. Initially, the UF induced hypovolemia will activate compensatory mechanisms such as an increase in the heart rate [24]. However, when these mechanisms prove insufficient, the blood pressure will drop and patients will experience intradialytic hypotension (IDH) which can be accompanied by symptoms such as dizziness, syncope, muscle cramps, nausea, chest pain, abdominal pain and a feeling of restlessness. [4, 25]. Despite the multiparametric approach that incorporates different monitoring techniques, IDH is still encountered frequently at the dialysis department in the LUMC.



(a) Hemodialysis and ultrafiltration.

(b) Vascular refilling.

Figure 2: Schematic presentation of hemodialysis with ultrafiltration (a) and vascular refilling (b).

Intradialytic hypotension

The effects of IDH extent beyond the evident discomfort during HD sessions. IDH can lead to hypoperfusion of vital organs, such as the heart, brain and kidney [9]. This can have many consequences, such as faster deterioration of the residual kidney function, increased risk of myocardial stunning and cerebral ischemia [9, 26]. A recent study has shown that ischemic brain injury, caused by reduced brain blood flow, can already occur within one single dialysis session [27]. In patients with frequent occurrence of IDH, an association has been found with an increased 5 years risk of new-onset dementia and on the long term, high occurrence of IDH is associated with increased risk of mortality [4, 28]. Identified factors that are of influence on the occurrence of IDH are presence of diabetes, certain medications, life style, malignancies, presence of heart failure, autoimmune overactivity, hemoglobine levels, hypoalbumina, older age, female sex lower body weight and longer dialysis vintage (time on dialysis) [24, 29–31]. The assessment and prevention of IDH are complicated by the lack of consensus on a clear definition. Depending on the used definition, the prevalence of IDH ranges from 5% to 40% [7].

Interventions that are frequently applied when patients are experiencing hypotension, are positioning of the patient in Trendelenburg, administration of a bolus of saline, reduction of the UFR or termination of UF or dialysis treatment [25]. However, frequent early termination of dialysis treatment will result in the patient receiving an inadequate dialysis dose

and untreated fluid accumulation. This indicates the complexity of managing the fluid balance in HD patients. On the one hand, high UFRs are required to prevent cardiovascular disease by removal of all excess ECV within the treatment duration, but on the other hand, high UFRs can lead to IDH, which can result in direct negative health consequences and premature termination of HD treatment.

The amount of factors associated with IDH and the many monitoring techniques that are applied for prevention emphasize the complexity of assessing individualized risks. When determining the target UF volume (UFV) and UFR, clinicians can not take all these factors in consideration. Therefore, a possible explanation for the high occurrence of IDH is the large amount of parameters that are of influence in this process and the many hidden interconnections between these parameters. Consequently, assessment of the risk of IDH is too complex. An individualized fluid assessment that is able to incorporate a large amount of patient variables would therefore be of great added value in assessing a patients risk of IDH. This might be achieved by application of machine learning based prediction models.

Machine learning for the prediction of intradialytic hypotension

Machine learning is a branch of artificial intelligence that can be applied to discover complex patterns within large quantities of data [32]. Machine learning algorithms can generate predictions by learning from data and refining various models through optimization techniques. Applications of machine learning in nephrology have shown to be successful in predicting acute kidney injury and kidney disease progression rates [33, 34]. While application of machine learning in the prediction of IDH is a relatively new field of research, an advantage of HD treatment is that large quantities of data are available as result of the high treatment frequency per patient and the fact that data is automatically measured and stored. This large amount of data lends itself well to the development of machine learning algorithms.

While many papers applying machine learning methods for diagnosis and prediction of chronic kidney disease (CKD) progression have been published over the last 20 years, prediction of IDH with machine learning has mostly appeared in literature in the last five years. In one of the first publications on the prediction of IDH, linear regression was applied on 279 dialysis sessions to analyze the interaction between variables and IDH and a deep neural network was developed in order to find the variables of influence in the occurrence of IDH [35]. The specific outcome of this model was occurrence of rapid IDH, defined as IDH occurring within the first 120 after HD initiation. The most successful model was a 4-factor interaction model with UFV, UFR, UF coefficient and hypertension comorbidity. However, according to the authors, the accuracy of the deep neural network was not sufficient (accuracy 94.97%, true positive rate (TPR) 87.97% and positive predictive value 66.74%).

More accurate results were obtained by the development of a deep learning algorithm based on 261.647 HD records from 9292 patients, all accompanied by blood test results. [36]. The goal of this model was to predict the risk of IDH within one hour at any timepoint during HD. The used features included demographics, vital signs, comorbidities, medications, and blood test results. The authors developed a recurrent neural network (RNN) model with the gated recurrent unit (GRU) in order to process both current and previous datapoints. This model was compared to a multilayer perceptron, light gradient boosting machine, and logistic regression model. These models did not use all the previous data within the current session, but only the current and former time point (mostly 1 hour). The RNN showed sufficient and slightly better performance than the other models with an Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve of 0.94, 0.87 and 0.79, for three different definitions of IDH.

One of the most recent studies aimed to develop a model based on data only from the dialysis machine, without demographic data, in order to assess the application of a completely anonymized database that could be more easily implemented in dialysis machines [37]. The time-varying datapoints were saved every minute and combined with several time-invariant variables such as dialysate temperature and pre-dialysis systolic blood pressure (SBP) in a convolutional neural network. Again, the deep learning model showed a better performance than other machine learning models.

These studies demonstrate the potential of machine learning models, particularly neural networks, for real-time prediction of IDH. Recurrent neural networks are a type of neural networks that are specifically valuable for time series data. Although promising results with a RNN were obtained by Lee et al., the authors stated that the applicability to other patient datasets might be limited due to differences in patient characteristics and parameter acquisition methods [36]. At the HD department in the LUMC, many different parameters, such as the blood pressure, pulse, blood flow rate, and RBV, are automatically stored by the dialysis machines every 30 minutes. The extensive available amount of data in the LUMC provides the opportunity for development of similar machine learning models.

Therefore, the goal of this master thesis is to assess the feasibility of machine learning-based real-time prediction of IDH by use of data from HD sessions from in LUMC and compare the performance of a recurrent neural network with other models.

Logistic regression

Logistic regression is a widely applied statistical machine learning algorithm employed for classification tasks [38]. Unlike linear regression, which predicts continuous numeric values, logistic regression applies a sigmoid function to transform linear combination of input features into a probability between 0 and 1. This value represents the probability of belonging to a certain class. The final prediction is dependent on the threshold set for classification. The classifier will predict a 1 for probabilities above this threshold and 0 for probabilities below this threshold.

Logistic regression is a relatively simple and interpretable method as it assigns coefficients to each feature, signifying their respective contribution to the prediction. However, it assumes a linear relationship between features and the log-odds of the target labels. Consequently, its applicability is somewhat restricted when dealing with complex datasets as complex patterns can not be captured.

Decision tree classifiers

Decision tree classifiers are supervised machine learning algorithms that build trees from combinations of features in order to make a classification [39]. Every node in a tree represents a feature that is used to split the dataset, based on a certain value of this feature. By adding features to a tree, different branches can be created. The last nodes of the tree, after which no more splits occur, are called leaf nodes. A visual representation of a simple decision tree is provided in Figure 3.

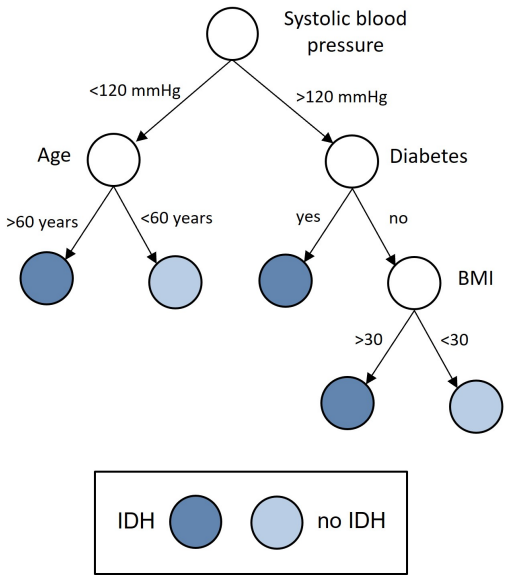


Figure 3: Example of a simple decision tree. Abbreviations: BMI, body mass index; IDH, intradialytic hypotension.

A single leaf node will provide a probability for all the classes in the dataset, based on the final distribution of training samples that were used to build the tree. Consequently, when a new sample is run through the tree and ends in a certain leaf node, the class with the highest probability in that node will be assigned to the sample.

The feature and threshold to use at each node are selected by calculating a measure of impurity or information gain. This is mostly achieved by either the Gini impurity or entropy. The feature that results in the greatest reduction in impurity or error, and thus better classification at each node, is chosen. In the training process, it is important to set certain parameters in order to prevent the trees from growing indefinitely and prevent overfitting. This can be established by tuning hyperparameters such as the maximal tree depth and minimal number of samples at a leaf node. Another generalization technique is bootstrapping, where different subsets of training data are selected when building the trees. This randomness contributes to improved generalization performance.

The two types of decision tree-based classifiers that are applied in the current project are a random forest and an extreme gradient boosting (XGBoost) classifier [40–42]. These classifiers build of a collection of multiple trees, called a forest, that all provide a vote for a class. The final classification of a datapoint is all those votes combined, weighted by the probability estimates. A random forest consists of trees that are all built individually, but a XGBoost classifier builds its forest of trees sequentially, where each new tree is improved based on the errors of the previous trees. Some hyperparameters that should be tuned for an optimal learning process of an XGBoost are the learning rate, the step size by which the trees are adjusted based on the previous errors, the ratio of the training samples to use per boosting iteration and the ratio of features to use for creating the trees.

Random forest classifiers are known for their robustness, are less prone to overfitting than other classifiers due to the ensembling of several trees, can handle missing data and can provide feature importances. However, when applied on complex datasets, they can be computationally expensive and their interpretability is compromised as the prediction is a combination of many different trees.

XGBoost classifiers share some of the advantages of the random forest, such as the ability to handle missing data, but are generally considered to be more efficient and have high predictive accuracy in many classification tasks. Additionally, XGBoost algorithms apply regularization techniques to prevent overfitting. When training an XGBoost classifier, it is important to consider that they are computationally complex and have a wide range of hyperparameters that require tuning.

Recurrent Neural Network

A RNN is a type of deep learning models specifically designed to learn from sequential or time-varying data [43, 44]. These algorithms are capable of retaining memory of previous timepoints and have been applied in a wide range of applications, such as in textual and financial predictions. Generally, deep networks contain neurons that compute

$$y = f(w_1 \cdot x_1 + w_2 \cdot x_2 + b)$$

with x_1 and x_2 representing numerical inputs, w_1 and w_2 weights associated with these inputs and b the bias, a constant added to the sum of the weights. For obtaining output y, the weighted sum of the inputs is transformed by an activation function f in order to introduce non-linearity in the model, thus allowing the neuron to learn complex relationships. A visual representation of a single neuron is provided in Figure 4.

RNNs work with hidden states that save information from the previous timesteps and apply these as input in the new prediction, as is visualized in Figure 5a. The hidden state is updated for every timestep, thus for each element in the sequence, and is calculated as

$$\mathbf{h}_t = tanh(\mathbf{x}_t \cdot \mathbf{W}_{ih}^T + b_{ih} + \mathbf{h}_{t-1} \cdot \mathbf{W}_{hh}^T + b_h h)$$

with $\mathbf{h}_t$ and $\mathbf{x}_t$ the hidden state and input at timestep t, $\mathbf{W}_{ih}$ and b_{ih} the weight matrix and bias that are applied to the input data at each time step to compute the hidden state, $\mathbf{h}_{t-1}$ the hidden state of the previous layer at time $t-1$ and $\mathbf{W}_{hh}$ and b_{hh} the weights and bias applied to the previous hidden state to compute the current hidden state. The output L_t is then calculated as

$$\mathbf{L}_t = \mathbf{W}_{hy} \cdot \mathbf{h}_t$$

with W_{hy} the weight vector at the output layer. The simplified theory behind a RNN is visualized in Figure 5a. Figure 5b illustrates application of a RNN to data of one dialysis session.

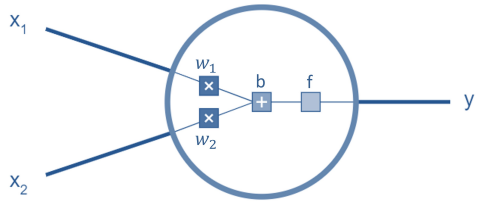
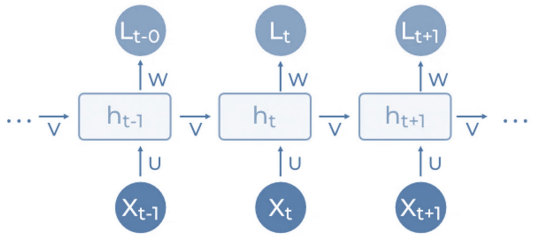
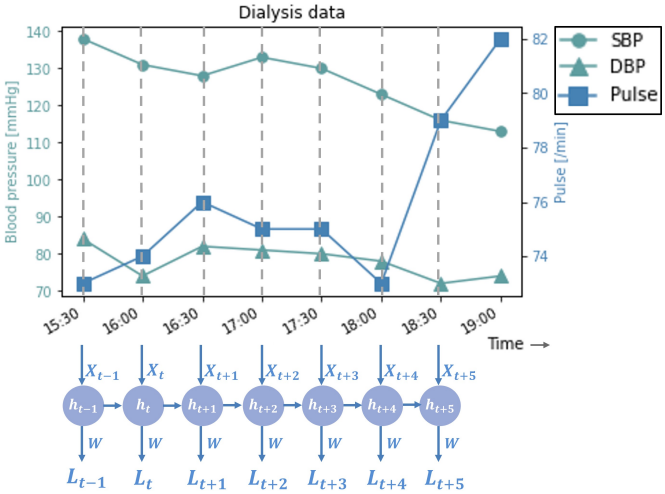


Figure 4: A single neuron where x_1 and x_2 are inputs, w_1 and w_2 the weights applied to the inputs, b the bias, f the activation function and y the output.

When a model is developed and trained, the model's output is compared to the true labels, and the loss is calculated. The loss indicates how "wrong" the predictions are and is used to update the weights through backpropagation. During the training process, the weights and biases will be updated to minimize the difference between the output and the true labels. To accomplish this, a loss function and optimizer are required. The optimizer updates the model parameters sequentially over each hidden state for every timestep.



(a) Principle idea of recurrent neural network.



(b) Visualization of predictions with an RNN on dialysis data.

Figure 5: Visualization of theory and application of a RNN in the prediction of outputs L_t, with inputs X_t and hidden state h_t at time t. V, U and W are parameters of the network that are the same for each timestep. Abbreviations: DBP, diastolic blood pressure; SBP, systolic blood pressure; RNN, recurrent neural network.

Methods

Data collection

Data was obtained from all HD patients of the LUMC with an age of ≥ 18 years. The protocol (ProjectID 132628, acronym "Intradialytische Hypotensie") for data collection and analysis was approved by the Institutional Review Board of Internal Medicine of the LUMC. Demographics and dialysis data were collected from Hix, the standard electronic health record (EHR) in the LUMC, and Diamant, the dialysis-specific electronic record. In Diamant, all data as recorded by the dialysis machines as well as nursing interventions and notes are stored. The raw data was collected from the data platform of the LUMC, which could be accessed using Structured Query Language (SQL) Management Studio. Several queries were written in SQL, a language for retrieving data from this type of data platforms, to extract the desired variables [45]. An example of a query to obtain dialysis summary data is provided in Appendix A. Simultaneously, a formal database was designed in the data management platform Castor to meet the requirements for retrospective data analysis protocols of the LUMC [46].

