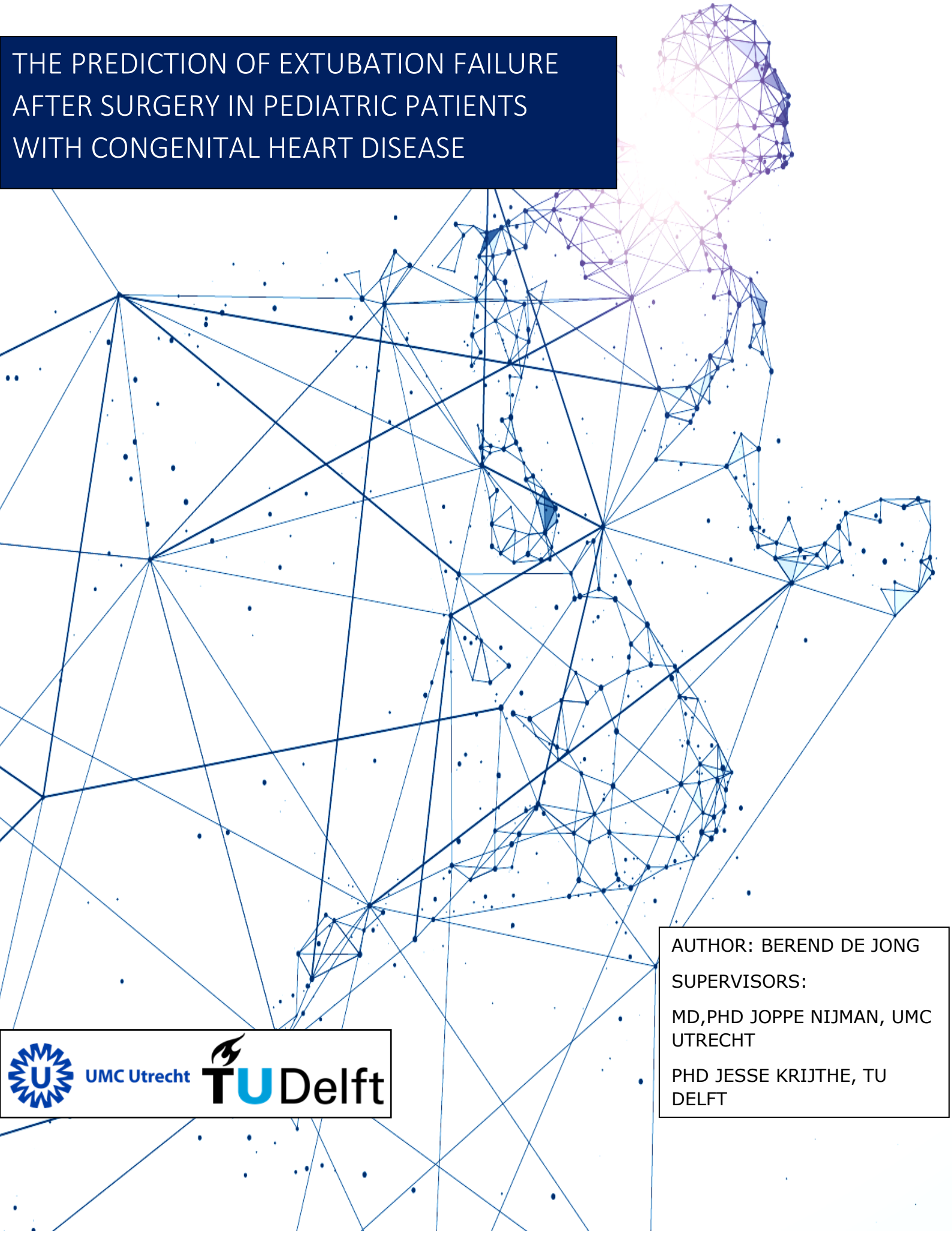
THE PREDICTION OF EXTUBATION FAILURE AFTER SURGERY IN PEDIATRIC PATIENTS WITH CONGENITAL HEART DISEASE



AUTHOR: BEREND DE JONG

SUPERVISORS:

MD,PHD JOPPE NIJMAN, UMC UTRECHT

PHD JESSE KRIJTHE, TU DELFT

UMC Utrecht TU Delft

THE USE OF MACHINE LEARNING ON PICU DATA FOR THE PREDICTION OF EXTUBATION FAILURE AFTER SURGERY IN PEDIATRIC PATIENTS WITH CONGENITAL HEART DISEASE

Berend de Jong

Student number: 4448081

13-7-2021

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)

Pediatric intensive care unit, Wilhelmina Children's Hospital (WKZ)

January 2021 – July 2021

Supervisor(s):

MD, PhD Joppe Nijman, Wilhelmina children's hospital (UMC Utrecht)
PhD Jesse Krijthe, TU Delft

Thesis committee members:

Prof. dr. ir. Jaap Harlaar, Erasmus MC (chair)

MD. PhD Joppe Nijman, Whilhelmina children's hospital (UMC Utrecht)
PhD Jesse Krijthe, TU Delft

Abstract

Background and aims

The timing of extubation is a difficult decision for the medical team on the PICU. With negative impact on patient outcome when extubating too late or too early. The aim of this study was to create machine learning models for extubation failure prediction after surgery in patients with congenital heart disease. The goal was to assess the influence of time variant features on the performance.

Methods

Data from post cardiac surgery patients admitted to the PICU of the University Medical Centre Utrecht, The Netherlands, between 2009 and 2018 was collected. Ventilator and monitor parameters were extracted in 12-hour segments. Different representations of time-variant features were calculated (per hour/ per 12-hour segment), these representations were tested against machine learning trained on only time-invariant features (age, weight diagnosis). Machine learning algorithms tested were: long short-term memory network (LSTM), logistic regression and random forest model. Models were evaluated by comparing the areas under the receiver operator curves

Results

With only time invariant features a performance of 75% [95%CI 81%-90%] using logistic regression. Adding the time-variant features to a LSTM model a performance was reached 77% [95%CI 80%-90%]. Important features from the logistic regression models were age, weight, heart rate and respiratory rate.

Conclusions

Based on the overall results we concluded that the chosen representations of time variant features did not significantly improve the performance of the models. To improve performance and implementation of machine learning models in the future, transparent and externally validated models need to be developed.

Introduction

The incidence of severe congenital heart disease (CHD) that will require expert cardiologic care is quite stable at about 2.5 to 3/1,000 live births (1). This group generally needs surgery in the first year of life, which always requires mechanical ventilation during and after the procedure. Patient outcome can be negatively impacted by prolonged mechanical ventilation (2). Prolonged mechanical ventilation is associated with infection, airway/ lung injury, increased exposure to sedatives, analgesic medications and increased intensive care unit utilization (3). On the contrary, extubation failure (defined as the need for reintubation within 48 hours) has also been independently associated with increased mortality, longer hospitalization, and more days on oxygen and ventilatory support (4). The experience of the medical team is a key factor in the success rate of extubation. Shinkawa, et al. found that the anaesthesiologist who performed the anaesthetic care during surgery was a significant predictor for immediate extubation after heart surgery (5). Scoring systems were developed to aid the medical teams in the evaluation of morbidities in the adult ICU. Schlapbach, et al tried to adapt these scores to age, since these scores were not validated for paediatric patients (6). They came to the conclusion that the scores lack specificity for children and thus new scores specifically designed for children are needed. Furthermore, these types of scores are created on data which is recorded sparsely and manually (6). With the adaptation to Electronic Health Records (EHR) an abundance of data is collected with a higher frequency. This exponential growth of available data could improve the monitoring and prediction of a better moment for extubation. The constant availability of large amounts of data, creates the possibility of dynamic predictions. Dynamic predictions could help with early recognition of extubation readiness as well as extubation unreadiness.

To improve extubation failure prediction, the aim of this study was to create machine learning models for extubation failure prediction after surgery in patients with congenital heart disease. The goal was to assess the influence of time variant features on the performance. In machine learning, the choice of what data to use is debatably more important than the choice of which algorithm. Therefore, the influence of time was investigated, alongside different algorithms. The influence of time was researched by choosing different representations of time variant features in the same period of time. The main algorithms which were investigated are a recurrent neural network (RNN) with a long-short-term memory (LSTM) layer, logistic regression and random forest.

A recurrent neural network is a type of neural network which can handle variable length sequences, which is the case in medical data (7). At every timestep it updates the importance (weight) of features, which results in good trend analysis over time. A diagram of a standard RNN is displayed in figure 1. An improvement on the standard RNN is the LSTM layer which can handle long sequences better due to a forget gate, which prevents a vanishing gradient (near zero gradient, which prevents weights from updating) (8). The vanishing gradient originates from the gradients in the derivatives of the activation functions in an RNN. If the gradients are small or high, the gradient over time will become nearly zero or nearly infinite. This makes the weights in the model unusable. Since the derivative of the forget gate does not have such a gradient it prevents it from vanishing (9).

Due to these properties of the RNN-LSTM, this type of artificial neural network has good time discriminant properties, which makes it suitable for medical prediction problems.

Logistic regression uses a logistic function to calculate the relation between an input variable and an outcome variable for every variable in a dataset. The logistic fuction produces a value between 0 and 1 for a binary classification problem. Optimisation can be performed by choosing the right boundary. The random forest classifier consists of a combination of tree classifiers where each classifier is genareted using a random vector sampled independently from the input vector and each tree casts a unit vote for the most popular class to classigy an input vector.

The main patient cohort in this study were the post-surgical patients with congenital heart disease (CHD), more specific, cardiac surgeries which required cardiopulmonary bypass. Predicting extubation failure in patients with CHD is complicated, since different conditions cause different stable values for several parameters (i.e., blood pressure, saturation, respiratory rate) due to the differences in physiology/ anatomy. To assess the influence of this heterogeneity, a control cohort was investigated. The cohort was patients with bronchiolitis caused by respiratory syncytial virus (RSV). In the US bronchiolitis due to RSV causes approximately 125.000 hospitalisations and 250 deaths every year. Most children will have encountered the virus at least once between 3 months and 2 years, with the bulk of the patients being younger than 1 year (10). These patients generally have the same stable values for different parameters, thus creating a more homogenous cohort than the congenital heart disease cohort.