The demographics collected from HiX were gender, birthyear, length, all diagnoses and medication. From Diamant, the following dialysis summary data was obtained: pre-dialytic weight, blood pressure and pulse, the target weight and the date of the treatment. The intradialytic machine-stored data included: blood pressure, arterial pressure, venous pressure, bicarbonate conductivity, blood volume change, blood flow rate, conductivity, dialysate flow rate, ionic mass balance, Kt/V, pulse, temperature, ultrafiltration rate and ultrafiltration volume. Kt/V is a measure of dialysis adequacy, and in this case, it was directly measured by the dialysis machine using sodium concentration as a surrogate for urea, unlike the usual calculation from blood urea concentration. While potentially clinically relevant, BCM measurements could not be directly extracted from the data platform and were consequently excluded from the analysis.

From the obtained lists of medication, patients with diabetes were identified by the prescription of metformines, insulin, sulfonylurea derivatives or dipeptidyl peptidase-4 inhibitors. Additionally, it was derived whether patients used antihypertensive medication or had a cardiovascular history, identified from prescribed anticoagulants. The medication for this selection is specified in Appendix C. Lastly, patients with an autoimmune disease were identified based on the registered diagnoses and three additional features were calculated from the obtained baseline features:

- age at the dialysis treatment
- body mass index (BMI) at the dialysis treatment
- the amount of overhydration in liters at the start of dialysis, defined as
  overhydration = pre-dialytic weight - target weight

It is expected that the pre-dialytic SBP will have a great contribution to the prediction of IDH, as this feature is used in clinical practice is used to determine the ultrafiltration volume. However, as we are mostly interested if the machine learning models are able to make a proper prediction in addition to the a priori expectation at the start of dialysis, the pre-dialytic SBP was removed as a feature in the analysis.

Labels

Complicating the assessment and prevention of IDH is the lack of consensus on a clear definition. Consequently, in research regarding this topic, multiple definitions are applied and compared. The most commonly used definitions are provided in Table 1. For the current project, two different definitions for IDH are assessed. The first definition, defined as **IDH1**, is a drop of ≥ 20 mmHg compared to the pre-dialytic SBP. The second definition, **IDH2**, is an absolute intradialytic SBP of ≤ 90 mmHg.

The main analysis of the current project concerned predictions of these definitions for IDH 30 minutes ahead by application of a logistic regression, random forest, XGBoost and RNN. Additionally, two subanalyses are performed. The first is the prediction of both definitions, but 60 minutes ahead. The second is analysis of prediction 30 minutes ahead, but with blood pressure measurements excluded from the analysis to illustrate the impact of removing these parameters.

Table 1: Different definitions of intradialytic hypotension.

National Kidney Foundation Kidney Disease Outcomes Quality Initiative Guidelines 2005 [47]	Decrease in SBP ≥20 mmHg or MAP ≥10 mmHg
European Best Practice Guidelines 2007 [48]	Decrease in SBP ≥20 mmHg in combination with associated clinical and nursing interventions
UK Renal Association Guidelines 2009 [49]	An acute symptomatic fall in blood pressure during dialysis requiring immediate intervention to prevent syncope
Absolute nadir SBP [25]	≤90 mmHg

Abbreviations: MAP, mean arterial pressure; SBP, systolic blood pressure.

Data preprocessing

All data preprocessing steps, as well as training the machine learning algorithms, were performed Spyder, an interative Python environment [50].

Baseline variables

After all data was collected, feature distributions of all baseline variables (the pre-dialytic variables) were plotted in order to detect outliers. After visual inspection of these distributions, values outside a manually set range or more than 5 times the standard deviation from the mean were removed from the data to exclude non physiological data. These settings are provided in Appendix C.

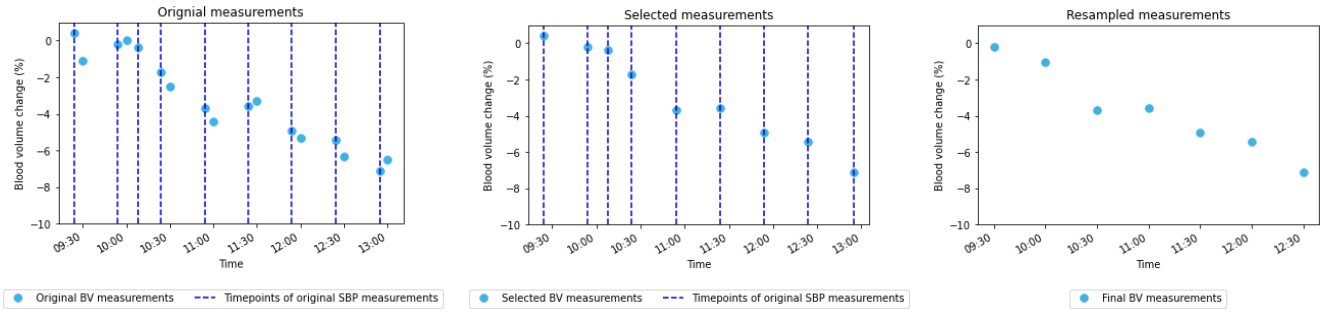
The data contained many missing target weight values due to administrative reasons. As not all machine learning models can handle missing data, these were filled by two different methods. Patients receiving HD without ultrafiltration were identified as having a total UF below 1L. For these patients, the target weight was set as the weight before dialysis, because these patients are generally considered not to be overhydrated due to preserved diuresis. For calculation of the remaining missing target weight values, the total UF of the treatment was subtracted from the weight before dialysis. Remaining missing values of pre-dialytic blood pressure, weight and target weight were interpolated by the previous and consecutive dialysis session of that same patient.

For both definitions of IDH, sessions in which IDH already occurred at the first measurement were excluded from the analysis. This scenario was considered to complicate the predictions later in the session.

Time-varying features

Sessions with a shorter duration than two hours or with less than five blood pressure measurements were removed from the data. Even though the intradialytic data was stored automatically, some outliers were present. Therefore, they were removed by application of the same method as described for the baseline variables. Deleted values were interpolated by linear interpolation of the previous and consecutive measurement within the same session. When analyzing the data, it appeared that the intradialytic machine measurements were stored with differences in resolution. To illustrate; one dialysis session of four hours could contain eight datapoints for blood pressure and 12 for arterial pressure. For generating input for the machine learning models, all datapoints were required to have the same timestamps. Therefore, the timepoints of the blood pressure measurements were selected for all features. These measurements were always conducted intervals of 30 minutes, with possibly some additional measurements on indication. If a feature had no measurement at exactly that time, the nearest datapoint was selected. Additional measurements were removed from the data. Consequently, one measurement per 30 minutes was selected to obtain the same resolution for every session. An example of selection of datapoints of blood volume measurements is visualized in Figure 6.

Training a logistic regression, random forest and XGBoost classifier requires an input of shape $(n_{samples}, n_{features})$. As the dataset consisted of n sessions, with n timepoints of n features, reconstruction was required. Therefore, one sample was set as one timepoint in a dialysis session. A session of four hours will therefore result in 8 samples. The features for a single sample include the baseline variables of that session and the dialysis measurements at that timepoint. However, the change in parameters over time contains highly relevant information. Therefore, new features that summarize the behaviour of the time dependent features up until the specific timepoint of the current sample were created. For every 30 minutes from the start of HD, indicated with index i for every timepoint, the value of the feature at the current timepoint X_{t_i} minus the previous value $X_{t_{i-1}}$, the slope of the change in the feature from the start of the session t_0



(a) Original BV measurements and timepoints of SBP measurements.

(b) Selected BV measurements based on SBP timepoints.

(c) Resampled BV measurements.

Figure 6: Methods for selecting datapoints in a dialysis session. Abbreviations: BV, blood volume; SBP, systolic blood pressure.

to t, and the standard deviation of all measurements between the start of the session t_0 and t were calculated by the following formulas

$$X_{t_i} - X_{t_{i-1}} \qquad X_{dt_i} = \frac{X_{t_i} - X_{t_0}}{t_i - t_0} \qquad \sigma = \sqrt{\frac{\sum (X_{t_i} - \overline{X})^2}{n}} \qquad \text{with } \bar{X} = \sum_{i=1}^{n} X_{t_i}$$

with X_{dt_i} the slope of the feature values from t_0 to t_i, X_{t_0} the value of the parameter at t_0, σ the standard deviation, $\bar{X}$ the mean of the measurements and n the number of measurements from t_0 to t_i. To ensure there are no unexpected relations between the features, a feature correlation plot was created and visually inspected.

Data preparation for the RNN was slightly different than for the other models, as a RNN layer expects an input of shape (N, L, H_{in}), where N is the batch size, L the sequence length and H_{in} the number of features. The labels that are used for evaluation of the predictions have shape $(N, L, 1)$, containing either a 0 or 1 for each element in the sequence. This allows the RNN to go over the whole sequence and make a new prediction for all consecutive timesteps. As a RNN expects sequences of the same length, the maximal time of the sequences is set to four hours ($L = 8$) and shorter sequences are extended with zeros.

The baseline variables were transformed to constant vectors and concatenated with the time-varying features. The final feature set for the RNN was smaller than for the other models, because the feature calculations that were conducted for the other models were not used as input since it was expected that the RNN has the ability to learn this by itself.

Model development

The Scikit-Learn library was used for development of a logistic regression, random forest and XGBoost classifier [51]. For training and evaluation of the models, 5-fold cross-validation was applied, in which the final database was split in a train and test set, with all sessions of the same patient in either the train or test set. The data was then scaled by a standard scalar and a randomized search was performed to find the optimal hyperparameters for each model. In a randomized search, several different combinations of hyperparameters are assessed and the combination with the best performance is returned. The search was conducted over a manually set range of the following hyperparameters

- logistic regression: regularization parameter
- random forest: number of trees, criterion, perform bootstrapping, maximal tree depth, minimal samples per leaf
- XGBoost: maximal tree depth, learning rate, minimal loss reduction, subsampling ratio of samples, subsampling ratio of columns, l2 regularization term

The final hyperparameters were chosen based on the results of the randomized search, while considering prevention of overfitting and computational duration. As the current dataset is inbalanced for both IDH1 and IDH2, a sample weight was provided when fitting the models on the training data. This weight was determined by the ratio of the number of negative labels over the positive labels. In the current dataset, this ratio was 5 for IDH1 and 50 for IDH2. Subsequently, the model with optimal hyperparameter settings was fit on the train data and used for prediction of IDH on the test data.

Development of the RNN was performed with use of the PyTorch library, which is specifically designed for deep learning applications [52]. A RNN requires the following parameters

- input size: the number of features
- hidden size: the number of neurons
- the number of recurrent layers

The current model consisted of one recurrent layer with 50 neurons. In order to obtain the desired output dimension, which is 1 for binary classification, a linear transformation was applied after the recurrent layer. The initial hidden state contained zeros. During model training, the parameters are adjusted by assessing the input data in batches. In the current training loop, the batch size was set to 300 for IDH1 and 500 for IDH2. The latter is relatively high, but was required due to the low incidence of the labels. Lastly, the learning rate was set to $1e-4$ for IDH1 and $1e-5$ for IDH2 and the model was trained for 300 epochs.The Python code containing the model architecture is provided in Appendix D.

After the model was defined, it was trained for a certain number of epochs. Every epoch, the complete train dataset is assessed in batches. A padding function was applied over the selected batch to ensure all sequences within that batch were of the same length. Padding batches is computationally more efficient than padding the entire dataset from which batches are selected. Additionally, it influences the imbalance between the labels to a lesser extent. The model generates an output for every timestep in each sequence. At the start of the training, these output labels will mostly be "guessed" and will deviate from the true labels. For the current model, this deviation, the loss, was calculated as the binary cross entropy. Within the loss function, the same weights as for the other models were provided for the positive labels. The model was optimized with a stochastic gradient descent algorithm. Additionally, a linear scheduler was applied to the optimizer, in order to decrease the learning rate over a manually set amount of epochs. This allows faster convergence at the start of the training, and more specific convergence to the optimum later on.

Performance evaluation

For evaluation of the models, the ROC curves and corresponding AUCs were plotted and calculated. Additionally, the feature importance scores were obtained for the logistic regression, random forest and XGBoost and the 20 features with highest absolute scores were provided. To assess whether the RNN model was effectively learning from the data while overfitting was being prevented, the learning curve for the train and test data was plotted. This plot provides the change in loss across the training epochs. A proper learning curve should stabilize towards the end of the training process, indicating the loss is no longer significantly decreasing as it learns from the data, with a minimal gap between the training and test curves.

The precision, recall and specificity at the optimal probability threshold are calculated for for the final RNN and the model with the highest AUC. Precision is the ratio of correctly classified positive labels over all labels that are classified as positive. Recall is the ratio of correctly classified positive labels over the true number of positive labels and specificity the ratio of correctly classified negative labels over the true number of negative labels. These metrics are calculated as

$$precision = \frac{TP}{TP+FP} \qquad recall = \frac{TP}{TP+FN} \qquad specificity = \frac{TN}{TN+FP}$$

with TP the number of true positives, FP the false positives, FN the false negatives and TN the true negatives.

The choice of the prediction threshold for calculation of these metrics should be based on the clinical context. In the case of a low true positive rate in real-time intradialytic IDH predictions, IDH incidence would remain high due to belated intervention. For a high false positive rate (FPR), the ultrafiltration rate would unjustly be turned down, resulting in insufficient ultrafiltration. As it is undesirable that the feature predictions would lead to more patients not reaching their target weight, the FPR should be limited. In the paper of Zhang et al., which had similar objectives, interviews with nephrologists and dialysis staff led to the consensus that a FPR of 0.1 while maintaining a recall above 65% would be acceptable [53]. Such a threshold would limit the false alarm rates. To visualize the trade-off between these outcome, the precision, recall and specificity of the RNN were calculated for multiple prediction thresholds.

Lastly, several sessions were selected randomly in order to provide their predicted probabilities and corresponding labels based on the optimal threshold as described above. The data from these sessions was visually inspected to obtain insight in common mistakes made by the classifiers. This analysis was conducted from predictions made by the RNN and the classifier with the highest AUC. In addition to the predictions, the SBP, diastolic blood pressure (DBP), pulse and change in blood volume were plotted. These are the some of the most important parameters that are currently used for UF guidance in clinical practice. Additionally, the ultrafiltration rate over the complete session was provided. Analyzing these results and parameters may offer valuable insights in variations in predictions among the classifiers and the prevalent errors that arise. For illustrative purposes, these plots were provided for two sessions of each definition.

Results

Data collection and session characteristics

A total number of 457 patients with dialysis data from 2010 to June 2023 was identified. This comprised 64,437 dialysis sessions. After exclusion of patients with missing baseline variables, sessions with less than five intradialytic blood pressure measurements, sessions with longer duration than five hours and sessions with IDH at the first measurement, the remaining numbers of patients and sessions were

- IDH1: 436 patients and 37,025 sessions
- IDH2: 441 patients and 43,722 sessions

The discrepancy in these numbers was due to a higher number of sessions with IDH1 than IDH2 at the first measurement. In 24% of total sessions, a pre-dialytic pulse measurement was missing. Therefore, this baseline feature was excluded from the data. An overview of the final features is provided in Table 2.

6838 sessions without ultrafiltration were identified. For these sessions, the target weight was set as the weight before dialysis. Of the remaining 2677 missing target weight values, 1669 could be set by subtracting the the total ultrafiltration volume from the weight before treatment. The remaining missing target weight values, as well as the pre-dialysis weight and blood pressure were interpolated. The database that was used as input for prediction with the RNN contained a total of 27 features, to which 47 newly calculated features were added for the logistic regression, random forest and XGBoost models.

Due to data upload restrictions regarding the Castor database, not all available dialysis data could be entered in this database. Therefore, the data was obtained from the data platform directly. In the future, the Castor database could be completed and extended with additional parameters such as blood test results. An overview of the architecture of the Castor database is to be found in Appendix B.

Table 2: List of features.

Baseline variables	Machine data
Obtained	
Gender	Blood pressure
Length	Pulse
Pre-dialysis weight	Change in blood volume
Diabetes	Arterial pressure
Auto-immune disease	Venous pressure
Use of antihypertensives	Conductivity
	Ultrafiltration volume
	Ultrafiltration rate
	Kt/V
	Blood flow rate
	Dialysate temperature
	Dialysate flow rate
	Transmembrane pressure
	Bicarbonate conductivity
Derived	
Age at time of treatment	Difference with previous
BMI	Slope
Cardiovascular history	Standard deviation
Overhydration	

Kt/V is a measure of the dialysis adequacy.
Abbreviations: BMI, body mass index.

The baseline characteristics, calculated over the total number of sessions included per analysis, are provided in Table 3. Categorical variables are presented as the number and percentage of sessions and continuous variable as the mean and standard deviation. No unexpected feature correlations appeared in the feature correlation plot, which is provided in Appendix C. In the final databases used for analysis, 359 patients (82%) experienced one or more episodes of IDH1 and 172 patients (39%) of IDH2. IDH1 occurred in 57% and IDH2 in 10% of the sessions. This incidence was calculated before removal of sessions with IDH at the first measurement.

Table 3: Baseline characteristics.

	IDH1 (n=37025)		**IDH2** (n=43722)	
Characteristics	**Mean**	**± SD**	**Mean**	**± SD**
Age [years]	67.3	14.4	67.8	14.3
Length [cm]	171.2	10.2	171.1	9.9
BMI [kg/m^2]	26.0	5.2	26.0	5.1
Pre-dialytic SBP [mmHg]	136.8	24.6	141.9	24.1
Pre-dialytic DBP [mmHg]	74.2	14.5	75.4	14.3
Overhydration [liters]	+1.3	1.2	+1.3	1.2
	N	**%**	**N**	**%**
Male	23678	64.0	27624	63.2
Medical history				
Diabetes	7597	20.5	9023	20.6
Auto-immune disease	859	2.3	1076	2.5
Cardiovascular history	13110	35.4	16001	36.6
Antihypertensive medication	20065	54.2	24577	56.2

Calculated over the number of sessions per definition.
Abbreviations: BMI, body mass index; DBP, diastolic blood pressure; SBP, systolic blood pressure; SD, standard deviation.
IDH1 is intradialytic hypotension defined as a drop in systolic blood pressure of ≥20 mmHg compared to baseline, and IDH2 a absolute systolic blood pressure ≤90 mmHg.

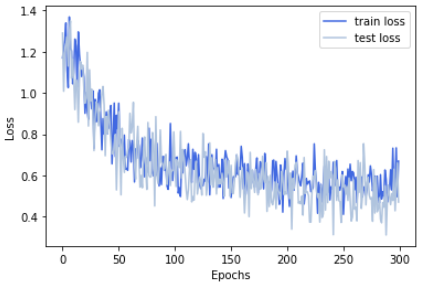


Figure 7: Loss curve of the recurrent neural network for prediction of IDH2.

Performance evaluation

An example of the train and test distribution in the different folds is provided in Appendix E to validate that a one patient can only occur in the train or test split within one fold. The hyperparameter options, results of the randomized search, and choice for the final settings are also provided in Appendix E. For the different definitions of IDH, different hyperparameter settings were applied in the final analysis. Since the outcome of the hyperparameter search was fairly consistent for the subanalyses, the same settings were used as for the main analysis. For illustrative purposes, the loss curve of training the RNN for the prediction of IDH2 is provided in 7.

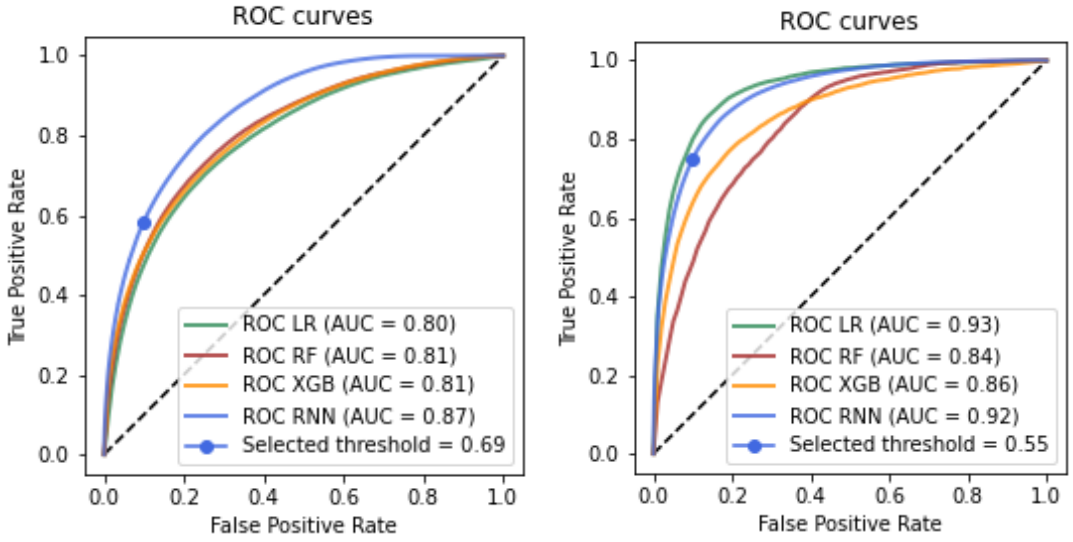
The ROC curves of all models are provided in Figure 8. The AUCs for the logistic regression, random forest, XGBoost and RNN were 0.80, 0.81, 0.81 and 0.87 respectively for the prediction of IDH1 and 0.93, 0.84, 0.86 and 0.92 for IDH2. The precision, recall and specificity for predictions with the RNN at different thresholds are provided in Table 4. The performance metrics of the RNN were compared to those of the logistic regression. This classifier was chosen for comparison, because it outperformed all other models in the prediction of IDH2. The threshold for classification with the RNN at a maximal FPR of 0.1 was 0.69 for IDH1 and 0.55 for IDH2. For the logistic regression this was 0.60 for IDH1 and 0.66 for IDH2. The precision, recall and specificity for the RNN with application of this threshold were 0.56, 0.59 and 0.90 for IDH1 respectively, and 0.10, 0.76 and 0.90 for IDH2. The precision, recall and specificity for the logistic regression were 0.63, 0.50 and 0.90 for IDH1 and 0.14, 0.80 and 0.90 for IDH2.

The logistic regression, random forest and XGBoost rendered fairly similar feature importance rankings. In all three models, the SBP slope over time was the feature with the greatest importance score in IDH1 and the SBP measurement at the timepoint of prediction was the highest form IDH2. The 20 most important features for classification are provided in Figure 9 and 10.

Subanalyses

Subanalysis of 60 minutes ahead instead of 30 minutes resulted in AUC values of 0.77, 0.78, 0.78, 0.86 for the logistic regression, random forest, XGBoost and RNN respectively for IDH1 and 0.89, 0.77, 0.79 and 0.91 for IDH2. When blood pressure measurements were excluded from the analysis, the AUCs dropped to 0.62, 0.62, 0.60 and 0.79 for IDH1 and 0.71, 0.57, 0.59 and 0.79 for IDH2. These ROC curves are provided in Appendix F.

Results of predictions on individual sessions were visually inspected for the RNN and logistic regression. Despite the fact that the logistic regression did not outperform the random forest and XGBoost for IDH1, this was still the classifier of choice for this analysis due to its high AUC. At this inspection, it appeared that for the logistic regression the predicted



(a) IDH1 (b) IDH2

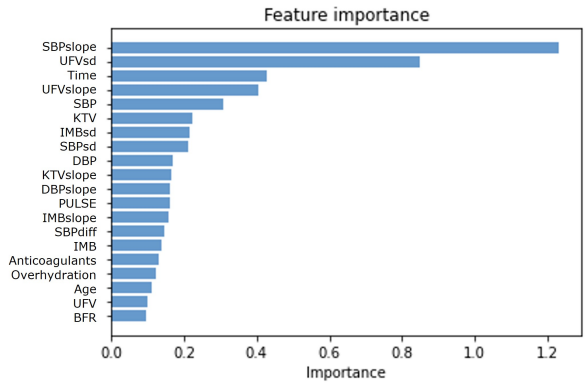
Figure 8: ROC curves of the logistic regressoin (LR), random forest (RF), XGBoost (XGB) and recurrent neural network (RNN) for prediction of IDH 30 minutes ahead. IDH1 is intradialytic hypotension defined as a drop in systolic blood pressure of $\geq$20 mmHg compared to baseline, and IDH2 a absolute systolic blood pressure $\leq$90 mmHg.

probability was highly associated with the SBP measurements. While this visual association was also apparent for the RNN, it was of a lesser extent. Other findings were that in many sessions, the RNN gave a relatively high number of false positives for IDH2 and that a positive prediction often occurred one timestep too late, regardless of the definition or classifier. Figure 11 shows the predictions of both classifiers, measurements of the intradialytic blood pressure, pulse, blood volume change and ultrafiltration rate for a session with IDH1 in which the RNN outperforms the logistic regression. Figure 12 shows an example of a session where both classifier are one timestep too late in their positive prediction of IDH1. Figure 13 shows a session where the logistic regression outperforms the RNN in the prediction of IDH2 and Figure 14 a session where the RNN provides false positive predictions based on the SBP. It is important to emphasize that this analysis was conducted randomly and that the representativeness of the findings could not be confirmed.

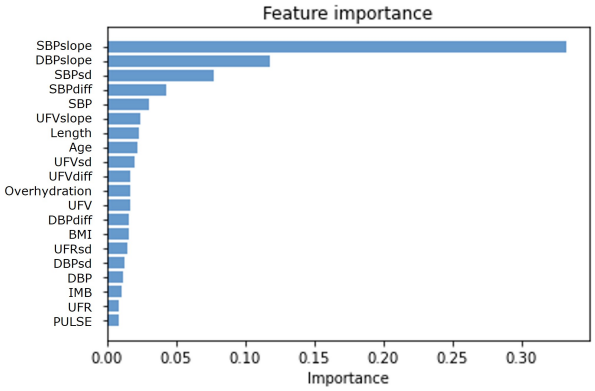
Table 4: Performance evaluation results of the RNN for different probability thresholds.

	IDH1			IDH2		
Threshold	**Precision**	**Recall**	**Specificity**	**Precision**	**Recall**	**Specificity**
0.0	0.18	1.0	0.0	0.02	1.0	0.0
0.1	0.24	1.0	0.31	0.02	0.99	0.38
0.2	0.27	0.98	0.31	0.03	0.87	0.54
0.3	0.30	0.95	0.51	0.04	0.94	0.66
0.4	0.35	0.90	0.62	0.06	0.90	0.77
0.5	0.41	0.82	0.73	0.09	0.80	0.87
0.6	0.48	0.70	0.83	0.13	0.69	0.92
0.7	0.57	0.58	0.90	0.17	0.58	0.95
0.8	0.67	0.43	0.95	0.24	0.46	0.98
0.9	0.79	0.24	0.99	0.39	0.28	0.99

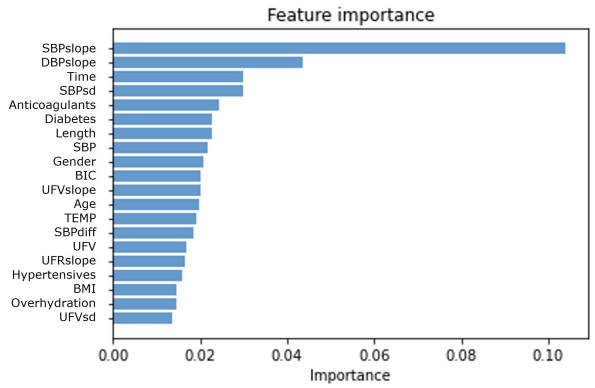
IDH1 is intradialytic hypotension defined as a drop in systolic blood pressure of $\geq$20 mmHg compared to baseline, and IDH2 a absolute systolic blood pressure $\leq$90 mmHg.



(a) Logistic regression

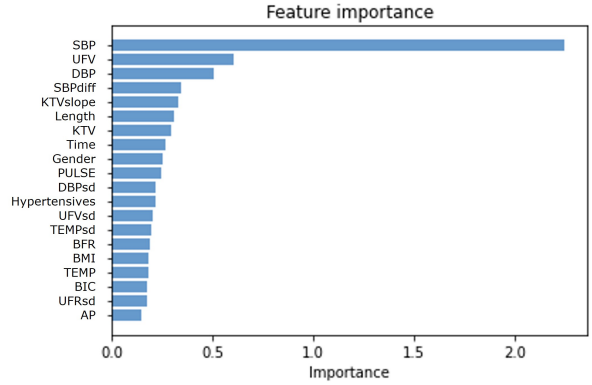


(b) Random forest

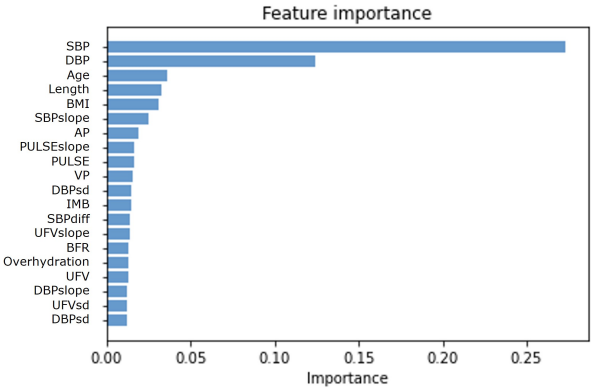


(c) XGBoost

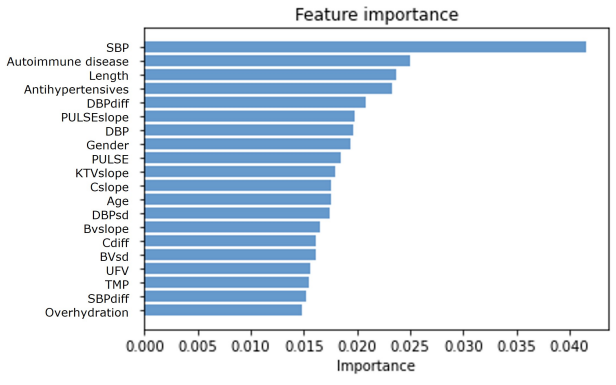
Figure 9: Mean feature importance scores of 5 fold cross-validation of logistic regression, random forest and XGBoost for prediction of IDH1 30 minutes ahead. The feature abbreviations are: BIC, bicarbonate conductivity; BFR, blood flow rate; BMI, body mass index; DBP, diastolic blood pressure; IMB, ionic mass balance; KTV, is a surrogate of Kt/V (a measure of the dialysis adequacy); SBP, systolic blood pressure; UFV, ultrafiltration volume. *Diff* indicates the difference of the feature value with the previous measurement, *sd* and *slope* the calculated standard deviation and slope of the feature over the sequence.