Related work

Factors associated with extubation failure on the PICU

In a large multi-centre study by Kurachek, et al. several patient factors were associated with extubation failure: <24months, dysgenic or syndromic condition, chronic respiratory or neurological disorder, epinephrine use, steroid use and length of mechanical ventilation. Factors that were not significantly associated with extubation failure were: gender, weight, race, intubating personnel, nasal or oral tube placement, cuff vs uncuffed tube, trauma patients, cardiac patients, oncology patients, RSV patients and oxygen therapies (11). In a study from Laham, et al. no association between ventilator settings and extubation failure or blood gas results and extubation failure were found, except for low respiratory rates on the ventilator (<8) within only 1 day of mechanical ventilation (12). Wratney, et al showed that the air leak test did not predict extubation failure in critically ill paediatric patients (13). Frutos-vivar, et al showed a weak predictive performance of the spontaneous breathing trial for extubation failure (14).

Research utilising machine learning for prediction of extubation failure

Machine learning has already been applied in several studies to predict extubation failure in adults or neonates/ preterm infants. Models either use a small number of vital signs or everything measured at the ICU (vital signs, medication, ventilator data, laboratory values). TingTing, et al used all measured parameters on the ICU and calculated the mean, minimum and maximum value per time variant parameter over the whole length of an admission. They achieved an AUROC of 81% with the XGBoost algorithm on an adult population (mimic-III dataset). Important features of the algorithm were pO2, haemoglobin level, paCO2, mean heart rate and age (15).

Kuo, et al. reached a similar performance with an AUROC of 83% on an adult population with an artificial neural network. They used age, intubation reason, APACHE II score and ventilator data to train their model. They used ventilator data from 30 minute spontaneous breathing trial, which was the only time variant representation used (16). Mikhno, et al investigated a neonatal population, 2 hours prior to the first extubation event, they used the minimum or the maximum value of a parameter based on the Youden index. They reached an AUROC of 87% with logistic regression using only 6 features (17). Mueller et al investigated extubation failure in premature infants, they used the whole admission from preterm infants with data directly from the EHR and all measurements from the neonatal ICU. It was unclear, how they handled the time variance of the features. They reached an AUROC of 76% (18).

Methods

Database

The PICU of the WKZ is one of seven Dutch PICUs, the data in their database called PICURED dates back to 2008. The database contains patient monitor data, ventilator settings, ventilator measurements and lab values. All parameters were stored once per minute, if there was no data available at a certain time point a NULL value was stored for every variable.

Patients with congenital heart disease were identified using a list made by the perfusion specialists at the WKZ. Operation date, surgeons, procedure, diagnosis and hospital number were registered by the perfusionists. Potential selection bias is introduced by this, since only patients (with CHD) which were connected to the cardiopulmonary bypass during surgery were included on the list. Some of the simpler surgical cardiac interventions were not included in this study. The operation dates and patient numbers were used to match the list with the PICURED database. To identify the bronchiolitis patients the Dutch Paediatric Intensive Care Evaluation (PICE) was used (19). PICE is an anonymized dataset of admission and treatment outcome information. The PICURED database was matched with the PICE database based on patient number and admission period.

The definition used for extubation failure in this report was the need for reintubation within 48 hours of the extubation, which conforms to other published works (16, 20, 21). The need for non-invasive ventilation after extubation was not considered as extubation failure. Identification of admissions with extubation failure was done by extracting all ventilator parameters from PICURED. Per admission the sections without ventilator data were identified. These sections were flagged if the gap was longer than 15 minutes and did not exceed 48 hours (based on the extubation failure definition). If a patient had a flagged section, the admissions were manually checked in the physician notes of the electronic health record to confirm extubation failure. Only the admission after surgery was checked for extubation failure for the CHD cohort. Only the bronchiolitis admission was checked for the bronchiolitis cohort. The flagging procedure was the same for both cohorts. If a patient had a failed extubation during admission, all 12-hour segments in the same admission were registered as extubation failure. This as opposed to only labelling the specific window before reintubation as extubation failure.

This prevents extensive imbalance, since the segments on the ventilator without extubation failure were abundantly more frequent than the segments with extubation failure.

Exploratory data analysis

Prior to the development of the prediction models an exploratory data analysis was performed, to learn about the dataset and to reduce the number of parameters from PICURED (~70 parameters, appendix A). Reducing the parameters was performed to improve interpretability of the models. To remove features and keep performance an informed decision should be made which parameters to keep and which to discard. The numeric columns were investigated using Pearson correlation. To complement the exploratory data analysis discussions with physicians about valuable and redundant parameters were performed, to also address the variable selection from a medical perspective. We arbitrary chose to use only the top ten physical parameters which came from the correlation analysis and the discussions with physicians. Aside from the top ten features, demographic features such as age, weight and diagnosis were added to the dataset. Gender was not added, since it was not present in the dataset.

Data preparation

The first pre-processing step we took was eliminating physiologically impossible values. Impossible values are for example a heart rate below 0 or over 300 or a saturation value below 0% or above 100%. These values were replaced by the maximum accepted value (i.e., 0 or 300 for heart rate). Weight data was not available for every patient. To generate a weight the formulas in the article from Tinning, et al. were used which calculate the weight based on a combination of exact age and age category. The diagnosis of each patient (i.e., ventricle septum defect, pulmonary stenosis) were transformed into a risk score using the risk score of the PICU of the Wilhelmina children's hospital, as seen in table 1. After the outlier removing, z-score standardization was applied to every continuous numerical variable.

The admissions were subdivided in periods of 12 hours, thus a 72-hour admission results in 6 different segments. If extubation failure happened during one of those segments, as mentioned before, all 6 segments were considered as extubation failure. Admissions not fully divisible in 12-hour segments were padded with zeros. As stated before, every segment in an admission in which reintubation occurred was labelled as reintubation. No oversampling or under sampling methods were applied to solve the imbalance in group size. The 12-hour segments were further subdivided into 1-hour segments to provide to the models as timesteps. all data where a patient was not connected to the ventilator was discarded, since the goal was to predict extubation failure before a patient is extubated. This was done by removing the timepoints where the ventilator did not measure a respiratory rate and an expiratory tidal volume. This created some gaps in time in the data. To correct for the created gaps a parameter was added which was defined as the time needed to get 720 data points (12 hours' worth of data, one point each minute). If a gap existed of 10 minutes in one segment block, the added parameter would thus have a value of 730 (before z-score standardization).

Model building and training

As a base RNN-LSTM model the model from Kaji,et al. was chosen (22). Kaji et al created a model to predict daily sepsis, myocardial infarction (MI), and vancomycin antibiotic administration over two-week patient ICU courses in the MIMIC-III dataset.

These models achieved next-day predictive AUC of 0.876 for sepsis,0.823 for MI, and 0.833 for vancomycin administration. The RNN-LSTM was implemented in Keras with a TensorFlow backend. 3-dimensional data with patient ICU admissions, time steps (n = 12), and features (n=14) served as input. This input layer was put into an attention layer (23) that weights the inputs. The output of the input layer was put into a masking layer which disregard empty timesteps (for segments without 12 hours' worth of data). This produces variable length sequences which an RNN-LSTM model can handle. Lastly, the output of the masking layer was fed directly to an LSTM layer with 256 hidden units. A hyperbolic tangent was used as an activation function. The output of the network features one dense neuron with a SoftMax activation. A RMSProp optimizer with a learning rate of 0.001, rho of 0.9, epsilon of 1e-08 and no decay were used in all models. From this base model, tuning of the RNN was done manually. The layer types, the number of hidden nodes and the type of activation function were tuned. Binary cross-entropy was used as the loss function for all models.

Simpler models like logistic regression and random forest were used as control. The input for the simpler models needed some extra pre-processing steps then the steps described in the previous paragraph. The simpler models cannot create their own features from the pre-processed data, the features needed to be created manually. We chose to calculate the mean and the standard deviation for each feature for each hour of data. Additionally, the slope of a fitted linear regression for each hour of the data was used to analyse the trend over the time period. Meaning each observation/ sample consisted of 3 values per feature for every hour in 12 hours of an admission. 12-hour segments of a failed extubation and successful extubation for each parameter are shown in appendix B. No extra feature selection method was applied. The hyperparameters of both models were tuned using a 3-fold cross validated randomised search, 20 hyperparameter combinations were evaluated per fold. For logistic regression the following hyperparameters were optimised: solver type, class weight, penalty method and the regularization parameter. For random forest the hyperparameters which were optimised are: the number of estimators, the maximum depth of the tree, minimum samples per split, minimum samples per leaf node. Added characteristic of the simpler models is the ability to identify feature importance. Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. For logistic regression this is the weight of each input feature. Feature importance in random forest corresponds to the reduction of the impurity (uncertainty) on average for a feature in all decision trees in a random forest.

Model validation and evaluation

Hold-out validation was chosen to validate the models (after cross-validated optimization). This was chosen to separate the training and evaluation to prevent overfitting due to leakage. It was opted to split the data in three parts training/ validation/ testing with a percentage split of 70%/ 10%/ 20%.

The train set will be used to train the data. the validation set was used to tune the hyperparameters of the RNN_LSTM. For the simpler models the validation set was added to the training set, since cross validation of the complete training set was used to optimise the hyperparameters. The main performance measure of the model was the area under the receiver operating curve (AUROC). The secondary outcome measures include: F1 Score, average precision score and recall. Accuracy was disregarded as a performance measure due to the imbalance in the data.
F1 score is a measure for accuracy of the model, calculated from the precision and recall, it is sometimes referred to as the harmonic mean between precision and recall. Precision is the ability of the classifier to recognise a negative sample as not a positive sample (distinction between true and false positives). Recall is the ability of the classifier to find all the positive samples.

Experimental set ups

During evaluation of the models, the choice of data was also evaluated. The models reviewed data from the entire admission in 12-hour segments. From a medical perspective this choice was argued, since patients in both groups (failed/ successful extubation) could have had the same ventilator settings at the start of the admission, but could have very different settings at the moment of extubation. Having the same setting in the beginning of the admission results in similar segments, but different outcomes, which could cause a bad performance. Hence why, if the overall performance of most models with 12-hour segments of the entire admission as input was considered bad (maximum AUROC <70%) the choice will be made to only research the 12 hours before the first extubation attempt. In the patient group with failed extubation this means the last 12 hours before the first extubation attempt will be investigated, in the group without extubation failure, it indicates the last 12 hours on the ventilator. Medically, patients are most stimulated to breath on their own when the ventilator settings are the lowest (little support). This is generally right before the extubation moment, thus analysing the last segment could provide better results.

To investigate the influence of time variant features on the model, several representations of the 12 hours were tested. The representations are: Static features, 12-hour features, per hour features and RNN-LSTM. Static features refer to the features which do not change over time during an admission (time invariant). For this study the static features were diagnosis, weight, age and size of the time gap (indication for missingness). The 12-hour features were the mean, SD and slope calculated over the whole 12-hour segment for each parameter, instead of the per hour described above. The per hour features were the mean, SD and slope for every parameter for every hour of the 12-hour segment. The static features were also a part in both the 12-hour features models and the hourly features models. RNN-LSTM creates its own features for the same 12-hour segments as the other models, the static features were also a part of this model.

Python version 3.7.9 was used for the creation of the models/ processing the data. The following python packages were used: Anaconda (version 4.10.3), Keras (version 2.4.3), TensorFlow (version 2.5.0), Scikit-learn (version 0.23.4), SciPy (version 1.5.4), pandas (version 1.2.4) and NumPy (version 1.19.5). The python scripts will be made available on git-hub repository: Berend789/ Extubation_Failure_Prediction

Results

Patient cohorts

After matching the list of the perfusionists and PICURED, 2149 admissions with congenital heart disease were identified. From the 2149 admissions, extubation failure was registered in 127 admissions (5.9%). Detailed information about the patient group is displayed in table 2. On a group level several significant differences were seen between the groups. Since the parameters in the table are not corrected for age, the difference between groups might be exaggerated. The simpler group of bronchiolitis patients consisted of 327 admissions of which 36 admissions (11%) reported extubation failure. Detailed information about the patient group is displayed in table 3. On a group level, less significant differences are seen in comparison with the patients with CHD. This could have resulted from the small group size of the bronchiolitis cohort.

Feature selection

In the PICURED database (parameters in appendix A) laboratory values, ventilator parameters, observations and monitor data are gathered. The sparsity of the laboratory values created an artificially high correlation, due to the high number of zeros in the data. All laboratory values were discarded due to the lack of availability. In the PICURED database, the ventilator parameters are divided in two large groups. The settings as chosen by the medical team and the measured parameters from the machine. High correlation (>90%) was seen between the settings and the measured parameters. Therefore, we chose to discard all settings parameters and keep only the measured ventilator parameters. After discussion with physicians, positive end expiratory pressure (PEEP), peak pressure, respiratory rate, inspired oxygen percentage (fio2) and expiratory tidal volume were selected as parameters. These parameters were complemented by the monitor parameters: heart rate, respiratory rate, end-tidal CO2, saturation and invasive blood pressure (which was only measured if arterial line was present). Other parameters such as blood gas parameters, blood measurement and urine measurement were not incorporated, due to sparse availability of data. In table 4 all used parameters are noted. In appendix B, patient trajectories of various parameters are shown over the 12 hours for both the failed extubation as the successful extubation group.

Full admission

As mentioned in the method section, the choice of time period was also evaluated. In figures 2 and 3, 4 receiver operator curves are seen of different models tested with the 12-hour segments for the entire admission. For both datasets, the models learn relatively little, with the best performing models having AUROC's of around 60-66%. Based on these results, the choice was made to investigate only the last 12 hours on the ventilator or the 12 hours before the first extubation attempt as mentioned in the methods.

Simple models

In figures 4 and 5 the results for the best performing model per time representation for both the CHD cohort and bronchiolitis group respectively are displayed. The performance for the different simpler models were very similar.

The static features in the CHD and bronchiolitis cohorts reached a similar performance to the RNN-LSTM with an AUROC 75% and 71% respectively. The addition of the time variant features for both the 12-hour representation as the hourly corresponded with a marginal improvement of the AUROC for the CHD cohort. In the bronchiolitis cohort dynamic features resulted in a loss of AUROC, with the hourly representation reaching an AUROC of 54% 95%CI [38%-68%]. The overall performance for the bronchiolitis cohort was worse than the CHD cohort. The best performing algorithm in the bronchiolitis cohort was the static feature model, thus the model basing its prediction only on age, weight, time gap and diagnosis. Figures 6, 7 and 8 show the important features for the best performing models on the CHD cohort for the static model, the 12-hour representation and the hourly representation, respectively. In all models weight and age were important features. Moreover, in the 12-hour model and the per hour model, heart rate also was an influential feature. Expiratory tidal volume and the positive end-expiratory pressure (PEEP) were the most influential ventilator parameters.

RNN-LSTM

Figure 9 shows the results for different RNN-LSTM configurations on the CHD cohort. The configurations shown are the base RNN-LSTM with a different optimizer called Adam (24), the base RNN-LSTM with forward and back propagation (bidirectional) (25), a simple RNN without LSTM layer, a RNN with a gated recurrent unit (LSTM without an output gate and fewer parameters) (26) and the base RNN-LSTM without an attention layer (23). Without the attention layer the AUROC dropped from 77% 95%CI [76%-78%] to about 59% 95%CI [57%-61%]. Altering the standard forward propagated recurrent neural network to a bidirectional neural network did not improve performance with an AUROC of 74% 95%CI [72%-75%]. It was clear that the base model from Kaji, et al. reached the highest AUROC on this cohort (22).
The results for the bronchiolitis cohort are displayed in figure 10. Performance for all models ranged between 48%-68%. It was clear that the neural networks performed better on the CHD cohort. The best performing model on the bronchiolitis group is the bidirectional RNN-LSTM with an AUROC of 68% 95%CI [64%-71%].

The best model of the report was the base RNN-LSTM model on the CHD cohort. In figure 11 the distribution of the probabilities of the prediction (the probability of extubation failure) for the successful extubations and the failed extubations are shown as percentage of the group. The spread of the distribution is a measure of uncertainty around the prediction. Mean probability for the failed extubation group was 0.6 95%CI [0.40-0.78], for the successful extubation group the mean probability was 0.33 95%CI [0.001-0.71] The sensitivity and specificity of the model were 0.935, 0.527 respectively. The positive predictive value (PPV) and negative predictive value (NPV) were 0.094, 0.993, when the model was optimized for the highest AUROC. In table 5, precision, recall and f1 score are shown.

Discussion

Main findings

the aim of this study was to create machine learning models for extubation failure prediction after surgery in patients with congenital heart disease. The goal was to assess the influence of time variant features on the performance. The best model was the base RNN-LSTM with an AUROC of 77%, the best performing simpler model was a logistic regression model on the per hour feature data. This model had an AUROC of 76%. The performance of the models was worse than some of the models reported in the introduction where model performances were between 76%-87%, although these results were on adults and neonate populations (15-18).

The models performed well without the usage of the dynamic features, the static model for the CHD cohort reached an AUROC of 75% 95%CI [67%-82%] and for the bronchiolitis cohort the AUROC was 71% 95%CI [60%-82%]. The influence of all added dynamic features was marginal. The separability within the features could not generalise to the whole population due to the limited size of the data and especially the limited size of the failed extubation group. This lack of generalisation could be enhanced by the differences in physiology between age groups. However, the incorporation of weight and age in the feature set should counteract the physiological differences between the age groups. Secondly, patients are challenged the most with low ventilator settings and high patient effort one or two hours before extubation. The overlap between ventilator measurements and monitor parameters in the first half of the 12 hours between the failed and successful extubation group might have added noise to the prediction, which caused the bad performance. Analysing different time windows could give a definitive answer.

The results of the bronchiolitis cohort were not comparable with the CHD cohort. This is probably due to the limited size of the bronchiolitis group. The number of extubation failures in the test group became too small for reliable prediction and comparison with the CHD cohort. It was therefore difficult to form a conclusion about the influence of the heterogeneity in the congenital heart disease cohort on the performance.

Analysing the 12 hours before the first extubation attempt instead of the full admission leaded to improved AUROC. Medically, this choice does not impact the simpler procedures like VSD and ASD corrections since the majority of patients would not have received more than 12 hours of ventilation (3). However, after more complicated surgery or after a postoperative complication, large amounts of data were lost. Physicians do not usually incorporate detailed information (i.e., heart rate at admission) of the full admission into their decision to extubate, especially when a patient is admitted weeks/ months. They do incorporate a previous extubation attempt, postoperative complication and surgical procedure. So, from current medical views it is relevant to analyse only the last 12 hours of extubation with the addition of such impactful information. Nevertheless, the question could be asked if a baseline should be incorporated for better performance.

Age, weight, heart rate and respiratory rate seemed to be the most influential parameters across the different models.
It is noteworthy that in the static model, weight and age are counteracting each other, but in the dynamic models the features complement each other. From a medical perspective younger age is associated with a lighter weight, which both are associated with a worse extubation outcome (11). The difference could have originated from the interactions between features.

Limitations

Without external validation the performance of machine learning models cannot be reliably determined. The 1 centre data as used in this study cannot prevent overfitting or bias. Despite using techniques like hold-out or cross-validation since the test population is probably very similar to the training population, overfitting cannot be completely disregarded. National external validation and international external validation could prove the worth of the machine learning model.

Configuration tuning of the RNN-LSTM was not done optimally. The aim of the research was to investigate the influence of the dynamic representations. Because of this, optimisation of the neural network was done manually and not every possible configuration was considered. Further optimization could result in an algorithm which performs a bit better. However, with extensive optimisation, there is also an increased risk of overfitting.

Due to the mathematical background of the LR/ RF the input is restricted. The model essentially needs a 1D representation of an observation. The observation in this study was chosen as 12-hour intervals for every admission. This 1D representation limits the amount of temporal information that can be given per variable. The representation as done in this report gives the information in feature form. The algorithm does not know the temporal relation between the features, so a heart rate at 1 hour will also be linked to a respiratory rate at 9 hours, while this linkage is not considered relevant. Adding to many features/ temporal information could therefore cause noise in the data, which could possibly result in a worse performance. A countermeasure could be to use features derived from simpler trend analysis, like the slope of the fitted linear regression used in this model.

The used data currently lacks one major aspect of care, namely interventions. If a patient gets worse during admission, the medical team will intervene to prevent the patient from getting sicker. One of the main interventions on the PICU is medication. Types of medication groups used on the CHD cohort are inotropes and sedatives. Both of these medication groups will directly impact a variety of the used parameters (i.e., respiratory rate/ heart rate). Moreover, if at a certain point medication is given, it could possibly mean that a patient is not ready to be extubated yet or have an influence on the extubation readiness. Knowing this could improve the models. Dosage information and precise medication (i.e., noradrenaline or phenylephrine) will probably not be of use, since this information will be widely dispersed over time and different for most patients. Knowing which medication group is given at a certain time point will probably have enough information to be used by a machine learning algorithm. With a dataset sufficiently large it could possibly learn the interactions between the medication and the

other parameters. Diagnostic interventions like a thorax x-ray will probably have less predictive value, since it will mostly indicate a bad or worsened state of the patient. A risk with implementing intervention in prediction models is that physicians change or wait with their interventions based on the model, which could lead to poor predictions of the model and suboptimal healthcare. For example, unnecessary risk can be taken with lowering the dosage of sedatives, since a lowering the dose could be associated with better outcome. However, quickly lowering the dose of sedatives could lead to delirium and other withdrawal symptoms which could prolong the ventilator duration. Prospective research and a good implementation plan could identify and potentially prevent such a relation.