(a) Logistic regression

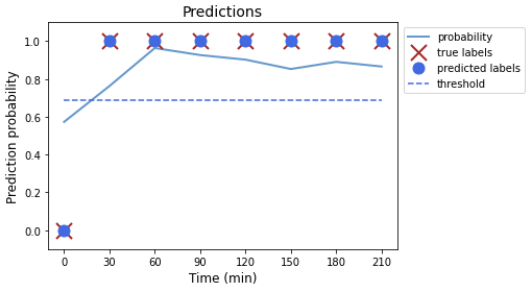


(b) Random forest

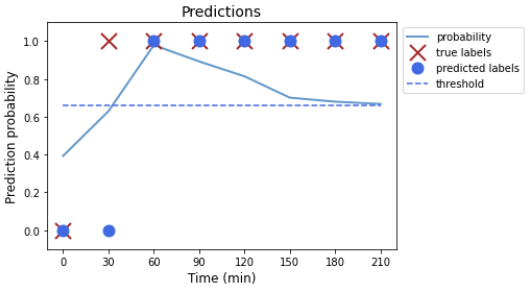


(c) XGBoost

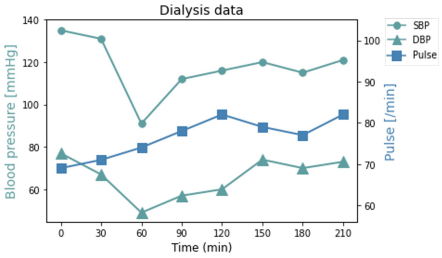
Figure 10: Mean feature importance scores of 5 fold cross-validation of logistic regression, random forest and XGBoost for prediction of IDH2 30 minutes ahead. The feature abbreviations are: AP, arterial pressure; BIC, bicarbonate conductivity; BFR, blood flow rate; BMI, body mass index; C, dialysate conductivity; DBP, diastolic blood pressure; IMB, ionic mass balance; KTV, is a surrogate of Kt/V (a measure of the dialysis adequacy); SBP, systolic blood pressure; UFR, ultrafiltration rate; UFV, ultrafiltration volume; VP, venous pressure. *Diff* indicates the difference of the feature value with the previous measurement, *sd* and *slope* the calculated standard deviation and slope of the feature over the sequence.



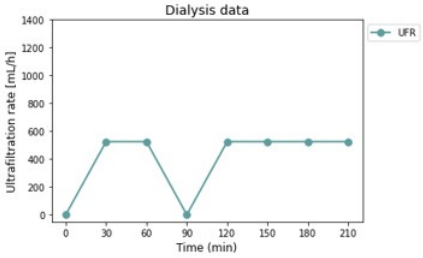
(a) Recurrent neural network.



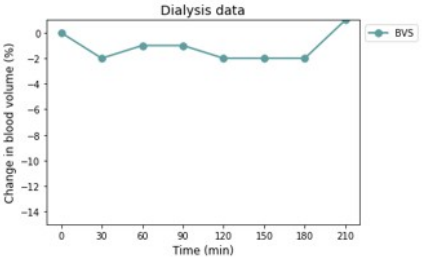
(b) Logistic regression.



(c) Blood pressure and pulse.

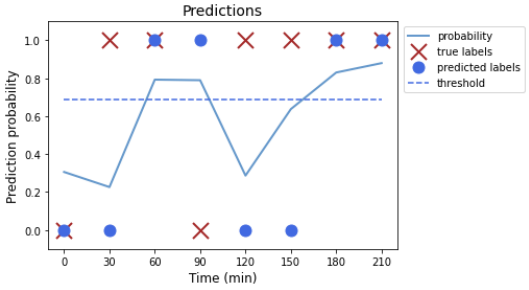


(d) Change in blood volume.

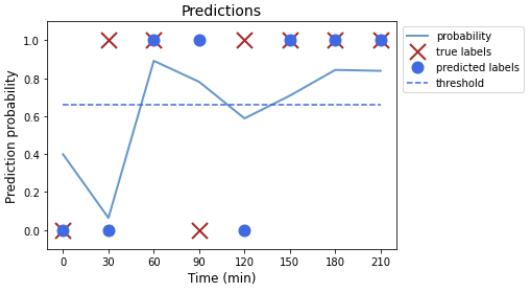


(e) Ultrafiltration rate.

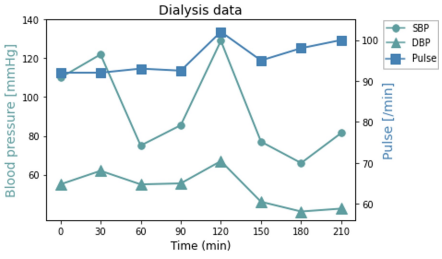
Figure 11: Predictions and dialysis data of IDH1 in example session 1.
The figures provide the probabilities, predicted labels and true labels of the RNN (a) and logistic regression (b), the blood pressure and pulse measurements (c), blood volume sensoring data (d) and ultrafiltration rate (e) over the dialysis session. IDH1 is intradialytic hypotension defined as a drop in systolic blood pressure of $\geq$20 mmHg compared to baseline. Abbreviations: BVS, blood volume sensoring; DBP, diastolic blood pressure; SBP, systolic blood pressure; UFR, ultrafiltration rate.



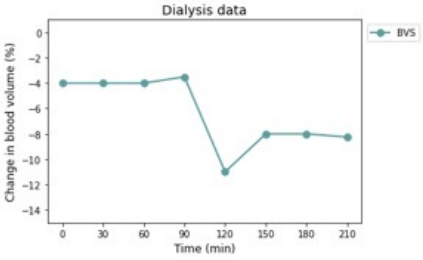
(a) Recurrent neural network.



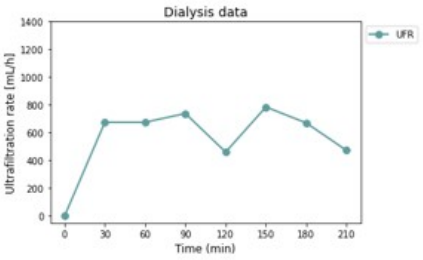
(b) Logistic regression.



(c) Blood pressure and pulse.

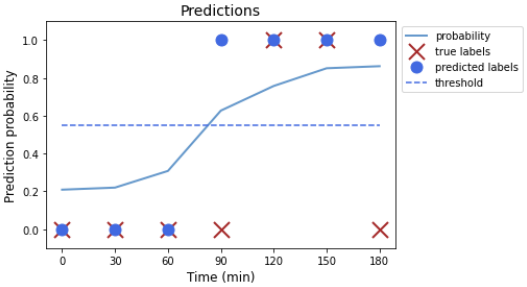


(d) Change in blood volume.

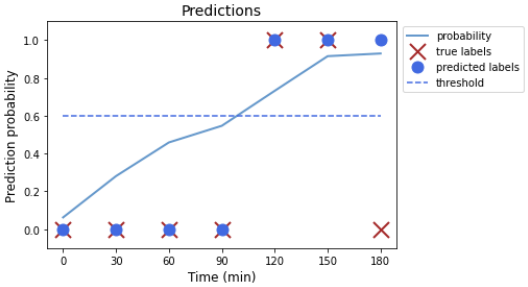


(e) Ultrafiltration rate.

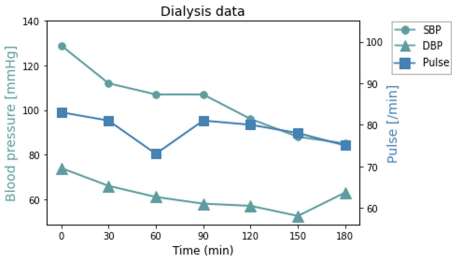
Figure 12: Predictions and dialysis data of IDH1 example session 2.
The figures provide the probabilities, predicted labels and true labels of the RNN (a) and logistic regression (b), the blood pressure and pulse measurements (c), blood volume sensoring data (d) and ultrafiltration rate (e) over the dialysis session. IDH1 is intradialytic hypotension defined as a drop in systolic blood pressure of $\geq$20 mmHg compared to baseline. Abbreviations: BVS, blood volume sensoring; DBP, diastolic blood pressure; SBP, systolic blood pressure; UFR, ultrafiltration rate.



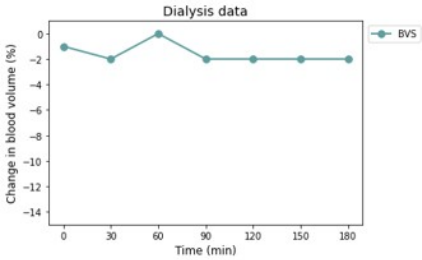
(a) Recurrent neural network



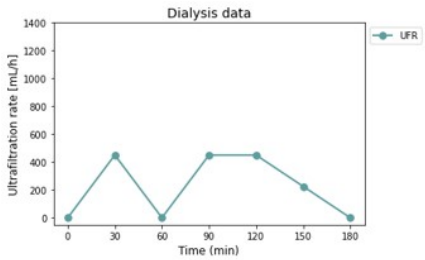
(b) Logistic regression.



(c) Blood pressure and pulse.

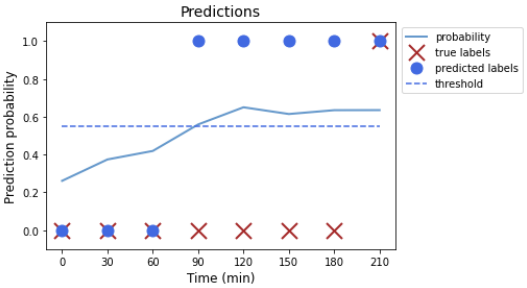


(d) Change in blood volume.

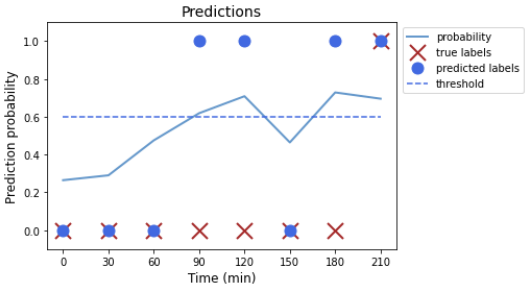


(e) Ultrafiltration rate.

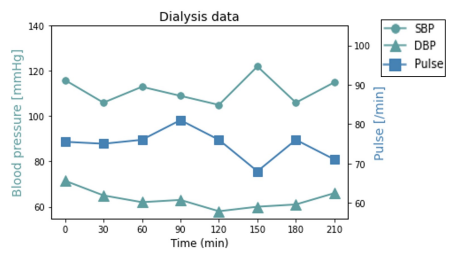
Figure 13: Predictions and dialysis data of IDH2 example session 3.
The figures provide the probabilities, predicted labels and true labels of the RNN (a) and logistic regression (b), the blood pressure and pulse measurements (c), blood volume sensoring data (d) and ultrafiltration rate (e) over the dialysis session. IDH2 is defined as a absolute systolic blood pressure $\leq$90 mmHg. Abbreviations: BVS, blood volume sensoring; DBP, diastolic blood pressure; SBP, systolic blood pressure; UFR, ultrafiltration rate.



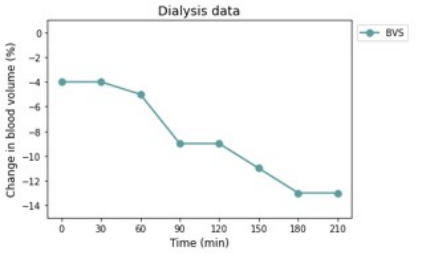
(a) Recurrent neural network.



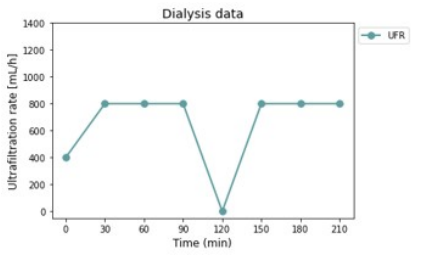
(b) Logistic regression.



(c) Blood pressure and pulse.



(d) Change in blood volume.



(e) Ultrafiltration rate.

Figure 14: Predictions and dialysis data of IDH2 example session 4.
The figures provide the probabilities, predicted labels and true labels of the RNN (a) and logistic regression (b), the blood pressure and pulse measurements (c), blood volume sensoring data (d) and ultrafiltration rate (e) over the dialysis session. IDH2 is defined as a absolute systolic blood pressure $\leq$90 mmHg. Abbreviations: BVS, blood volume sensoring; DBP, diastolic blood pressure; SBP, systolic blood pressure; UFR, ultrafiltration rate.

Discussion

In this project, both intradialytic hemodialysis measurements and baseline variables were used for development of machine learning algorithms that were capable of making predictions of IDH at 30 minute intervals. This study aimed to assess the potential of providing real-time warnings that would enable timely intervention and prevention of IDH. The results show that the predictions made by the current models are greatly based on the intradialytic SBP measurements and that only the RNN was capable of learning from other variables. However, evaluation of the model performances revealed that the predictions based on the the current database are not sufficient for clinical implementation.

Most of the findings of this thesis are in accordance with results from previous studies. Two recent studies showed that neural networks outperformed other machine learning models in the prediction of various definitions of IDH [36, 37]. Similar as in the current results, the AUCs of deep neural networks were higher for the prediction of IDH as an absolute value of $\leq$90 mmHg (0.94 and 0.91) than for a drop of $\geq$20 mmHg (0.87 and 0.87) in SBP. However, the Area Under the Precision-Recall Curves (AUPRC) in these studies differed for the definitions with AUPRCs of 0.62 and 0.29 for $\leq$90 mmHg and 0.78 and 0.80 for $\geq$20 mmHg. The study from the authors Lee et al., with the higher AUPCR for $\leq$90 mmHg, had availability over an extensive database of 261.647 HD sessions and many features [36]. Their feature ranking outcomes showed that in addition to SBP measurements and time to start of dialysis, body temperature and blood test results had high contribution to the predictions. It might therefore be feasible that additional features and sessions result in improved model performance. Similar to the current result, there was a higher trade-off between precision and recall for IDH as an absolute value ≤ 90 mmHg. The authors contributed these results to the low incidence of this definition in their data. In regard to the comparison of the performance between the models, the RNN of Lee et al. slightly outperformed the other models both for IDH as an absolute value ≤ 90 mmHg, but not for the other definitions that were applied. The study from Kim et al. showed similar results. The logistic regression and RNN had equal performance when evaluated internally. However, external validation provided superior performance of the RNN over the logistic regression. Additionally, for all definitions of IDH, the random forest and XGBoost had lower AUCs than the logistic regression and deep neural network.

The current study showed similar results when compared to the two studies described above, despite the fact that a smaller database with a limited number of features is applied. This suggests that the models have demonstrated strong performance on the present database and that there is potential for achieving even better results by data expansion involving additional sessions and features.

Despite the promising results, some limitations regarding the current methodology need to be considered. While most of the included features are stored by the dialysis machine automatically, some of the variables are filled in manually by dialysis staff. Consequently, there is a relatively high amount of missing data. While some variables could be excluded from the data without major consequences, others were considered to be of high clinical relevance and were preserved. One of these variables was the target weight, which is used for calculation of the estimated amount of overhydration at the start of dialysis. Several approaches have been applied to reduce the number of missing values and preserve this feature in the analysis. One of the methods was interpolation of these missing values by the previous and consecutive session. Although the period between these session mostly comprised only one to three days, the duration of these periods was not taken into account in the current methodology. Another method included subtraction of the total UFV from the pre-dialytic weight. However, when a patient is considered to be excessively overhydrated, the target weight is deliberately dropped and the target of the treatment is based on maximal volume extraction rather than on the weight. Therefore, in some cases this method might result in underestimated overhydration values. Additionally, not all baseline features could be collected directly, but required derivation from other data, such as the registered diagnoses and prescribed medication. The validity could therefore not be confirmed.

Another limitation in regard to the features, is that the spatial resolution of some intradialytic measured features, such as the change in blood volume, was reduced. Some features are initially measured and stored at a higher frequency or at different timepoints than the blood pressure data. Selecting the value at the timepoint closest to that of the blood pressure might have resulted in a shift in the data. In future research, this shift could be resolved by applying interpolation techniques. A disadvantage of this approach is that it is computationally expensive and therefore time consuming, while the shift is considered to be minimal.