Future research

The influence of the different representations in time as shown by this article, raised the question if analysing 12 hours before extubation is relevant. From a medical perspective, the patient is stimulated/ challenged the most just before extubation. At this point in time the ventilator parameters are at their lowest and the patient needs to show the physicians that he/ she can breathe on his/ her own. The tests performed by some physicians such as the aforementioned spontaneous breathing trial are not predictive enough for extubation failure, even when analysed with machine learning (13, 14). Devising a new protocol for the last hours for a weaning test may be an option to improve prediction. Prospectively collecting the data and comparing with the retrospective data could provide valuable insight in the mechanisms behind extubation failure.

Furthermore, besides the 12 hours before extubation which was chosen, it was also chosen to represent those 12 hours in hourly segments. A popular open source database for machine learning on adult ICU data is the mimic-III database only contains measurements per hour (27). Research with this database has led to good predictive performance for various subject (17, 28-30). However, it is imaginable that calculations per half-hour or quarters could lead to better results. The so-called dimensionality reduction from the minute data in the PICURED database was not investigated in this study, but is a field of interest to potentially increase performance. The choice of how data is presented is very important. This was partially shown in this report by the difference in results between analysing the full admission and only the last 12 hours before extubation.

External validation is a big problem in medical machine learning research. In a systematic review by Shillan, et al. only 10 of 258 studies (4%) were externally validated (31). The lack of externally validated methods hinders the implementation of machine learning methods in the hospital. Hospitals are hesitant in sharing their data, since the information is heavily privacy sensitive. To counteract the negatives of sharing data the idea is to send the machine learning algorithm across the world. With this method every hospital keeps control over their data and external validation is made possible. Future research should focus on minimizing or solving the differences in data between different hospitals to make wide spread external validation possible.

Furthermore, Van de Sande, et al. performed a systematic review and found that only 10 out of 494 studies (2%) clinically evaluated the machine learning model with most studies in the report not progressing after than the testing and prototyping environment (32). They identified several boundaries for further progression or implementation of machine learning models besides boundaries related to or solvable by external validation. Boundaries they named were: variations in local practices, transparency of models and patient safety. Transparency and patient safety are heavily correlated, if machine learning makes understandable predictions (transparency), physicians can make informed decisions on how to handle on the results (i.e., patient safety). Before model building it may be a useful step to create a clinical evaluation plan and possibly also an implementation plan. Doing this could prevent the creation of clinically irrelevant or infeasible models. It will also simplify incorporating commercial algorithms into the general workflow is a similar plan is already in place on the ward.

Conclusion

In this report machine learning models were created for extubation failure prediction after surgery in patients with congenital heart disease. The goal was to assess the influence of dynamic features (time variant) on the performance. Using an RNN-LSTM an AUROC was reached of 77%, while a simpler model with only static features (time invariant) reached an AUROC of 75%. Based on the overall results we concluded that the chosen representation of time in the datasets did not significantly improve the performance of the models. To improve performance and implementation of machine learning models in the future, transparent and externally validated models need to be developed. Furthermore, before developing machine learning models, a clinical evaluation/ implementation plan should be in place to objectively test actual performance of models.

Reference List

1. Hoffman JIE, Kaplan S. The incidence of congenital heart disease. Journal of the American College of Cardiology. 2002;39(12):1890-900.
2. Alghamdi AA, Singh SK, Hamilton BC, Yadava M, Holtby H, Van Arsdell GS, et al. Early extubation after pediatric cardiac surgery: systematic review, meta-analysis, and evidence-based recommendations. Journal of cardiac surgery. 2010;25(5):586-95.
3. Rooney SR, Donohue JE, Bush LB, Zhang W, Banerjee M, Pasquali SK, et al. Extubation Failure Rates After Pediatric Cardiac Surgery Vary Across Hospitals. Pediatric critical care medicine : a journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies. 2019;20(5):450-6.
4. Chawla S, Natarajan G, Shankaran S, Carper B, Brion LP, Keszler M, et al. Markers of Successful Extubation in Extremely Preterm Infants, and Morbidity After Failed Extubation. The Journal of pediatrics. 2017;189:113-9.e2.
5. Shinkawa T, Tang X, Gossett JM, Dasgupta R, Schmitz ML, Gupta P, et al. Incidence of Immediate Extubation After Pediatric Cardiac Surgery and Predictors for Reintubation. World journal for pediatric & congenital heart surgery. 2018;9(5):529-36.
6. Schlapbach LJ, Straney L, Bellomo R, MacLaren G, Pilcher D. Prognostic accuracy of age-adapted SOFA, SIRS, PELOD-2, and qSOFA for in-hospital mortality among children with suspected infection admitted to the intensive care unit. Intensive Care Medicine. 2018;44(2):179-88.
7. Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H. State-of-the-art in artificial neural network applications: A survey. Heliyon. 2018;4(11):e00938-e.

8. Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735-80.
9. Pascanu R, Mikolov T, Bengio Y, editors. On the difficulty of training recurrent neural networks2013: PMLR.
10. Piedimonte G, Perez MK. Respiratory syncytial virus infection and bronchiolitis. Pediatr Rev. 2014;35(12):519-30.
11. Kurachek SC, Newth CJ, Quasney MW, Rice T, Sachdeva RC, Patel NR, et al. Extubation failure in pediatric intensive care: a multiple-center study of risk factors and outcomes. Critical care medicine. 2003;31(11):2657-64.
12. Laham JL, Breheny PJ, Rush A. Do Clinical Parameters Predict First Planned Extubation Outcome in the Pediatric Intensive Care Unit? Journal of Intensive Care Medicine. 2013;30(2):89-96.
13. Wratney AT, Benjamin Jr DK, Slonim AD, He J, Hamel DS, Cheifetz IM. The endotracheal tube air leak test does not predict extubation outcome in critically ill pediatric patients. Pediatric critical care medicine: a journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies. 2008;9(5):490.
14. Frutos-Vivar F, Ferguson ND, Esteban A, Epstein SK, Arabi Y, Apezteguía C, et al. Risk Factors for Extubation Failure in Patients Following a Successful Spontaneous Breathing Trial. Chest. 2006;130(6):1664-71.
15. Chen T, Xu J, Ying H, Chen X, Feng R, Fang X, et al. Prediction of extubation failure for intensive care unit patients using light gradient boosting machine. IEEE Access. 2019;7:150960-8.
16. Kuo H-J, Chiu H-W, Lee C-N, Chen T-T, Chang C-C, Bien M-Y. Improvement in the Prediction of Ventilator Weaning Outcomes by an Artificial Neural Network in a Medical ICU. Respiratory Care. 2015;60(11):1560.
17. Mikhno A, Ennett CM, editors. Prediction of extubation failure for neonates with respiratory distress syndrome using the MIMIC-II clinical database2012: IEEE.
18. Mueller M, Almeida JS, Stanislaus R, Wagner CL. Can machine learning methods predict extubation outcome in premature infants as well as clinicians? Journal of neonatal biology. 2013;2.
19. PICE s. PICE Complicatie Registratie

Data- en kwaliteitsregistratie van de Nederlandse ICK www.PICE.nl: © 2020 PICE Pediatrische Intensive care evaluatie.; 22 December 2020 [
20. Savi A, Teixeira C, Silva JM, Borges LG, Pereira PA, Pinto KB, et al. Weaning predictors do not predict extubation failure in simple-to-wean patients. Journal of Critical Care. 2012;27(2):221.e1-.e8.
21. Saugel B, Rakette P, Hapfelmeier A, Schultheiss C, Phillip V, Thies P, et al. Prediction of extubation failure in medical intensive care unit patients. Journal of Critical Care. 2012;27(6):571-7.
22. Kaji DA, Zech JR, Kim JS, Cho SK, Dangayach NS, Costa AB, et al. An attention based deep learning model of clinical events in the intensive care unit. PLOS ONE. 2019;14(2):e0211057.
23. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:14090473. 2014.
24. Zhang Z, editor Improved adam optimizer for deep neural networks2018: IEEE.
25. Thomas. Benchmarking of LSTM Networks. arXiv pre-print server. 2015.
26. Cho K, Bart, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv pre-print server. 2014.
27. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016;3(1):160035.
28. Johnson AEW, Mark RG. Real-time mortality prediction in the Intensive Care Unit. AMIA Annu Symp Proc. 2018;2017:994-1003.
29. Celi LA, Galvin S, Davidzon G, Lee J, Scott D, Mark R. A Database-driven Decision Support System: Customized Mortality Prediction. Journal of Personalized Medicine. 2012;2(4):138-48.
30. Garcia-Gallo JE, Fonseca-Ruiz NJ, Celi LA, Duitama-Munoz JF. A machine learning-based model for 1-year mortality prediction in patients admitted to an Intensive Care Unit with a diagnosis of sepsis. Medicina Intensiva. 2020;44(3):160-70.
31. Shillan D, Sterne JAC, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Critical Care. 2019;23(1):284.
32. van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med. 2021;47(7):750-60.

Figures

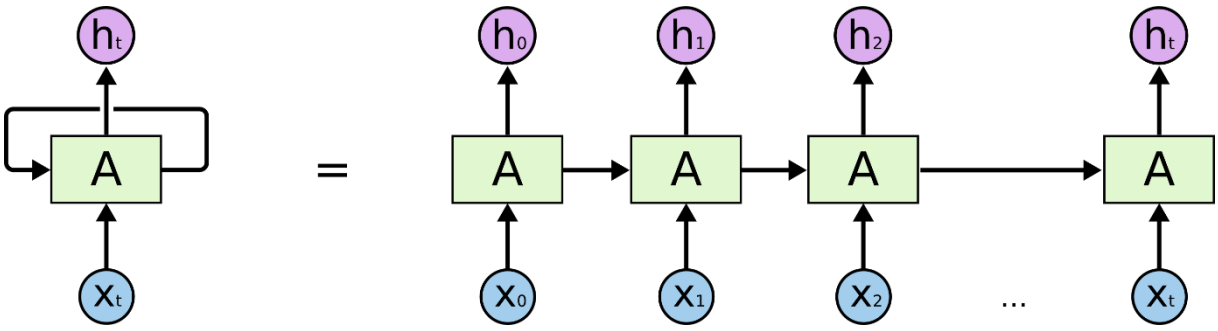


Figure 1: Diagram of standard recurrent neural network. X is the input at a certain timestep, A corresponds to a recurrent layer, h is the output at a certain timepoint. In a long short-term neural network (LSTM), A would correspond with the LSTM layers

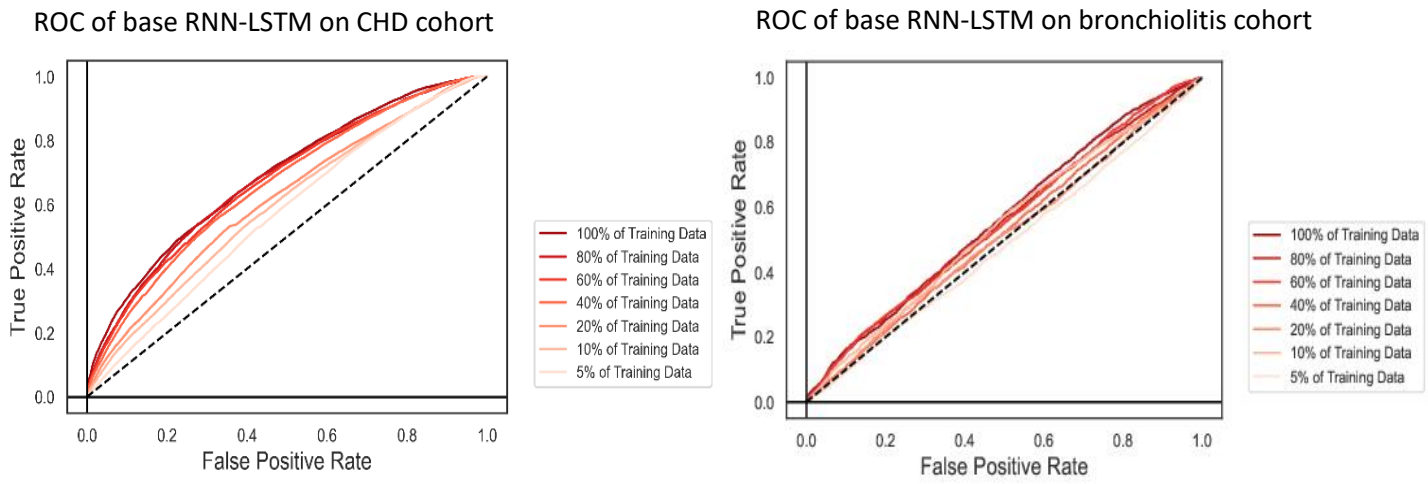


Figure 2: Results from the base RNN-LSTM model on the data from the full admission divided in 12 hour segments for CHD cohort (left figure) and bronchiolitis cohort (right figure).

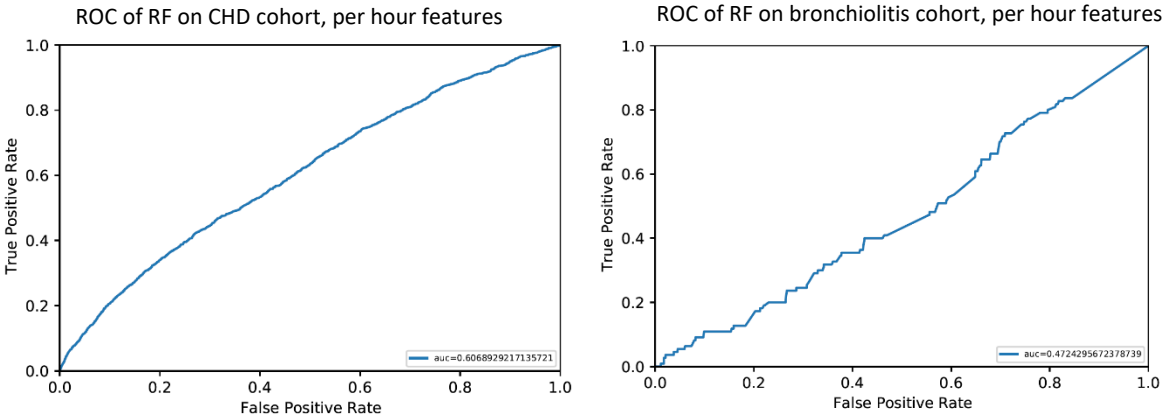


Figure 3:Results from the random forest models (RF) on the data from the full admission dividid in 12-hour segments. Results for the features calculated per hour (per 12-hour segment) on the left for the congenital heart disease cohort and for the bronchiolitis

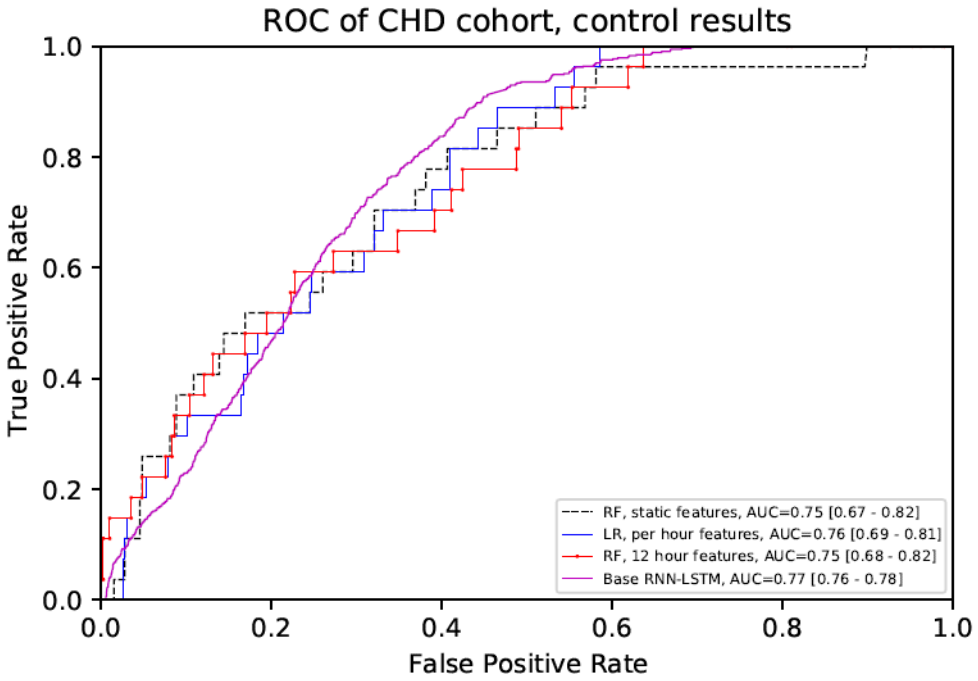


Figure 4:Receiver operating curves (ROC) of the control models, logistic regression (LR) and random forest (RF), on the congenital heart disease cohort. (AUC=Area under receiver operating curve, RNN-LSTM = recurrent neural network with long short-term memory)

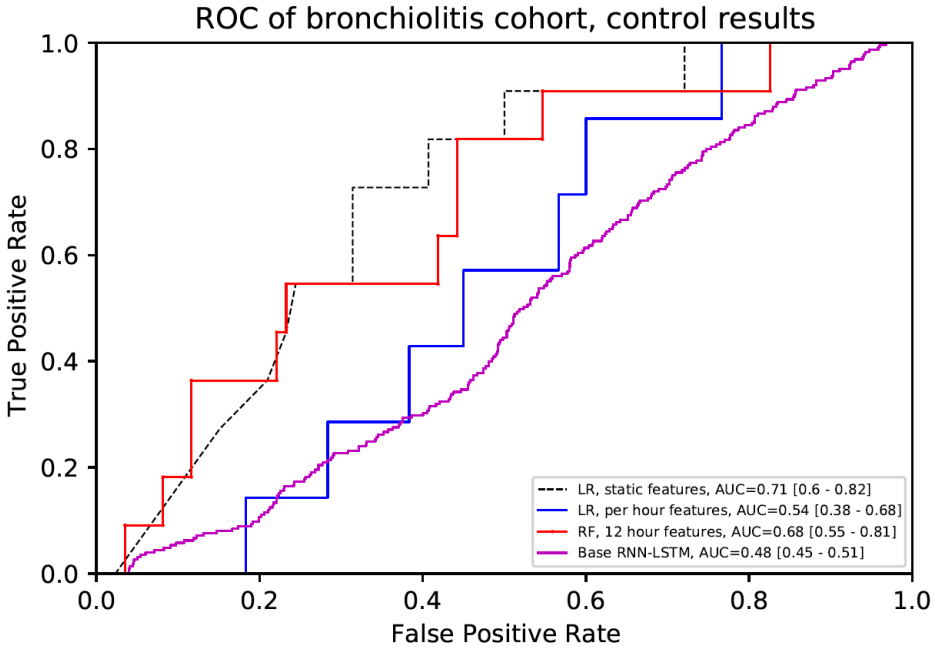


Figure 5: Receiver operating curves (ROC) of the control models, logistic regression (LR) and random forest (RF), on the bronchiolitis cohort. (AUC=Area under receiver operating curve, RNN-LSTM = recurrent neural network with long short-term memory)

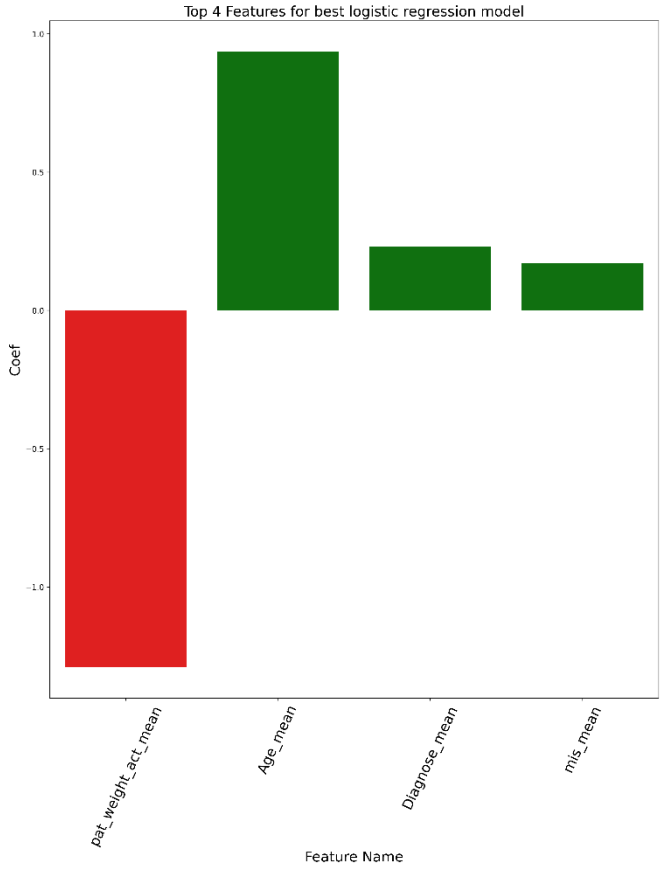


Figure 6: Feature importance's of control models for static models, 12 hour feature models and the per hour feature. Feature names are displayed as parameter_statistic (Abbreviations for parameters are explained in appendix A).