The hyperparameter settings for the logistic regression, random forest and XGBoost were determined by assessment of results from a randomized search over a pre-determined range of settings, since this is computationally less expensive than

a grid search. Consequently, only a limited number of combinations is evaluated. To confirm the optimal hyperparameter settings, a grid search should be performed to evaluate all parameter settings.

In the current structure of the applied RNN, all features are concatenated and fed into the recurrent layer. Consequently, time-invariant features, such as the weight at the dialysis session, were transformed into vectors with a constant value for every timestep. However, as recurrent layers are specifically designed for time-varying data, this approach does not align with their intended usage and may lead to higher losses during model training. Therefore, the architecture of the model could be improved by feeding only the time-varying features to the recurrent layer and the time-invariant features to a feed forward layer and then combine those in a new layer.

In addition the the methodological limitations, there are some matters regarding interpretability of the current result that require consideration. Firstly, an explanation for the high AUCs obtained in this project, is the absence of an external validation database. Possibly, the models are overfitting on the current database, which consists of data from one single center. Both dialysis practices and average patient characteristics might differ for other centers. When the performance of the current models would be evaluated on a database containing patients from other hospitals, the AUCs could be lower. Nonetheless, external validation is essential to validate the models' generalizability. Secondly, no feature importance is currently provided for the predictions made by the RNN. Gaining insight in features with high contribution to the predictions by application of readily available tools such as SHapley Additive exPlanation (SHAP) tools, could elucidate differences between the RNN and the other models and contribute in comparison of the models [54]. A SHAP value represents the difference between the prediction with the feature included in the model and the prediction with the feature set to a certain baseline value, such as the mean of the feature. Another approach for obtaining a explainable neural network that reveals which features are used for the predictions, is application of Gradient-weighted Class Activation Mapping (grad-CAM) [55, 56]. GradCAM calculates the gradient of an output with respect to the input to derive the contribution of the features for all timesteps. It generally provides heatmaps where colors indicate the contribution of a feature to the activation of the output.

Another matter of consideration is that a limited amount of features was applied in the current analysis. It is possible that parameters with a predictive value are now missed. There are many other factors that could lead to fluctuations in intradialytic blood pressure. Factors that have shown to be of influence on the blood pressure measured during HD are eating during the treatment, as more blood is directed towards the intestinal tract, and intradialytic exercise, which is also offered at the dialysis department in the LUMC [24, 57]. These aspects are not considered in the current analysis.

Lastly, a factor that is possibly of influence on the current predictions, is that some of the features contain information of an intervention on IDH, rather than the occurrence of IDH. A decrease in the UFR for example, is usually an intervention to a rapid decrease in intradialytic blood pressure or patient symptoms such as muscle cramps. Additionally, the dialysate temperature is mostly set lower for patients that are prone to developing IDH. In some of the models, these features are ranked highly and do have a significant contribution to the predictions. On the one hand, this is undesirable as it would be more relevant to assess the risk of IDH, regardless of these interventions, but on the other hand, for the risk assessment and prevention of IDH, there is always an interaction between objective measurements and their subjective interpretation. As subjective risk assessment will always be part of clinical care, incorporation of these interventions in the prediction matches the clinical context. In this respect, the true value of the prediction models, is in providing a correct positive predictions before it is noticed by and intervened on by dialysis staff. Consequently, a limitation of this project is that the distinction between predictions before and after a decrease in UFR has not been assessed.

One of the greatest drawbacks of machine learning in general is the lack of interpretability of the algorithms on which the predictions are based. Especially deep learning networks, such as the RNN, are considered to be a "black box" in which the input variables disappear. In some cases, this complicates direct clinical translation and interpretability. However, this could be compensated by a gain in prediction performance when compared to traditional statistical prediction models. Furthermore, although establishing a causal relationship is not possible, tools like grad-CAM do facilitate the identification of features that significantly contribute to predictions.

Given the outcomes of the current project, several steps could be considered for future development of the models. In the current analysis, only a subset of parameters that is available from maintenance HD patients has been included. There is a high additional available amount of data that could be added to the current models. In current clinical HD care, routine blood tests are performed monthly, containing results for sodium, potassium, hemoglobin and many others. While the individual measurements might not be of major influence, calculations on the change of these parameters over time could enhance performance of the models in the prediction of IDH. These blood test result are readily available in the data platform of the LUMC and should therefore be included in future analysis. Other possibly relevant parameters, such as the dialyzer type, dialysate prescription and type of vascular access could not be retrieved from the data platform. Since this information should be stored in the EHR Diamant, future attempts could be made to collect this data and add

it to the list of time-invariant features. The same applies for BCM measurements that are performed regularly for most patients. The results of the measurements are not stored systematically, but are reported textually in the dialysis record. Additional features that could be derived directly from the current database are the occurrence of IDH in previous sessions of the same patient and the dialysis vintage. For example, a new feature could be generated that indicates the occurrence of IDH in the previous 10 dialysis sessions. From a clinical perspective this is a feature with potential predictive value, which is confirmed in literature on similar machine learning models [53]. Additionally, some relevant parameters in the assessment of fluid status that could be considered for implementation in future analysis are the patients nutritional status and cardiovascular parameters. These variables have influence on the refilling capacity and thus the vascular response to hypotension.

Furthermore, in the current project, only two definitions of IDH were assessed. There are, however, many other definitions of clinical relevance. Instead of defining IDH as a drop of $\geq$ 20 mmHg compared to the pre-dialytic SBP, a drop of $\geq$20 mmHg within a certain amount of time, e.g. 15 minutes, could be clinically more relevant. For excessively overhydrated patients, an asymptomatic drop of $\geq$20 mmHg compared to baseline may be a desirable outcome. This would explain the high incidence of this definition in the current database. Alternatively, a percentual drop in SBP or inclusion of hypotensive symptoms or the MAP should be considered as alternatives for definition of IDH.

An approach that could be assessed for obtaining accurate predictions for IDH as $\leq$90 mmHg while averting a label inbalance, would be to redefine the target population. By only including HD sessions of patients with a high risk for development of IDH, for example by inclusion of sessions with below a certain threshold for the pre-dialytic blood pressure, the performance on this specific population could be evaluated.

Prior to the initiation of implementation of machine learning based prediction models, it should be ensured that the methods meet the requirements as defined in the Prediction model Risk Of Bias ASsessment Tool (PROBAST) [58]. This tool was developed for evaluation of prediction model studies, but could also serve as guideline for validation of all methodological steps in the model development. Currently, a PROBAST specifically designed for machine learning prediction studies is in development [59]. When published, this could serve as a valuable tool for validation of the methodology.

At interpretation of the data, it appeared that the features with the highest importance scores, and thus a great contribution to the prediction, were in line with expectations. For IDH1, the change in blood pressure during dialysis mostly contributed to the predictions, while change in UFV and the time since start of dialysis were also ranked highly. Unsurprisingly, the last SBP measurement had the highest importance score for the prediction of IDH2. The results indicate that the predictions are greatly based on the intradialytic blood pressure measurements. In current clinical practice, these measurements also guide the nursing staff in fluid management. Despite the fact that no feature importance scores were calculated for the RNN, the AUCs showed a smaller decline in value after removal of the blood pressure measurements compared to the other models. Furthermore, the performance of the models only slightly decreased when the prediction was done 60 minutes ahead, instead of 30, but witch a larger drop in AUC for the logistic regression, random forest and XGBoost than for the RNN. These results might suggest that the RNN has to ability to make a prediction based on other features than the blood pressure. It would therefore be interesting to assess which features contribute the predictions.

A possible explanation for the better performance of the logistic regression compared to the random forest and XGBoost could be found in the characteristics of the classifiers. The logistic regression might be less prone to overfitting due to its robustness compared to more complex ensemble methods like random forests and XGBoost, which require more data to generalize well.

Even though the AUCs are relatively high for prediction models in clinical context, inspection of predictions on some individual sessions revealed relevant shortcomings. In many of the sessions, the prediction of both definitions appeared to be one timestep too late. While the subsequent predictions might be correct and the performance results therefore high, this is clinically not acceptable. As one timestep is equivalent to 30 minutes, nursing staff would not have a sufficient time for interventions to prevent IDH. Even though the ROC outcomes for both definitions are high, it does not necessarily imply that the models could be readily applied in clinical setting. Evaluation of the performance metrics is complicated by the label distribution in the datasets, especially for IDH2. While the AUC of the RNN for IDH2 is extremely high, the precision at the threshold that is set to limit the FPR, is low. The cause for this counter-intuitive difference, is the high number of true negative values resulting from the inbalance in the labels in the dataset. This definition might therefore not be suitable for the development of machine learning models. Although the model performances for IDH1 are not yet sufficient, further improvements on data collection, data preprocessing and model development could potentially lead to their clinical applicability in the future. However, it is crucial to address the timing issue in predictions to ensure that interventions for preventing IDH are timely and effective.

In conclusion, the key findings of the current project can be summarized as follows. The machine learning algorithms have shown their ability to learn from both baseline variables and intradialytic measurements to make a prediction of IDH at different intradialytic timepoints. The current predictions are greatly based on the intradialytic blood pressure measurements. However, it is important to note that for IDH defined as a drop of $\geq$20 mmHg compared to baseline, the current models do not meet the necessary criteria for clinical implementation due to their limited performance metrics. In contrast, when IDH is defined as an absolute SBP $\leq$90 mmHg, these performance metrics were in a clinically acceptable range. Nonetheless, a significant challenge arises from the high label imbalance, resulting in very low precision outcomes. Consequently, the feasibility of achieving machine learning based real-time predictions for this definition remains uncertain. On the other hand, analysis of prediction 60 minutes ahead and after exclusion of blood pressure measurements, showed superior performance of the RNN over the other models. This suggest that the RNN has the ability to detect patterns that do not rely on the blood pressure. Therefore, further improvement of the RNN and expansion of the database by additional features and sessions holds significant promise for future research aimed at providing real-time predictions of IDH.

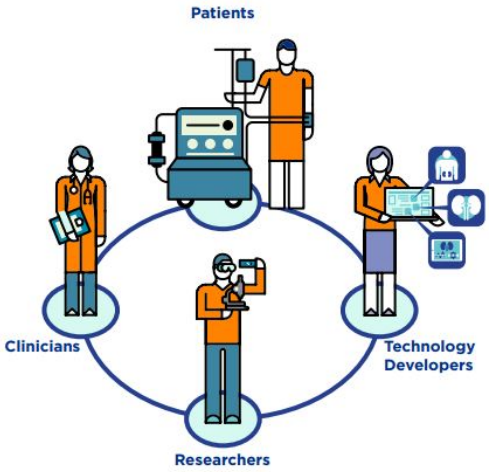


Figure 15: Cooperation future development from Fostering Innovation in Fluid Management [60].

Recently, a report named "Fostering Innovation in Fluid Management" was published by a working group from the Kidney Health Initiative [60]. This group, consisting of patients, clinicians, researchers and technology developers, aimed to provide an outline of the patient priorities on outcome improvement and new technological developments. Examples of such patient priorities were "reduced intradialytic and interdialytic symptoms" and "hemodynamic stabilization to avoid large variability in blood pressure and improve cardiovascular health". This working group specifically proposed technological requirements for development of technical solutions for improved of fluid management by intradialytic approaches. Their suggestions regarding intradialytic monitoring included: integration of treatment data (blood pressure, UFR, etc) (1), provision of recommendations to optimize amount and rate of volume removed during dialysis (2) and accommodate clinicians with information needed to recommend modifications to the dialysis regimen if volume is under or over the target based on input (3). The vision of the current project is considered to align well with the priorities stated in this report as it strives to combine both baseline variables and intradialytic measurements in order to provide clinicians information on the safe margins for fluid removal. This emphasizes the need for further development of machine learning models for the prediction and prevention of IDH to contribute to the current patient priorities.

In future development of the machine learning models, the first recommended step would be to determine the most important features on which the RNN based its predictions. This would render relevant information in regard to the additional value of development of this type of machine learning to a computationally less expensive logistic regression. In the case that the RNN appears to learn from other features in addition to the blood pressure measurements, it would be valuable to assess if the model performance could be further improved by extending the database with sessions from other institutions and additional features.

When comparing the RNN to other models, another statistical model which could be added in the analysis is a linear mixed model (LMM). A LMM can be used for data with repeated measures, e.g. the intradialytic measurements. This model can maintain information and variability in time-varying data and combine it with fixed measures, e.g. baseline variables, which account for individual differences at baseline level [61]. In future research, it should be determined whether such a model, requiring less computational power than an RNN could provide similar results as the current models.

A machine learning-based warning system to improve fluid management could contribute to enhancement of dialysis care and a reduction of adverse events related to fluid imbalance. When the current models would be trained on higher quantities of data to ensure a high variety of patients and practices is included, they could serve as a real-time warning system that is integrated in the dialysis machine. This would require well-developed software that incorporates the machine learning measurements and other patient specific information. Development of such software could be establish by cooperation with companies that provide dialysis equipment, such as Fresenius Medical Care, who have been known to collaborate with researchers for similar purposes [62].

While the current project has focused on the binary prediction of the event of IDH, an alternative approach for implementation could involve assessment of the shift in the predicted probability of the event rather than a yes/no prediction. Evaluation of the trend of predicted probability might provide dialysis staff an indication of the safe margins for fluid removal during dialysis. Issuing a warning when the probability of IDH exceeds a specific threshold could afford the medical staff valuable time to take measures, such as adjusting the patient's position to Trendelenburg or reducing the UFR. Such a tool would provide medical staff with more information for decision-making regarding fluid management,

which could consequently improve patient outcomes.

To enhance the integration of machine learning algorithms into clinical practice, one potential improvement is to increase the frequency of intradialytic measurements. Since blood pressure can show fast fluctuations over time, measurements at an interval of 30 minutes might leave relevant changes unnoticed. However, current blood pressure measurements are uncomfortable for the patient and increasing its frequency would therefore be undesirable. However, a future study on implementation of a multi-parametric wearable sensor will be conducted at the hemodialysis department in the LUMC. This sensor has the ability to measure bioimpedance, electrocardiography (ECG), relative blood pressure and pulse oximetry simultaneously every 5 minutes. This would render high-frequency data that provide more realistic parameter trends that could serve as input for future prediction models and improve their performance.

To provide a definitive answer to the question whether machine learning models would contribute to improved patient monitoring and prevention of IDH, a randomized multi-center clinical intervention trial should be conducted to compare fluid management guided by this warning system to conventional fluid management. In such a trial, patients should be randomly assigned to either the intervention or control group, where the interventions are based on either the warning system or clinical signs and symptoms. The primary outcome of this trial would be the incidence of IDH, and secondary outcomes that could be assessed are cardiovascular events, cerebral ischemia, changes in residual kidney function and mortality. Enrolling the study across multiple institutions would enhance warning system's applicability to a broad population.

Figure 16: Example of warning system incorporated in a dialysis machine.

Given the conclusion that blood pressure measurements play a crucial role in the machine learning-based prediction of IDH, the question rises whether machine learning has relevant added value over the current clinical practice. However, even though current fluid management is guided by many parameters such as the blood pressure and change in blood volume, the incidence of IDH remains high. It is therefore relevant to further investigate whether these models can effectively reduce the incidence rate of IDH by providing more accurate predictions, defining safe margins for fluid removal and enabling timely intervention.

References

[1] *Nefrovisie*, 2021. [Online]. Available: https://www.nefrovisie.nl/nefrodata/.

[2] A. K. Bello *et al.*, "Epidemiology of haemodialysis outcomes", *Nature Reviews Nephrology*, vol. 18, no. 6, pp. 378–395, Jun. 2022, ISSN: 1759-5061. DOI: 10.1038/s41581-022-00542-7.