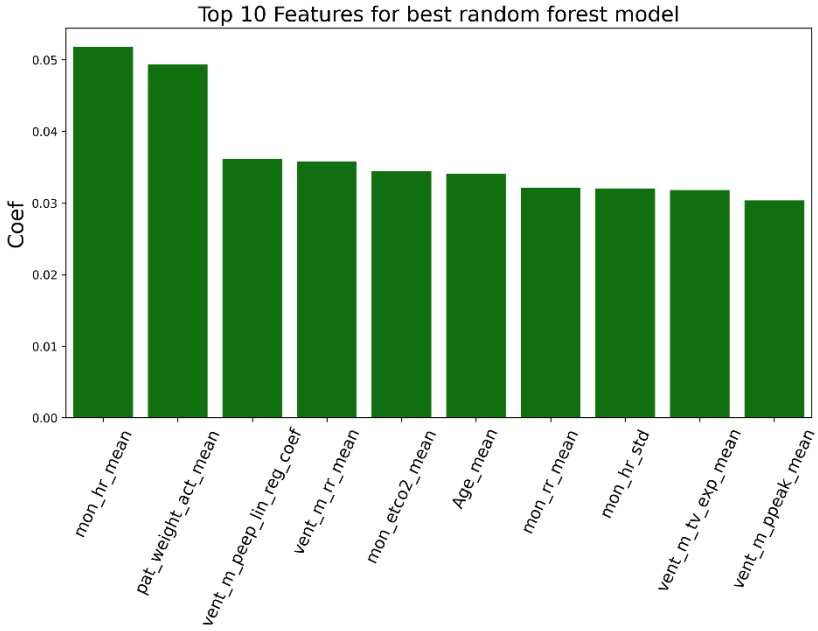


Figure 7:Feature importance's of control model for the 12 hour feature models and the per hour feature. Feature names are displayed as parameter_statistic (Abbreviations for parameters are explained in appendix A). (lin_reg_coef= Slope of fitted linear regression, std = standard deviation)

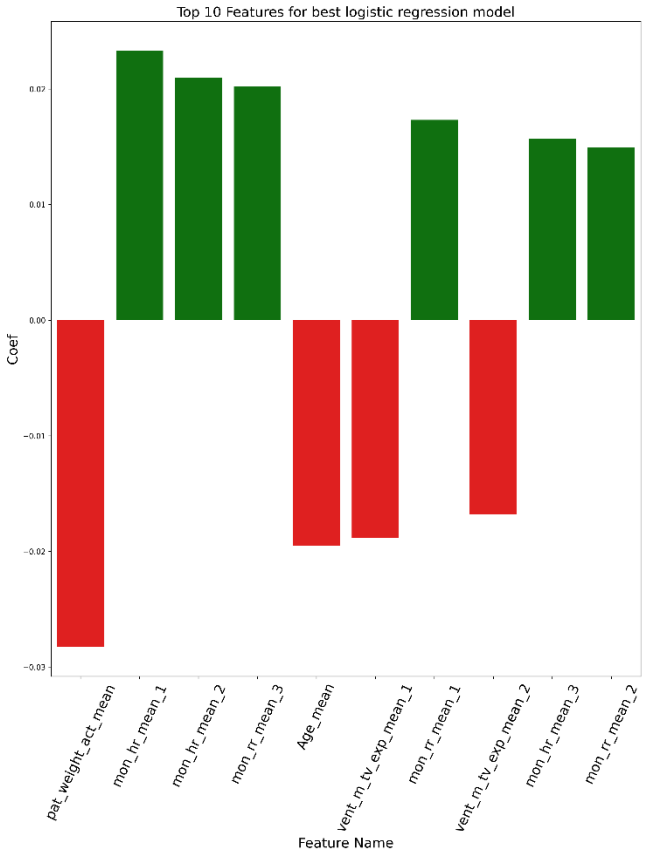


Figure 8: Feature importance's of control models for the per hour feature. Feature names are displayed as parameter_statistic_hour of admission (Abbreviations for parameters are explained in appendix A). Hour of admission 1 corresponds with 12 hours before extubation

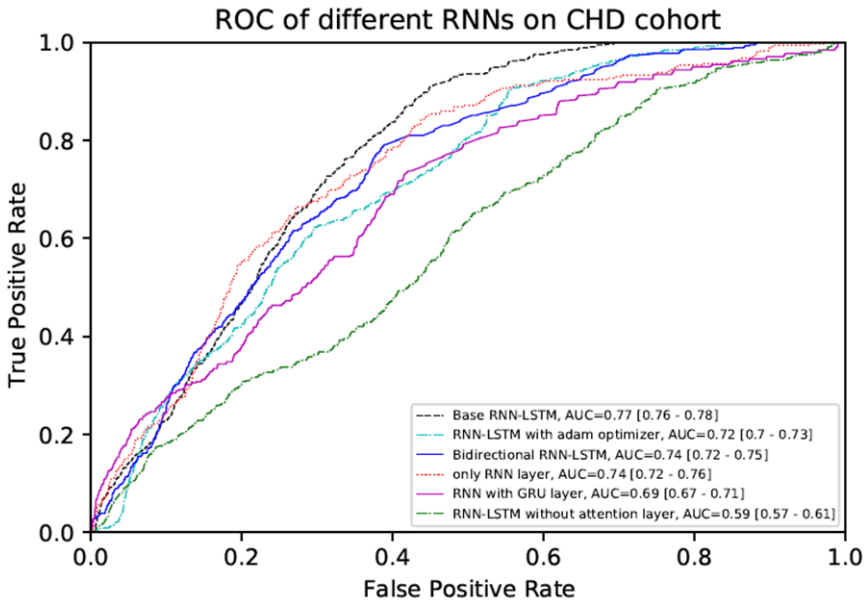


Figure 9: Receiver operating curves (ROC) of 5 different recurrent neural networks (RNN) layouts fitted on the congenital heart disease (CHD) cohort. (GRU= Gated recurrent unit, AUC=Area under receiver operating curve, LSTM= long short-term memory)

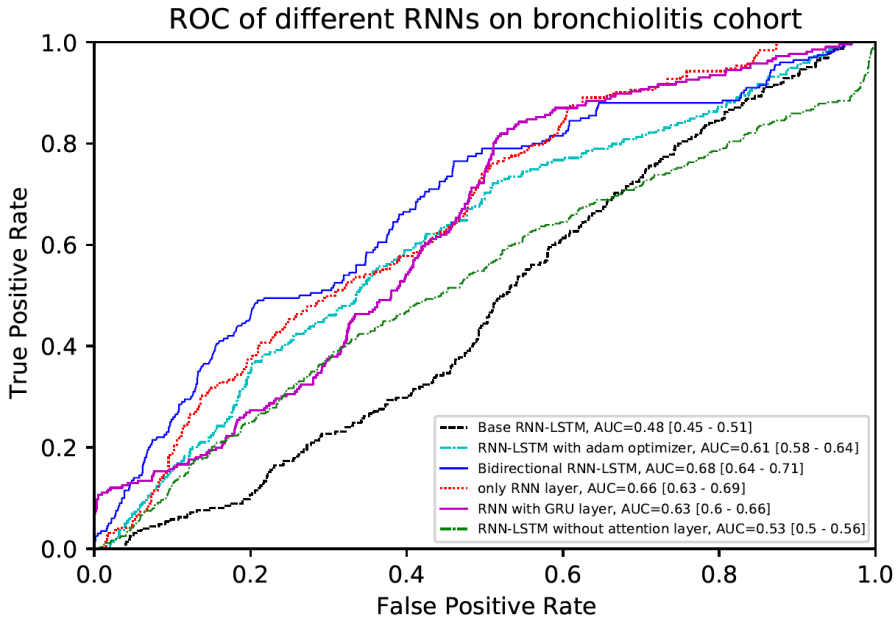


Figure 10: Receiver operating curves (ROC) of 5 different recurrent neural networks (RNN) layouts fitted on the bronchiolitis cohort. (GRU= Gated recurrent unit, AUC=Area under receiver operating curve, LSTM= long short-term memory)

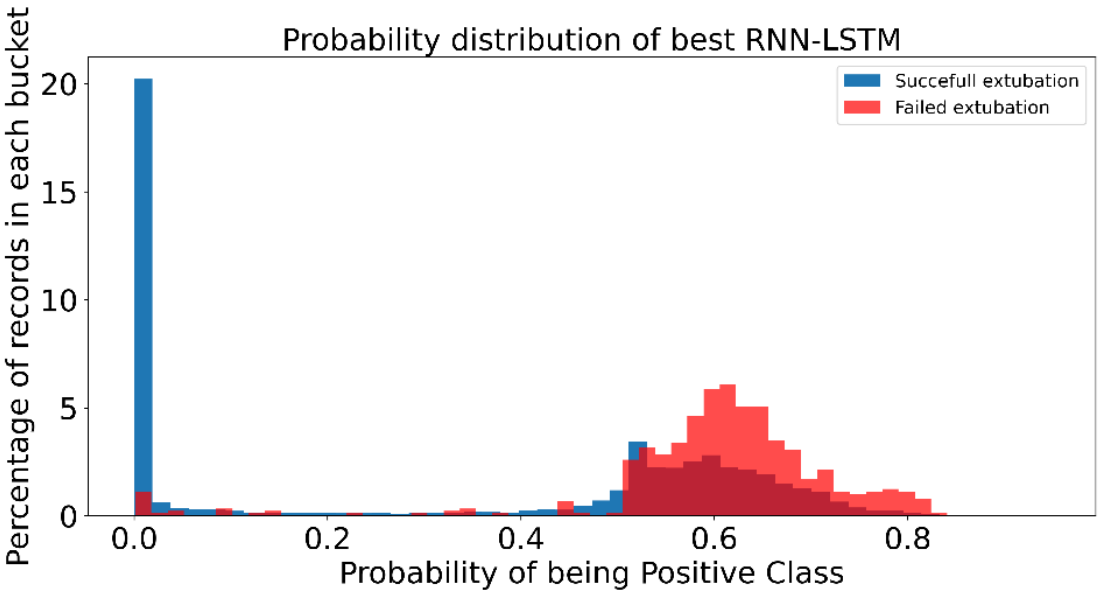


Figure 11: Probability distribution for the base RNN-LSTM model on the CHD cohort. Probabilities shown as percentage of the total group size for both successful extubation and failed extubation.

Tables

Table 1: Risk score table based on the diagnosis/ surgical procedure

High risk (score =2)	Medium risk (score=1)	Low risk (score=0)
Norwood/ Single Ventricle	Transposition of the great arteries, corrected using arterial switch operation.	Atrial/ ventricle septal defect (ASD/ VSD)
Biventriculair repair	Correction Tetralogy of Fallot	Partial Cavo Pulmonary Connection (PCPC)
Neonatal arch recontruction	Atrial-Ventral Septal Defect correction	Total cavopulmonary connection (TCPC)
Neonatale valvereconstruction (i.e. Ross-konno procdure or mitral valve reconstruction)	Homograft replacement	Coarctation of the aorta
Truncus arteriosus	Arch reconstruction	

Table 2: The demographics and distribution of the congenital heart disease cohort. Used features and their units are stated on the left. Mean and standard deviation (SD) are shown for each feature. The mean/ SD were calculated over the overall population, the group without reintubation (group 0) and the group with reintubation (group 1). The p-value displayed for the continuous numerical variables was calculated using the two-sample t-test, for the categorical value the chi-squared test was used.

		Extubation			
		Overall	Success	Failure	P-Value
n (%)		2149 (100)	2022 (94.1)	127 (5.9)	
Age (years), mean (SD)		2.9 (4.5)	3.0 (4.6)	1.1 (3.0)	<0.001
Diagnosis, n (%)	0	826 (38.4)	794 (39.3)	32 (25.2)	<0.001
	1	1194 (55.6)	1115 (55.1)	79 (62.2)	
	2	129 (6.0)	113 (5.6)	16 (12.6)	
Weight (kgs), mean (SD)		13.0 (12.3)	13.4 (12.5)	7.4 (7.6)	<0.001
End tidal CO2 (mmHg), mean (SD)		40.3 (6.0)	40.4 (6.0)	39.2 (5.9)	0.028
Heart rate (bpm), mean (SD)		124.8 (23.1)	124.0 (23.1)	138.1 (18.9)	<0.001
Invasive blood pressure (mmHg), mean (SD)		64.8 (13.8)	65.0 (13.9)	60.8 (10.7)	<0.001
Respiratory rate (freq/min), mean (SD)		29.1 (8.6)	28.7 (8.6)	34.4 (7.2)	<0.001
Saturation (%), mean (SD)		95.7 (7.2)	95.8 (7.2)	94.2 (7.5)	0.023
Inspiratory oxygen (%), mean (SD)		33.9 (9.5)	33.9 (9.4)	34.4 (11.0)	0.630
Ventilator PEEP (mmHg), mean (SD)		5.0 (0.8)	5.0 (0.8)	5.3 (1.0)	0.003
Ventilator peak pressure (mmHg), mean (SD)		15.7 (3.5)	15.7 (3.5)	16.2 (3.2)	0.095
Ventilator respiratory rate (freq/min), mean (SD)		26.6 (14.7)	26.2 (14.7)	33.2 (13.8)	<0.001
Ventilator expiratory tidal volume (mL), mean (SD)		95.1 (107.2)	97.9 (108.5)	50.7 (69.0)	<0.001
Time gap (min), mean (SD)		466.6 (292.8)	457.7 (295.4)	608.5 (201.4)	<0.001

Table 3: The demographics and distribution of the bronchiolitis cohort. Used features and their units are stated on the left. Mean and standard deviation (SD) are shown for each feature. The mean/ SD were calculated over the overall population, the group without reintubation (group 0) and the group with reintubation (group 1). The p-value displayed for the continuous data was the two-sample t-test.

	Extubation			
	Overall	**Success**	**Failure**	**P-Value**
n (%)	327 (100)	291 (89.0)	36 (11.0)	
Age (years), mean (SD)	0.3 (0.5)	0.3 (0.5)	0.4 (0.5)	0.407
Weight (kgs), mean (SD)	5.7 (2.5)	5.6 (2.4)	6.3 (2.7)	0.169
Respiratory rate, mean (SD)	39.1 (6.9)	39.1 (7.0)	39.2 (6.6)	0.990
Heart rate, mean (SD)	140.5 (16.9)	141.2 (16.7)	135.1 (17.2)	0.049
Saturation (%), mean (SD)	96.8 (3.4)	96.7 (3.5)	96.8 (1.7)	0.940
End tidal CO2 (mmHg), mean (SD)	40.9 (6.1)	41.0 (6.1)	39.9 (6.1)	0.308
Inspiratory oxygen (%), mean (SD)	35.4 (10.2)	35.1 (10.3)	37.6 (8.9)	0.124
Ventilator peak pressure (mmHg), mean (SD)	16.2 (5.3)	15.9 (5.2)	18.8 (5.1)	0.002
Ventilator PEEP (mmHg), mean (SD)	5.6 (1.5)	5.6 (1.4)	5.8 (2.0)	0.579
Invasive blood pressure (mmHg), mean (SD)	66.8 (9.6)	66.5 (9.0)	69.0 (13.3)	0.294
Ventilator respiratory rate, mean (SD)	38.9 (12.7)	39.0 (12.8)	37.7 (12.0)	0.543
Ventilator expiratory tidal volume (mL), mean (SD)	34.1 (24.8)	33.6 (24.9)	38.2 (23.7)	0.282

Table 4: Used features in the model after feature selection with the corresponding units. A distinction is made between the static (time invariant) and time variant features (features which change over time)

All features
Static features (time invariant)
Age (years)
Weight (kgs)
Diagnosis (risk score)
Dynamic features (time variant)
End tidal CO2 (mmHg)
Heart rate (bpm)
Invasive blood pressure (mmHg)
Respiratory rate (freq/min)
Saturation (%)
Inspiratory oxygen (%)
Ventilator PEEP (mmHg)
Ventilator peak pressure (mmHg)
Ventilator respiratory rate (freq/min)
Ventilator expiratory tidal volume (mL)
Time gap (min)

Table 5: Precision, recall and f1-score calculated for the base RNN-LSTM model on the congenital heart disease cohort.

	Precision	Recall	F1-score
Successful extubation	0.99	0.50	0.67
Failed extubation	0.09	0.94	0.17

Appendix A

Table 10: Parameters in PICURED database.

Origin of variable	Name	Explanation	Kind of Variable	Data Type
Laboratory	lab_bg_be	Blood gas Base Excess	Continuous	Float
Laboratory	lab_bg_hco3	Blood gas HCO3	Continuous	Float
Laboratory	lab_bg_mode	Laboratory mode	Continuous	Float
Laboratory	lab_bg_origin	Origin of blood sample	Continuous	Float
Laboratory	lab_bg_pco2	Blood gas pCO2	Continuous	Float
Laboratory	lab_bg_ph	Blood gas pH	Continuous	Float
Laboratory	lab_bg_po2	Blood gas pO2	Continuous	Float
Laboratory	lab_bg_sat	Blood gas Saturation	Continuous	Float
Laboratory	lab_bl_b2m	Beta-2-microglobulin (in blood)	Continuous	Float
Laboratory	lab_bl_bil_d	Direct bilirubin (in blood)	Continuous	Float
Laboratory	lab_bl_bil_i	Indirect bilirubin (in blood)	Continuous	Float
Laboratory	lab_bl_ca2	Ionized calcium (in blood)	Continuous	Float
Laboratory	lab_bl_catot	Total calcium (in blood)	Continuous	Float
Laboratory	lab_bl_cc	Cystatin C (in blood)	Continuous	Float
Laboratory	lab_bl_cl	Chloride (in blood)	Continuous	Float
Laboratory	lab_bl_cr	Creatinin (in blood)	Continuous	Float
Laboratory	lab_bl_CRP	C-reactive protein (in blood)	Continuous	Float
Laboratory	lab_bl_f	Phosphate (in blood)	Continuous	Float
Laboratory	lab_bl_gluc	Glucose (in blood)	Continuous	Float
Laboratory	lab_bl_hb	Hemoglobulin (in blood)	Continuous	Float
Laboratory	lab_bl_ht	Haematocrit (in blood)	Continuous	Float
Laboratory	lab_bl_k	Kalium (in blood)	Continuous	Float
Laboratory	lab_bl_lactate	Lactate (in blood)	Continuous	Float
Laboratory	lab_bl_leuco	Leucocytes (in blood)	Continuous	Float
Laboratory	lab_bl_mg	Magnesium (in blood)	Continuous	Float
Laboratory	lab_bl_na	Natrium (in blood)	Continuous	Float
Laboratory	lab_bl_tr	Thrombocytes (in blood)	Continuous	Float
Laboratory	lab_bl_ur	Ureum (in blood)	Continuous	Float

Table 11: Parameters in PICURED (continued)

Origin of variable	Name	Explanation	Kind of Variable	Data Type
Monitor	mon_etco2	End tidal CO2	Continuous	Float
Monitor	mon_hr	Heart rate	Continuous	Float
Monitor	mon_ibp_dia	Invasive diastolic blood pressure	Continuous	Float
Monitor	mon_ibp_mean	Invasive mean blood pressure	Continuous	Float
Monitor	mon_ibp_sys	Invasive systolic blood pressure	Continuous	Float
Monitor	mon_nibp_dia	Non-invasive diastolic blood pressure	Continuous	Float
Monitor	mon_nibp_mean	Non-invasive mean blood pressure	Continuous	Float
Monitor	mon_nibp_sys	Non-invasive systolic blood pressure	Continuous	Float
Monitor	mon_rr	Respiratory rate	Continuous	Float
Monitor	mon_sat	Oxygen saturation	Continuous	Float
Monitor	mon_temp	Body temperature	Continuous	Float
Monitor	mon_temp_mode	Body temperature mode	Nominal	Object
Monitor	mon_temp_skin	Skin temperature	Continuous	Float
Observation	obs_pup_dia	Pupillary diameter	Continuous	Float
Observation	obs_pup_light	Pupillary light response	Binary	Object
Patient	pat_adm_start	Start date of admission	Date	DateTime
Patient	pat_datetime	Timestamp of observations	Date	DateTime
Patient	pat_hosp_id	Unique hospital ID	Discrete	Int
Ventilator	vent_cat	Patient category on ventilator	Nominal	Object
Ventilator	vent_fio2	Percentage of oxygen therapy	Continuous	Float
Ventilator	vent_fio2_flow	Flow	Continuous	Float
Ventilator	vent_fio2_mod	Modality oxygen therapy	Nominal	Object
Ventilator	vent_m_fio2	Oxygen therapy (measured)	Continuous	Float
Ventilator	vent_m_no	NO therapy (measured)	Continuous	Float
Ventilator	vent_m_peep	PEEP (measured)	Continuous	Float
Ventilator	vent_m_ppeak	Peak pressure (measured)	Continuous	Float
Ventilator	vent_m_pplat	Plateau pressure (measured)	Continuous	Float
Ventilator	vent_m_rr	Respiratory rate (measured)	Continuous	Float
Ventilator	vent_m_tv_exp	Expiratory Tidal Volume (measured)	Continuous	Float
Ventilator	vent_m_tv_insp	Inspiratory Tidal Volume (measured)	Continuous	Float
Ventilator	vent_machine	Machine type	Nominal	Object
Ventilator	vent_mode	Ventilator mode	Nominal	Object
Ventilator	vent_tube	Tube size	Continuous	Float

Appendix B

Patient trajectories for different parameters

Patient trajectories for all parameters used in the study. Last 12 hours on the ventilator are displayed for the successful extubation group, and 12 hours before first attempt are shown for the failed extubation group.