[3] S. Ahmadmehrabi and W. H. W. Tang, "Hemodialysis-induced cardiovascular disease", *Seminars in Dialysis*, vol. 31, no. 3, pp. 258–267, May 2018, ISSN: 08940959. DOI: 10.1111/sdi.12694.

[4] J. A. Chou, K. Kalantar-Zadeh, and A. T. Mathew, "A brief review of intradialytic hypotension with a focus on survival", *Seminars in Dialysis*, vol. 30, no. 6, pp. 473–480, Nov. 2017, ISSN: 08940959. DOI: 10.1111/sdi.12627.

[5] C. W. McIntyre, "Recurrent Circulatory Stress: The Dark Side of Dialysis", *Seminars in Dialysis*, vol. 23, no. 5, pp. 449–451, Sep. 2010, ISSN: 08940959. DOI: 10.1111/j.1525-139X.2010.00782.x.

[6] J. Yu, Z. Liu, B. Shen, J. Teng, J. Zou, and X. Ding, "Intradialytic Hypotension as an Independent Risk Factor for Long-Term Mortality in Maintaining Hemodialysis Patients: A 5-Year Follow-Up Cohort Study", *Blood Purification*, vol. 45, no. 4, pp. 320–326, 2018, ISSN: 0253-5068. DOI: 10.1159/000486231.

[7] J. Kuipers *et al.*, "The Prevalence of Intradialytic Hypotension in Patients on Conventional Hemodialysis: A Systematic Review with Meta-Analysis", *American Journal of Nephrology*, vol. 49, no. 6, pp. 497–506, 2019, ISSN: 0250-8095. DOI: 10.1159/000500877.

[8] T. Shoji, Y. Tsubakihara, M. Fujii, and E. Imai, "Hemodialysis-associated hypotension as an independent risk factor for two-year mortality in hemodialysis patients", *Kidney International*, vol. 66, no. 3, pp. 1212–1220, Sep. 2004, ISSN: 00852538. DOI: 10.1111/j.1523-1755.2004.00812.x.

[9] B. Charra, "Fluid balance, dry weight, and blood pressure in dialysis", *Hemodialysis International*, vol. 11, no. 1, Jan. 2007, ISSN: 1492-7535. DOI: 10.1111/j.1542-4758.2007.00148.x.

[10] Y. Dou, F. Zhu, and P. Kotanko, "Assessment of Extracellular Fluid Volume and Fluid Status in Hemodialysis Patients: Current Status and Technical Advances", *Seminars in Dialysis*, vol. 25, no. 4, pp. 377–387, Jul. 2012, ISSN: 08940959. DOI: 10.1111/j.1525-139X.2012.01095.x.

[11] B. G. Stegmayr, "Ultrafiltration and Dry Weight-What Are the Cardiovascular Effects?", *Artificial Organs*, vol. 27, no. 3, pp. 227–229, Mar. 2003, ISSN: 0160-564X. DOI: 10.1046/j.1525-1594.2003.07205.x.

[12] M. J. Dekker *et al.*, "Impact of fluid status and inflammation and their interaction on survival: a study in an international hemodialysis patient cohort", *Kidney International*, vol. 91, no. 5, pp. 1214–1223, May 2017, ISSN: 00852538. DOI: 10.1016/j.kint.2016.12.008.

[13] V. Wizemann *et al.*, "The mortality risk of overhydration in haemodialysis patients", *Nephrology Dialysis Transplantation*, vol. 24, no. 5, pp. 1574–1579, May 2009, ISSN: 1460-2385. DOI: 10.1093/ndt/gfn707.

[14] M. Pinsky, P. Brophy, J. Padilla, E. Paganini, and N. Pannu, "Fluid and volume monitoring", *The International Journal of Artificial Organs*, vol. 31, no. 2, pp. 111–126, 2008.

[15] G. E. Thomson, K. Waterhouse, H. P. McDonald, and E. A. Friedman, "Hemodialysis for chronic renal failure. Clinical observations.", *Archives of internal medicine*, vol. 120, no. 2, pp. 153–67, Aug. 1967, ISSN: 0003-9926.

[16] F. M. van der Sande, E. R. van de Wal-Visscher, S. Stuard, U. Moissl, and J. P. Kooman, "Using Bioimpedance Spectroscopy to Assess Volume Status in Dialysis Patients", *Blood Purification*, vol. 49, no. 1-2, pp. 178–184, 2020, ISSN: 0253-5068. DOI: 10.1159/000504079.

[17] S. Ishibe and A. J. Peixoto, "Methods of Assessment of Volume Status and Intercompartmental Fluid Shifts in Hemodialysis Patients: Implications in Clinical Practice", *Seminars in Dialysis*, vol. 17, no. 1, pp. 37–43, Jan. 2004, ISSN: 08940959. DOI: 10.1111/j.1525-139X.2004.17112.x.

[18] J. Schotman *et al.*, "Accuracy of Bioimpedance Spectroscopy in the Detection of Hydration Changes in Patients on Hemodialysis", *Journal of Renal Nutrition*, Dec. 2021, ISSN: 10512276. DOI: 10.1053/j.jrn.2021.11.004.

[19] J. J. Dasselaar, F. M. van der Sande, and C. F. Franssen, "Critical Evaluation of Blood Volume Measurements during Hemodialysis", *Blood Purification*, vol. 33, no. 1-3, pp. 177–182, 2012, ISSN: 0253-5068. DOI: 10.1159/000334142.

[20] L. Pstras, M. Debowska, A. Wojcik-Zaluska, W. Zaluska, and J. Waniewski, "Hemodialysis-induced changes in hematocrit, hemoglobin and total protein: Implications for relative blood volume monitoring", *PLOS ONE*, vol. 14, no. 8, e0220764, Aug. 2019, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0220764.

[21] S. Kron, D. Schneditz, T. Leimbach, J. Czerny, S. Aign, and J. Kron, "Determination of the critical absolute blood volume for intradialytic morbid events", *Hemodialysis International*, vol. 20, no. 2, pp. 321–326, Apr. 2016, ISSN: 14927535. DOI: 10.1111/hdi.12375.

[22] J. Del Risco-Zevallos *et al.*, "Management of acute renal replacement therapy in critically ill cirrhotic patients", *Clinical Kidney Journal*, Jan. 2022, ISSN: 2048-8505. DOI: 10.1093/ckj/sfac025.

[23] B. Sars, F. M. van der Sande, and J. P. Kooman, "Intradialytic Hypotension: Mechanisms and Outcome", *Blood Purification*, vol. 49, no. 1-2, pp. 158–167, 2020, ISSN: 0253-5068. DOI: 10.1159/000503776.

[24] A. Davenport, "Why is Intradialytic Hypotension the Commonest Complication of Outpatient Dialysis Treatments?", *Kidney International Reports*, vol. 8, no. 3, pp. 405–418, Mar. 2023, ISSN: 24680249. DOI: 10.1016/j.ekir.2022.10.031.

[25] M. M. Assimon and J. E. Flythe, "Definitions of intradialytic hypotension", *Seminars in Dialysis*, vol. 30, no. 6, pp. 464–472, Nov. 2017, ISSN: 08940959. DOI: 10.1111/sdi.12626.

[26] C. MacEwen, S. Sutherland, J. Daly, C. Pugh, and L. Tarassenko, "Relationship between Hypotension and Cerebral Ischemia during Hemodialysis", *Journal of the American Society of Nephrology*, vol. 28, no. 8, pp. 2511–2520, Aug. 2017, ISSN: 1046-6673. DOI: 10.1681/ASN.2016060704.

[27] U. C. Anazodo *et al.*, "Hemodialysis-Related Acute Brain Injury Demonstrated by Application of Intradialytic Magnetic Resonance Imaging and Spectroscopy", *Journal of the American Society of Nephrology*, Mar. 2023, ISSN: 1046-6673. DOI: 10.1681/ASN.0000000000000105.

[28] M. M. Assimon, L. Wang, and J. E. Flythe, "Cumulative Exposure to Frequent Intradialytic Hypotension Associates With New-Onset Dementia Among Elderly Hemodialysis Patients", *Kidney International Reports*, vol. 4, no. 4, pp. 603–606, Apr. 2019, ISSN: 24680249. DOI: 10.1016/j.ekir.2019.01.001.

[29] W. Liu, L. Wang, X. Huang, W. He, Z. Song, and J. Yang, "Impaired orthostatic blood pressure stabilization and reduced hemoglobin in chronic kidney disease", *The Journal of Clinical Hypertension*, vol. 21, no. 9, pp. 1317–1324, Sep. 2019, ISSN: 1524-6175. DOI: 10.1111/jch.13658.

[30] C. Zoccali *et al.*, "Chronic Fluid Overload and Mortality in ESRD", *Journal of the American Society of Nephrology*, vol. 28, no. 8, pp. 2491–2497, Aug. 2017, ISSN: 1046-6673. DOI: 10.1681/ASN.2016121341.

[31] J. J. Sands *et al.*, "Intradialytic hypotension: Frequency, sources of variation and correlation with clinical outcome", *Hemodialysis International*, vol. 18, no. 2, pp. 415–422, Apr. 2014, ISSN: 14927535. DOI: 10.1111/hdi.12138.

[32] O. Niel and P. Bastard, "Artificial Intelligence in Nephrology: Core Concepts, Clinical Applications, and Perspectives", *American Journal of Kidney Diseases*, vol. 74, no. 6, pp. 803–810, Dec. 2019, ISSN: 02726386. DOI: 10.1053/j.ajkd.2019.05.020.

[33] N. Tomašev *et al.*, "A clinically applicable approach to continuous prediction of future acute kidney injury", *Nature*, vol. 572, no. 7767, pp. 116–119, Aug. 2019, ISSN: 0028-0836. DOI: 10.1038/s41586-019-1390-1.

[34] N. Lei *et al.*, "Machine learning algorithms' accuracy in predicting kidney disease progression: a systematic review and meta-analysis", *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 205, Dec. 2022, ISSN: 1472-6947. DOI: 10.1186/s12911-022-01951-1.

[35] J.-B. Chen, K.-C. Wu, S.-H. Moi, L.-Y. Chuang, and C.-H. Yang, "Deep Learning for Intradialytic Hypotension Prediction in Hemodialysis Patients", *IEEE Access*, vol. 8, pp. 82 382–82 390, 2020, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2988993.

[36] H. Lee *et al.*, "Deep Learning Model for Real-Time Prediction of Intradialytic Hypotension", *Clinical Journal of the American Society of Nephrology*, vol. 16, no. 3, pp. 396–406, Mar. 2021, ISSN: 1555-9041. DOI: 10.2215/CJN.09280620.

[37] H. W. Kim *et al.*, "Deep Learning Model for Predicting Intradialytic Hypotension Without Privacy Infringement: A Retrospective Two-Center Study", *Frontiers in Medicine*, vol. 9, Jul. 2022, ISSN: 2296-858X. DOI: 10.3389/fmed.2022.878858.

[38] D. A. Freedman, *Statistical Models*. Cambridge: Cambridge University Press, 2009, ISBN: 9780511815867. DOI: 10.1017/CBO9780511815867.

[39] L. Rokach and O. Maimon, *Data Mining with Decision Trees*. WORLD SCIENTIFIC, Oct. 2014, vol. 81, ISBN: 978-981-4590-07-5. DOI: 10.1142/9097.

[40] L. Breiman, "Random Forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, ISSN: 08856125. DOI: 10.1023/A:1010933404324.

[41] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY: Springer New York, 2009, ISBN: 978-0-387-84857-0. DOI: 10.1007/b94608.

[42] T. Chen and C. Guestrin, "XGBoost", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794, ISBN: 9781450342322. DOI: 10.1145/2939672.2939785.

[43] L. R. Medsker and L. Jain, *Recurrent neural networks, Design and Applications*. CRC Press, 2001, vol. 5, pp. 64–67.

[44] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network", Aug. 2018. DOI: 10.1016/j.physd.2019.132306.

[45] *W3schools: Introduction to SQL*. [Online]. Available: w3schools.com/sql/sql_intro.asp.

[46] Castor EDC., *Castor Electronic Data Capture*. 2019. [Online]. Available: https://www.castoredc.com.

[47] K. D. Workgroup, "K/DOQI clinical practice guidelines for cardiovascular disease in dialysis patients", *Am J Kidney Dis*, vol. 45, no. S1–S153, 2005.

[48] J. Kooman *et al.*, "EBPG guideline on haemodynamic instability", *Nephrology Dialysis Transplantation*, vol. 22, no. Supplement 2, pp. ii22–ii44, May 2007, ISSN: 0931-0509. DOI: 10.1093/ndt/gfm019.

[49] R. Mactier, N. Hoenich Ph.D, and C. Breen, "Renal Association Clinical Practice Guideline on Haemodialysis", *Nephron Clinical Practice*, vol. 118, no. s1, pp. c241–c286, 2011, ISSN: 1660-2110. DOI: 10.1159/000328072.

[50] P. Raybaut, *Spyder-documentation*, 2009.

[51] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python", *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[52] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library", Dec. 2019.

[53] H. Zhang *et al.*, "Real-time prediction of intradialytic hypotension using machine learning and cloud computing infrastructure", *Nephrology Dialysis Transplantation*, vol. 38, no. 7, pp. 1761–1769, Jun. 2023, ISSN: 0931-0509. DOI: 10.1093/ndt/gfad070.

[54] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions", May 2017.

[55] R. Assaf and A. Schumann, "Explainable Deep Neural Networks for Multivariate Time Series Predictions", in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, California: International Joint Conferences on Artificial Intelligence Organization, Aug. 2019, pp. 6488–6490, ISBN: 978-0-9992411-4-1. DOI: 10.24963/ijcai.2019/932.

[56] K. Fauvel, T. Lin, V. Masson, É. Fromont, and A. Termier, "XCM: An Explainable Convolutional Neural Network for Multivariate Time Series Classification", *Mathematics*, vol. 9, no. 23, p. 3137, Dec. 2021, ISSN: 2227-7390. DOI: 10.3390/math9233137.

[57] J. Strong, M. Burgett, M. L. Buss, M. Carver, S. Kwankin, and D. Walker, "Effects of calorie and fluid intake on adverse events during hemodialysis", *Journal of Renal Nutrition*, vol. 11, no. 2, pp. 97–100, Apr. 2001, ISSN: 10512276. DOI: 10.1016/S1051-2276(01)51664-7.

[58] R. F. Wolff *et al.*, "PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies", *Annals of Internal Medicine*, vol. 170, no. 1, p. 51, Jan. 2019, ISSN: 0003-4819. DOI: 10.7326/M18-1376.

[59] G. S. Collins *et al.*, "Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence", *BMJ Open*, vol. 11, no. 7, e048008, Jul. 2021, ISSN: 2044-6055. DOI: 10.1136/bmjopen-2020-048008.

[60] Kidney Health Initiative, "Fostering Innovation in Fluid Management", Tech. Rep., 2019. [Online]. Available: https://khi.asn-online.org/uploads/KHI_InnovationsInFluidManagement.pdf.

[61] C. J. Janmaat, M. van Diepen, R. Tsonaka, K. J. Jager, C. Zoccali, and F. W. Dekker, "Pitfalls of linear regression for estimating slopes over time and how to avoid them by using linear mixed-effects models", *Nephrology Dialysis Transplantation*, vol. 34, no. 4, pp. 561–566, Apr. 2019, ISSN: 0931-0509. DOI: 10.1093/ndt/gfy128.

[62] U. Kuhlmann, A. Maierhofer, B. Canaud, J. Hoyer, and M. Gross, "Zero Diffusive Sodium Balance in Hemodialysis Provided by an Algorithm-Based Electrolyte Balancing Controller: A Proof of Principle Clinical Study", *Artificial Organs*, vol. 43, no. 2, pp. 150–158, Feb. 2019, ISSN: 0160-564X. DOI: 10.1111/aor.13328.