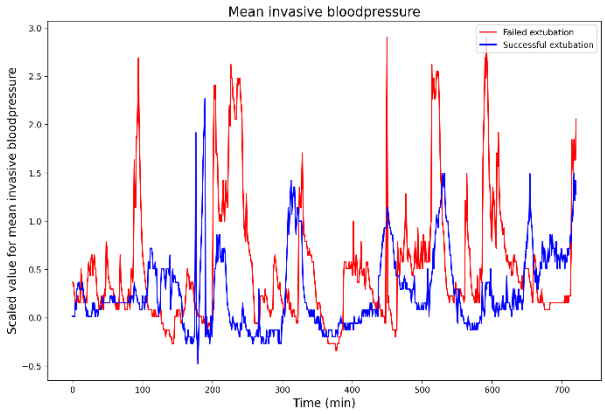


Figure 12: Patient trajectories for the mean invasive blood pressure. (min = minutes)

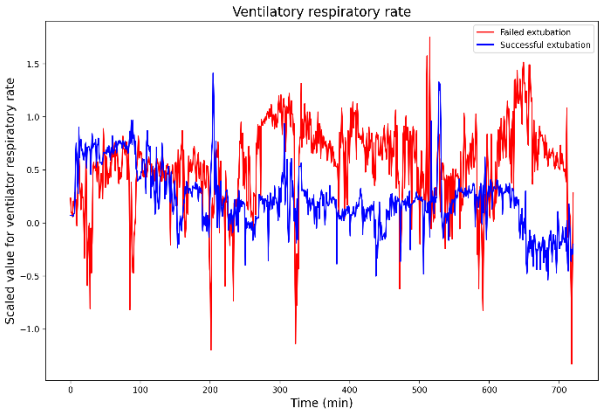


Figure 13: Patient trajectories for the ventilator respiratory rate. (min = minutes)

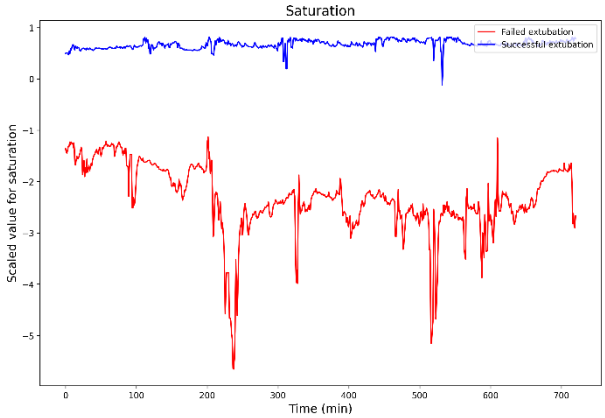


Figure 14: Patient trajectories for the saturation. (min = minutes)

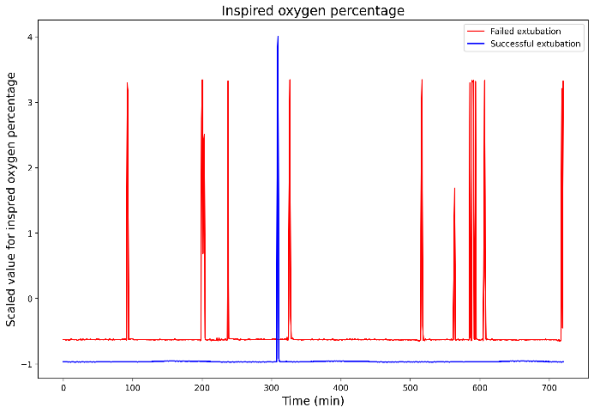


Figure 15: Patient trajectories for the inspired oxygen percentage (min = minutes)

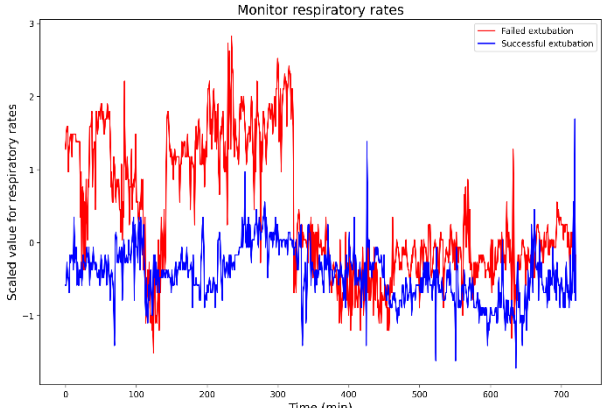


Figure 16: Patient trajectories for the monitor respiratory rates (min= minutes)

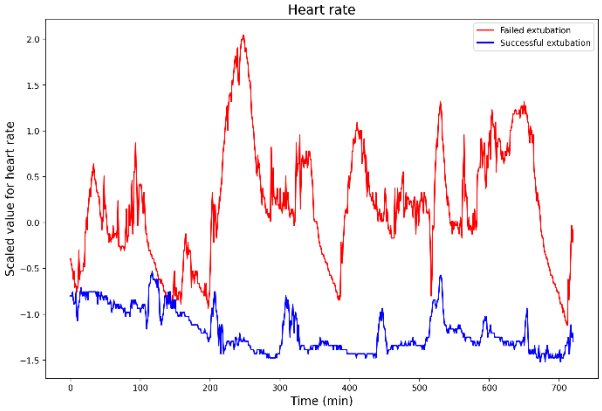


Figure 17: Patient trajectories for the heart rate (min= minutes)

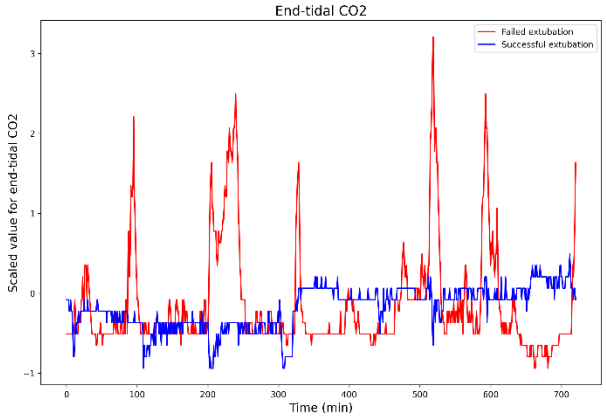


Figure 18: Patient trajectories for the end-tidal CO2. (min= minutes)

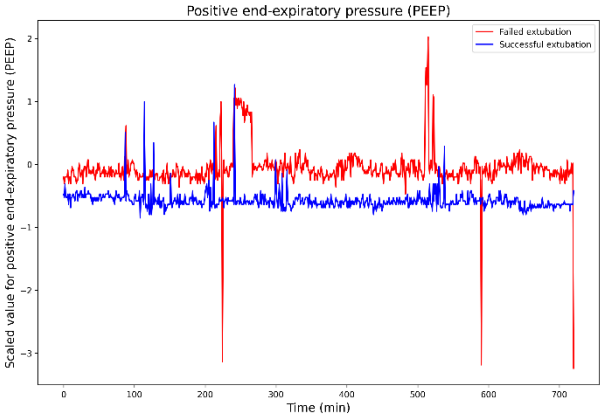


Figure 12: Patient trajectories for the positive end-expiratory pressure. (min= minutes)

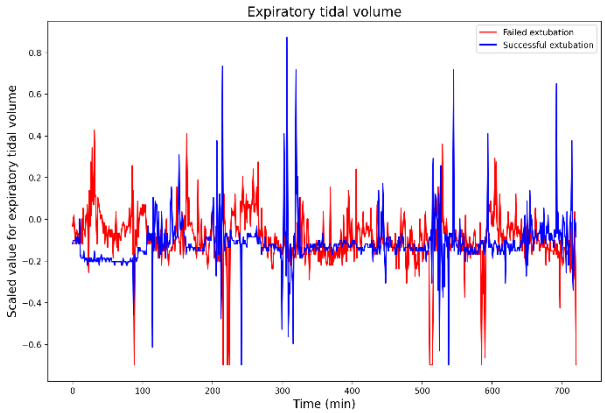


Figure 20: Patient trajectories for the expiratory tidal volume. (min= minutes)

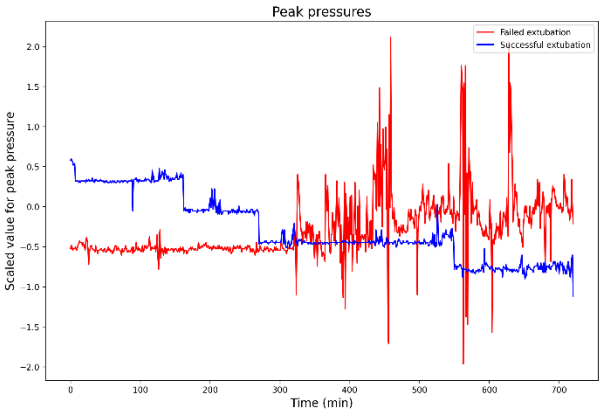


Figure 21: Patient trajectories for the peak pressures of the ventilator (min= minutes)