Appendix

Example query for retrieving dialysis summary data

This Appendix contains an example of Structured Query Language (SQL) code that was developed to extract dialysis data. This specific query was used to extract dialysis summary data, including measured weight and blood pressure before and after the treatment.

```
/*
DIALYSIS BEFORE AFTER VALUES
Contains: RR before/after, weight before/after, total UF, target weight, fluid administration,
↪  pulse before/after, dialysis start and end time
*/

SELECT
        monitor.patientID AS PatientID
        ,monitor.hdTreatment AS session_treatment
        ,cast(monitor.rangeStart as date) AS date_treatment        --Select the date
        ,cast(monitor.rangeStart as time) AS start_time_treatment --Select the time
        ,cast(monitor.rangeStop as time) AS end_time_treatment --Select the time
        ,treatment.weightBefore AS weight_before_treatment
        ,treatment.weightAfter AS weight_after_treatment
        ,treatment.totalBloodVolume AS total_bv_treatment
        ,treatment.totalUF as total_uf_treatment
        ,treatment.nettoUF AS netto_uf_treatment
        ,treatment.rrBeforesystolic AS sys_before_treatment
        ,treatment.rrBeforediastolic AS dia_before_treatment
        ,treatment.rrAftersystolic AS sys_after_treatment
        ,treatment.rrAfterdiastolic AS dia_after_treatment
        ,treatment.targetWeight AS target_weight_treatment
        ,treatment.cachedFluidAdmin AS fluid_admin_treatment
        ,treatment.pulseBefore AS pulse_before_treatment
        ,treatment.pulseAfter AS pulse_after_treatment

FROM DIAMANT.dbo.HDTreatment treatment
left outer JOIN Diamant.dbo.HDMonitorSession monitor ON monitor.hdTreatment = treatment.id
WHERE monitor.patientID IS NOT NULL
        AND DATEDIFF(hour, monitor.rangeStart, monitor.rangeStop) > 1 --Only obtain session
↪  with duration above one hour.
ORDER BY monitor.patientID, treatment.rangeStart
```

The figures below are screenshots of the Castor database that was developed for data storage, with Figure 17 the patient characteristics, Figure 18 the dialysis data and Figure 19 BCM measurements and blood test results.

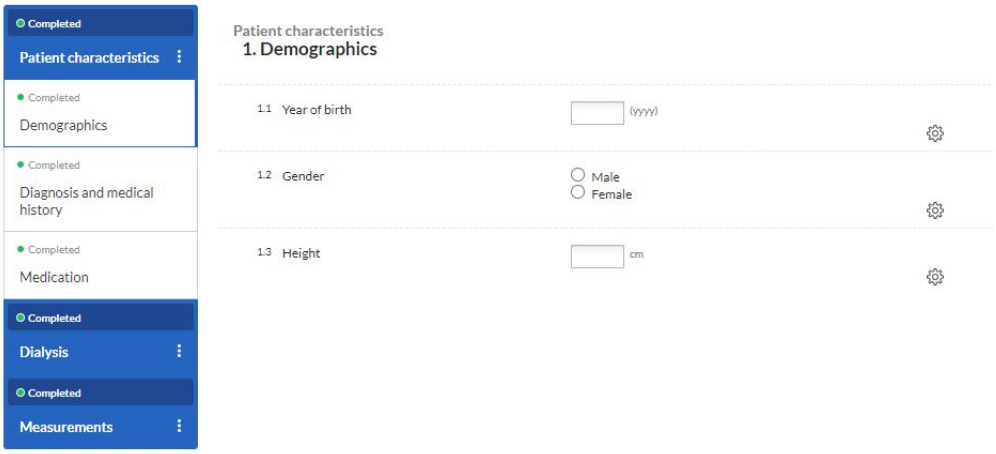


Figure 17: Patient characteristics

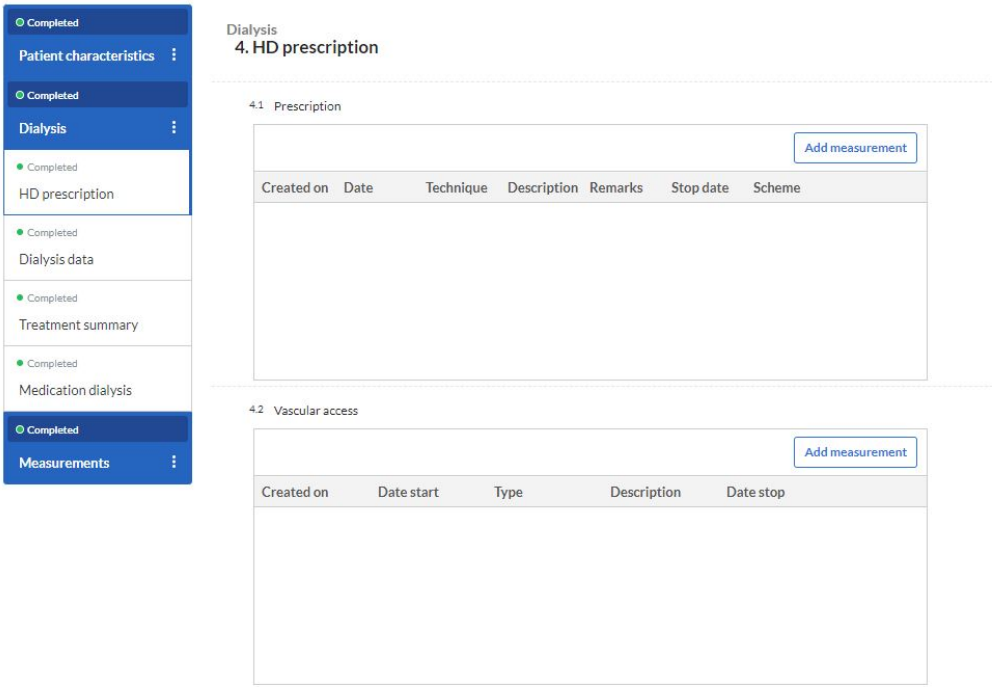


Figure 18: Dialysis data

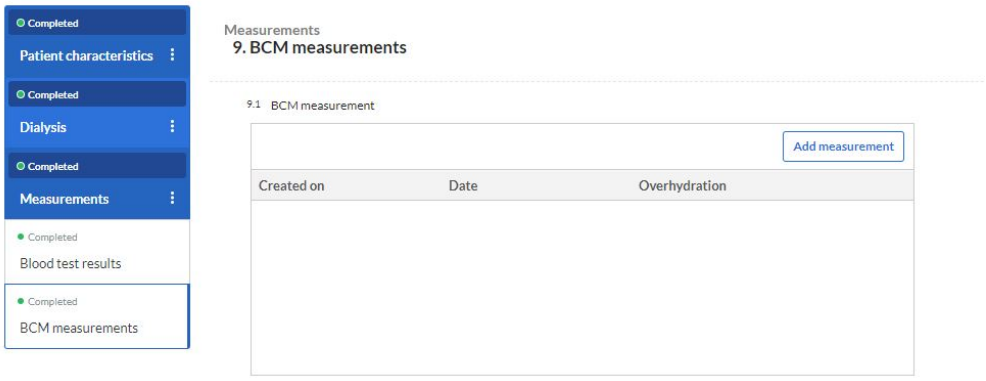


Figure 19: Measurements

Derived features

The medication lists of all included patients were collected to derive whether the patients had diabetes, a cardiovascular history or used antihypertensives

- Diabetes
  - Metformin
  - Insulin
  - Sulfonylurea derivatives: gliclazide, tolbutamine
  - Dipeptidyl peptidase-4 inhibitors: sitagliptine, linagliptine
- Cardiovascular history
  - Anticoagulants: carbasalaatcalcium, acetylsalicylzuur, ascal, clopidogrel, cangrelor, prasugrel, ticagrelor, dipyridamol.
- Antihypertensives
  - Calcium channel blockers: diltiazem, verapamil, isoptin, amlodipine, barnidipine, clevidipine, felodipine, lacidipine, lercanidipine, nicardipine, nifedipine, nimodipine
  - Beta-blockers: acebutolol, atenolol, bisoprolol, carvedilol, celiprolol, esmolol, labetalol, landiolol, metoprolol, nebivolol, propanolol, sotalol.
  - Antihypertensives: clonidine, guanfacine, methyldopa, moxinidine.

Three additional features were calculated form the collected data

Age at dialysis = birthyear - date of dialysis
BMI = weight / length^2
Overhydration = weight before dialysis - target weight

Outliers

Prior to using the features as input of the machine learning models, outliers outside the limits that are provided in Table 5 were removed.

Table 5: Limits for deleting outliers per feature.

Feature	Lower limit	Upper limit
Weight	10	200
Systolic blood pressure	40	300
Diastolic blood pressure	30	150
Overhydration	-20	20
Arterial pressure	-500	0
Venous pressure	0	600
Bicarbonate	0	5
Blood volume change	-20	20
Blood flow rate	100	700
Conductivity	5	20
Dialysate flow rate	200	800
Ionic mass balance	-800	800
Kt/V	0	3
Pulse	30	200
Temperature	35	38
Transmembrane pressure	0	50
Ultrafiltration rate	0	2000
Ultrafiltration volume	0	6000

Kt/V is a measure of the dialysis adequacy.

Feature correlation heatmap

Figure 20 shows a correlation heatmap of all features that were included in the analysis. Red colors indicate high correlations.

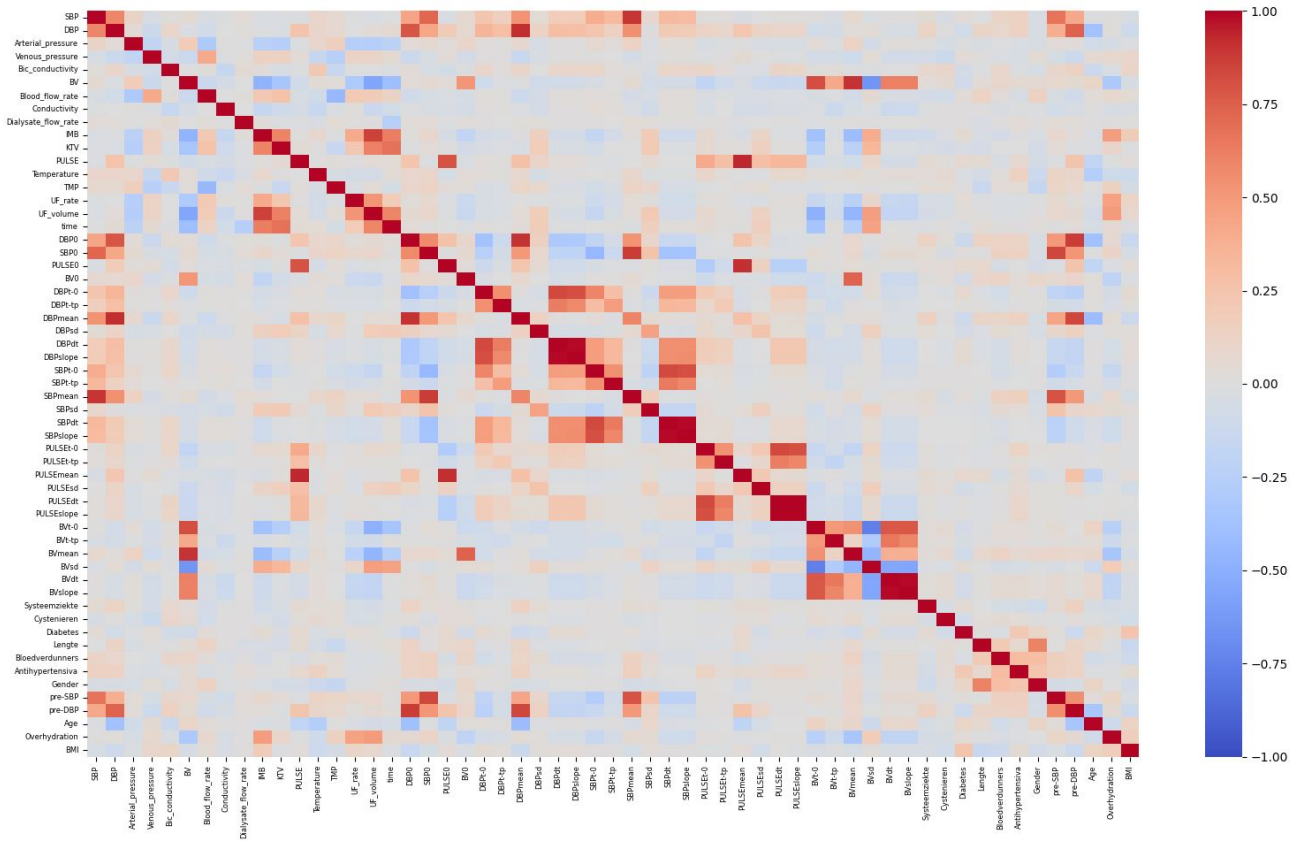


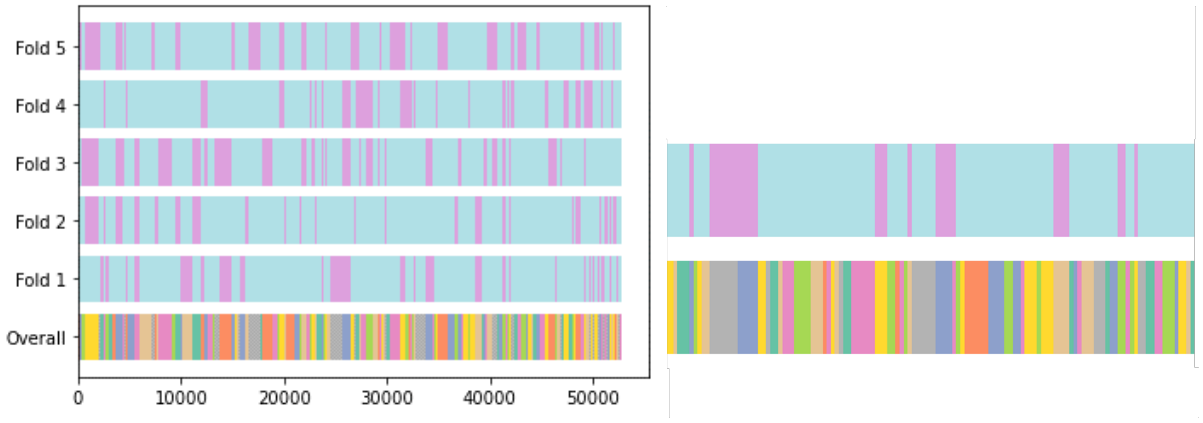
Figure 20: Correlation heatmap of features

Recurrent neural network architecture

```
"""
Architecture and training of recurrent neural network
"""

input_size = 27                  # Number of features
hidden_size = 50                 # Number of neurons
num_layers = 1                   # Number of hidden layers

output_dim = 1                   # 1 for binary classification

batch_size = 300
learning_rate = 0.00001
pos_weight = 5                   # 5 for IDH1 and 50 for IDH2
num_epochs = 300

# Architecture
class RNNmodel(nn.Module):
    def __init__(self, input_size, hidden_size, num_layers):
        super(RNNmodel, self).__init__()
        self.hidden_size = hidden_size
        self.num_layers = num_layers
        self.rnn = nn.RNN(input_size, hidden_size, num_layers, batch_first=True)
        self.fc = nn.Linear(hidden_size, output_dim)

    def forward(self, x):
        h0 = torch.zeros(self.num_layers, x.size(0), self.hidden_size)
        out, h = self.rnn(x, h0)
        out = self.fc(out)

        return out, h


# Model training
def train_rnn(train_dataloader) :
    rnn = RNNmodel(input_size, hidden_size, num_layers)

    criterion = nn.BCEWithLogitsLoss(pos_weight=torch.tensor(pos_weight))
    optimizer = torch.optim.Adam(rnn.parameters(), lr=learning_rate, weight_decay=0.000001)
    scheduler = torch.optim.lr_scheduler.LinearLR(optimizer, start_factor=1, end_factor=0.5,
    total_iters=200, verbose=True)

    for epoch in range(num_epochs):
        rnn.train()
        for i, (inputs, labels) in enumerate(train_dataloader):
            # Clear the existing gradients, else gradients will accumulate to existing gradients.
            optimizer.zero_grad()
            outputs, h = rnn(inputs.float())
            loss = criterion(outputs, labels)
            loss.backward()
            # Initiate gradient descent.
            optimizer.step()

            # Obtain the predicted labels
            train_proba = torch.sigmoid(outputs)
            train_pred = torch.where(train_proba > threshold, torch.tensor(1), torch.tensor(0))

        scheduler.step()
```

Visualization of cross validation

Figure 21, visualizing the train and test distributions over the different folds of the 5-fold cross-validation, was used to validate that sessions of the same patient did not occur in both the train and test set within one fold.



(a) Complete patient set.

(b) Zoomed in.

Figure 21: Visualization of train test split during cross validation. The lower bar shows the number shows the all samples of one patient in the same color. The upper bars show the train (blue) and test (pink) distributions.

Hyperparameters

The options that were assessed in the hyperparameter search are provided in Table 6. The hyperparameter search was performed for both definitions of IDH 30 minutes ahead, 60 minutes ahead and without blood pressure measurements. The results of this search over different folds of the cross-validation are provided in Table 7, 8 and 9 for the logistic regression, random forest and XGBoost respectively.

Table 6: Options for the hyperparameter search.

Classifier	Hyperparameter	Options
Logistic regression	C	0.001, 0.1, 1, 10, 100, 1000
Random forest	Number of trees	1 - 300
	Criterion	Gini, Entropy
	Bootstrap	True, False
	Max depth	2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, None
	Min samples leaf	1, 2, 3, 4, 5
XGBoost	Max depth	2, 3, 4, 5, 6, 7, 8
	Learning rate	0.1, 0.01, 0.05, 0.1, 0.15
	Gamma	0, 0.1, 0.25, 0.5, 1
	Reg lambda	0, 0.5, 1, 10
	Subsample	0.6, 0.7, 0.8
	Colsample bytree	0.5, 0.6, 0.7, 0.8

C is the inverse regularization strength of the logistic regression classifier.

Table 7: Results of the hyperparameter search with the logistic regression.

			Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
IDH1	**30 min**	C	0.001	10	1000	0.1	1
		AUC	**0.80**	**0.76**	**0.81**	**0.79**	**0.82**
	60 min	C	1000	1	1000	100	100
		AUC	**0.75**	**0.76**	**0.78**	**0.78**	**0.78**
	30 min no BP	C	0.001	1000	100	0.001	100
		AUC	**0.64**	**0.62**	**0.66**	**0.56**	**0.64**
IDH2	**30 min**	C	1000	1000	100	100	100
		AUC	**0.93**	**0.93**	**0.94**	**0.91**	**0.94**
	60 min	C	1000	100	10	1000	1000
		AUC	**0.91**	**0.91**	**0.90**	**0.88**	**0.88**
	30 min no BP	C	1000	0.1	1	0.1	0.1
		AUC	**0.81**	**0.70**	**0.57**	**0.67**	**0.81**

Abbreviations: AUC, Area Under the Curve; BP, blood pressure.
IDH1 is intradialytic hypotension defined as a drop in systolic blood pressure of $\geq$20 mmHg compared to baseline, and IDH2, defined as a absolute systolic blood pressure $\leq$90 mmHg. C is the inverse regularization strength of the logistic regression classifier.

Table 8: Results of the hyperparameter search with the random forest.

			Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
IDH1	**30 min**	Number of trees	198	284	25	148	110
		Criterion	gini	entropy	entropy	entropy	gini
		Bootstrap	False	False	True	True	False
		Max depth	20	None	None	None	None
		Min samples leaf	2	3	3	1	3
		AUC	**0.80**	**0.78**	**0.81**	**0.83**	**0.83**
	60 min	Number of trees	81	237	137	172	235
		Criterion	gini	gini	gini	gini	gini
		Bootstrap	True	True	True	True	True
		Max depth	None	None	20	None	None
		Min samples leaf	4	2	2	1	1
		AUC	**0.77**	**0.75**	**0.77**	**0.80**	**0.78**
	30 min no BP	Number of trees	292	101	154	88	290
		Criterion	gini	gini	entropy	entropy	gini
		Bootstrap	False	False	True	True	True
		Max depth	None	30	None	None	None
		Min samples leaf	1	3	1	1	1
		AUC	**0.64**	**0.56**	**0.65**	**0.62**	**0.61**
IDH2	**30 min**	Number of trees	68	90	54	266	83
		Criterion	entropy	entropy	gini	gini	gini
		Bootstrap	False	False	False	True	True
		Max depth	None	30	None	30	None
		Min samples leaf	3	3	2	3	1
		AUC	**0.88**	**0.86**	**0.82**	**0.85**	**0.87**
	60 min	Number of trees	189	120	52	235	176
		Criterion	entropy	entropy	gini	entropy	entropy
		Bootstrap	False	True	True	True	True
		Max depth	20	None	30	30	None
		Min samples leaf	4	2	3	2	1
		AUC	**0.80**	**0.86**	**0.80**	**0.80**	**0.83**
	30 min no BP	Number of trees	218	123	272	51	252
		Criterion	gini	entropy	entropy	entropy	gini
		Bootstrap	True	True	True	False	False
		Max depth	None	None	None	9	None
		Min samples leaf	3	4	1	1	5
		AUC	**0.64**	**0.58**	**0.55**	**0.58**	**0.75**

Abbreviations: AUC, Area Under the Curve; BP, blood pressure.
IDH1 is intradialytic hypotension defined as a drop in systolic blood pressure of $\geq$20 mmHg compared to baseline, and IDH2, defined as a absolute systolic blood pressure $\leq$90 mmHg.

Table 9: Resultsof the hyperparameter search with the XGBoost.

			Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
IDH1	30 min	Max depth	8	8	8	8	8
		Learning rate	0.15	0.15	0.15	0.15	0.15
		Gamma	0	0	1	0.1	0
		Reg lambda	0	0	0	1	0.5
		Subsample	0.8	0.6	0.6	0.8	0.8
		Colsample bytree	0.6	0.8	0.6	0.5	0.6
		AUC	**0.81**	**0.76**	**0.81**	**0.82**	**0.83**
	60 min	Max depth	8	8	8	8	8
		Learning rate	0.15	0.15	0.15	0.15	0.15
		Gamma	0.5	1	0	0	0.1
		Reg lambda	0	1	1	1	1
		Subsample	0.8	0.8	0.6	0.8	0.8
		Colsample bytree	0.8	0.6	0.6	0.8	0.7
		AUC	**0.76**	**0.75**	**0.79**	**0.81**	**0.78**
	30 min no BP	Max depth	8	8	8	8	8
		Learning rate	0.15	0.15	0.15	0.15	0.15
		Gamma	1	0.5	0	0.1	0.5
		Reg lambda	1	10	0	0	1
		Subsample	0.8	0.6	0.7	0.6	0.7
		Colsample bytree	0.5	0.6	0.8	0.8	0.5
		AUC	**0.64**	**0.58**	**0.65**	**0.56**	**0.57**
IDH2	30 min	Max depth	8	8	8	8	8
		Learning rate	0.15	0.15	0.15	0.15	0.15
		Gamma	0.5	1	1	0.25	0.5
		Reg lambda	1	0	0	0	1
		Subsample	0.7	0.8	0.8	0.8	0.7
		Colsample bytree	0.5	0.5	0.5	0.5	0.5
		AUC	**0.87**	**0.79**	**0.86**	**0.81**	**0.89**
	60 min	Max depth	8	8	8	8	8
		Learning rate	0.15	0.15	0.15	0.15	0.15
		Gamma	0.1	1	0	0.25	1
		Reg lambda	0	0	0.5	0.5	1
		Subsample	0.6	0.8	0.8	0.8	0.8
		Colsample bytree	0.7	0.5	0.8	0.7	0.7
		AUC	**0.74**	**0.76**	**0.80**	**0.76**	**0.84**
	30 min no BP	Max depth	8	8	8	8	8
		Learning rate	0.15	0.15	0.15	0.15	0.15
		Gamma	0	0.1	0.1	0.5	0
		Reg lambda	1	0	0.5	10	10
		Subsample	0.8	0.8	0.6	0.7	0.7
		Colsample bytree	0.6	0.8	0.6	0.8	0.7
		AUC	**0.57**	**0.44**	**0.54**	**0.59**	**0.69**

Abbreviations: AUC, Area Under the Curve; BP, blood pressure.
IDH1 is intradialytic hypotension defined as a drop in systolic blood pressure of $\geq$20 mmHg compared to baseline, and IDH2, defined as a absolute systolic blood pressure $\leq$90 mmHg.

Final hyperparameter settings

IDH1

Logistic Regression
- regularization parameter = 10

Random Forest
- number of trees = 150
- criterion = entropy
- bootstrap = True
- maximal tree depth = 10
- minimal samples per leaf = 3

XGBoost
- maximal tree depth = 8
- learning rate = 0.15
- minimal loss reduction = 0.1
- l2 regularization term = 0
- subsampling ratio of samples = 0.8
- subsampling ratio of columns = 0.6

Recurrent Neural Network
- input size = 27
- hidden size = 50
- number of layers = 1
- batch size = 300
- weight of positive samples = 5
- learning rate = $1*10^{-4}$
- weight decay = $1*10^{-6}$
- learning rate scheduler = linear
- number of epochs = 300

IDH2

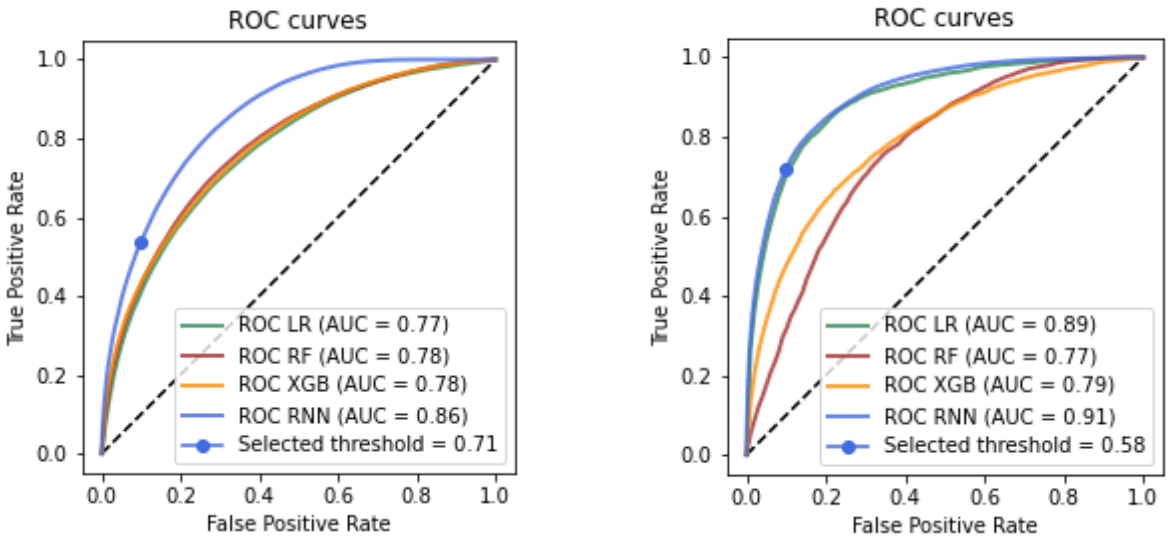
Logistic Regression
- regularization parameter = 100

Random Forest
- number of trees = 100
- criterion = gini
- bootstrap = True
- maximal tree depth = 10
- minimal samples per leaf = 3

XGBoost
- maximal tree depth = 8
- learning rate = 0.15
- minimal loss reduction = 0.25
- l2 regularization term = 0
- subsampling ratio of samples = 0.8
- subsampling ratio of columns = 0.5

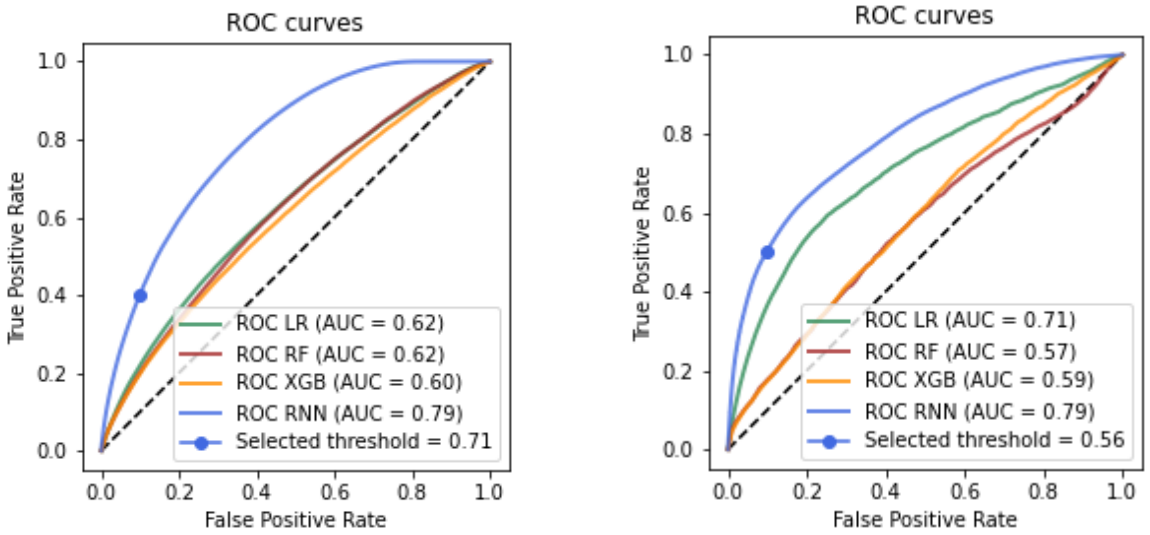
Recurrent Neural Network
- input size = 27
- hidden size = 50
- number of layers = 1
- batch size = 500
- weight of positive samples = 50
- learning rate = $1*10^{-5}$
- weight decay = $1*10^{-6}$
- learning rate scheduler = linear
- number of epochs = 300

The figures below show the ROC curves of the subanalyses that were performed. Figure 22 provides the results of prediction of IDH 60 minutes ahead, instead of 30 minutes and Figure 23 the results after exclusion of all blood pressure measurements.



(a) IDH1 (b) IDH2

Figure 22: ROC curves of the logistic regressoin (LR), random forest (RF), XGBoost (XGB) and recurrent neural network (RNN) for prediction of IDH 60 minutes ahead. IDH1 is intradialytic hypotension defined as a drop in systolic blood pressure of $\geq$20 mmHg compared to baseline, and IDH2 a absolute systolic blood pressure $\leq$90 mmHg.



(a) IDH1 (b) IDH2

Figure 23: ROC curves of the logistic regressoin (LR), random forest (RF), XGBoost (XGB) and recurrent neural network (RNN) for prediction of IDH 30 minutes ahead, but without blood pressure measurements. IDH1 is intradialytic hypotension defined as a drop in systolic blood pressure of $\geq$20 mmHg compared to baseline, and IDH2 a absolute systolic blood pressure $\leq$90 mmHg.