Predicting Haemodynamic Instability in Critical Congenital Heart Disease Patients: A Proof of Concept

B. van Winden

TU Delft

Erasmus MC
Sophia Children's Hospital

Predicting Haemodynamic Instability in Critical Congenital Heart Disease Patients: A Proof of Concept

Brian van Winden

Student number: 4677277

19th February 2024

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)
Dept. of Paediatric Intensive Care Unit,
Sophia Children's Erasmus MC
19th June 2023 - 5th March 2024

Supervisors:
dr. R.C.J. de Jonge, MD, PhD, Erasmus MC
dr. D.M.J. Tax, Assistant professor, TU Delft
dr. J.W. Kuiper, MD, PhD, Erasmus MC
E. van Twist, MSc, Erasmus MC

Thesis committee members:
dr. R.C.J. de Jonge, MD, PhD, Erasmus MC
dr. D.M.J. Tax, Assistant professor, TU Delft
dr. J. Nijman, MD, PhD, UMC Utrecht
dr. J.W. Kuiper, MD, PhD, Erasmus MC

Contents

1 Abstract

Introduction

Approximately 9 in 1000 children are born with congenital heart disease (CHD), of whom a quarter are classified as critical CHD (CCHD) and require an intervention within their first year. Monitoring these patients in the Paediatric Intensive Care Unit (PICU) is crucial, yet with increasing amounts of data, detecting subtle changes that are important for the disease progression interpretation of all vital signs becomes difficult, even for skilled physicians. Machine learning (ML) offers potential solutions, however, challenges such as inter-patient variability and the absence of clear definitions for haemodynamic instability persist. This study aims to develop a ML algorithm for early prediction of haemodynamic instability in CCHD patients with high frequency vital-signs, addressing these challenges through objective labelling methods and stratification approaches.

Methods

Two approaches, on population and patient level, were developed with nested cross-validation (CV). Due to a high inter-patient variability, the patient specific approach was added.

A first iteration of objectively labelling haemodynamic instability was proposed, based on medical interventions such as medication administration and fluid therapy. Since it is difficult to retrospectively determine for how long patients were unstable, multiple values for instability duration (dT) were added to the analysis.

To capture the temporal dependency of time-series data, lag-analysis was performed, adding the relation between the vital signs and their previous values to the model development. Lag-analysis included a sliding window that moved over the data. The width of sliding window (W) was optimised during the model development. Additionally, a horizon (r) was implemented, so the data within the sliding window were predicting future timestamps.

Results

This retrospective study included a total of 224 admissions in the analysis. Two random forest classifiers were trained using a nested CV structure to detect haemodynamic instability in CCHD patients. For both approaches the same temporal settings (W: 50 minutes, r: 45 minutes, dT: 120 minutes) were used. This study has shown that the between-patient approach had notable differences between the mean train (85%, AUCPR) and test performance (40%, AUCPR).

The in-patient approach, while using 20% and 10% of the test data for training, still yielded a test performance of 96% (AUCPR) and 90% (AUCPR), respectively.

Discussion and conclusion

Generally speaking, the experiments suggest that the first iterations of the models were not robust and generalised poorly. It is most likely caused by a large inter-patient variability and a simple labelling system that is still depending on subjectivity.

This study has shown that the proposed prediction model, which combines high frequency vital signs, labels, and temporal settings (W, r, dT), requires additional refinement before it can be considered clinically feasible to implement this model as a reliable bedside tool for predicting haemodynamic instability.

2 Introduction

2.1 Background

Approximately 9 in 1000 children are born with a form of congenital heart disease (CHD) (1, 2). Of all infants born with CHD, approximately 25% is categorised as critical congenital heart disease (CCHD). Children with CCHD require a surgical intervention or heart catheter-based intervention within their first year of life (3).
Infants often need to undergo a transitional period before intervention, to grow and mature enhancing the likelihood of a successful intervention. Since these patients can rapidly become haemodynamically unstable, risking pre-surgical death, patients are monitored in the Paediatric Intensive Care Unit (PICU) (4, 5).

During monitoring, vital parameters are continuously measured with high frequency to closely track the physiological status of a perioperative patient. Currently, all vital signs, such as heart rate (HR), respiration rate (RR), oxygen saturation (SpO_2), and blood pressure (BP) are measured independently. This process where multiple parameters are measured and presented together, is called multi-modal monitoring. Multi-modal monitoring does not include tools that emphasize the correlation within the presented data, making interpretation of this information difficult, even for skilled physicians (6). Subtle changes in data may be overlooked among the high-frequent updates across screens that are displaying numerous of vital signs. These subtle changes, however, may be crucial for the disease progression. By using multi-modal monitoring, the physician's workload is increased, creating the need for real-time data processing (7, 8, 9).

To unburden the physicians of this workload, real-time data analyses can play an important role. Improvement of technology and the expanding volumes of collected data enhances the feasibility of machine learning (ML) or Artificial Intelligence (AI) approaches for real-time applications, including health care orientated applications (10). Unlike unsupervised learning, supervised learning is a ML technique which is more often implemented in healthcare (11). Supervised learning uses statistical techniques that can be programmed to classify or predict future examples, using the correlation between input and pre-defined output of interest (12).
With ML, multi-modal monitoring data can be used to develop detection algorithms of critical events, give context to the interpretation of the parameters and can be used to support decision making. As such, an accurate ML model can improve the workflow of the clinician by predicting aberrations of the patient before becoming clinically evident, resulting in early diagnosis or timely therapeutic treatment.

The clinical status that is closely monitored in CCHD patients is haemodynamic instability. Haemodynamic instability can be a result of wide variety of physiological presentations in the CCHD population. These differences introduce inter-patient variability posing difficulty for generalisation of the database, requiring additional attention in model development. In the literature, there is no universally accepted definition, utilizing objective measures, of haemodynamic instability. Without a clear definition, making automated and objective labels used in creating prediction or classification models is a true challenge.

When conducting a search in literature for algorithms for paediatric CHD patients, multiple algorithms have been conceptualised, although with different objectives. When comparing studies developing a prediction algorithm for cardiac patients, no general approach can be deduced. Different types of classifiers are used varying from simple linear discriminant analyses to complex ensembled methods of deep-learning models (13, 14).
Zoodsma et al. aimed to develop a diagnostic model using supervised ML to continuously detect clinical deterioration in CCHD patients admitted to the PICU (15).

Their model consists of a one-class support vector machine (OCSVM) using high-frequent physiological parameters, i.e., HR, RR, SpO_2, BP and regional cerebral saturation (rSO2). Utilizing solely high frequency vital signs, this model stands out as one of the few, if not the only one, documented in the literature for updating its real-time predictions for haemodynamic instability with the same sample rate as all input data. The model is evaluated using retrospective validation of two individual experts, lacking automation and objectivity of labels. The true positive rate of the unstable class (77%) and true positive rate of the stable class (93%) were calculated using the hours of data that was labelled correctly. Since, this evaluation is quite subjective, the statistical assessment is hard to reproduce and might not be reliable.

2.2 Goals

This study aims to develop a ML algorithm to predict haemodynamic instability in CCHD patients admitted to the PICU before physicians are capable of detecting the clinical deterioration. The goal includes pre-processing the retrospective data and creating an objective method for labelling haemodynamic instability.

Two stratification methods for the input data, called the between-patient and in-patient approach, will be explored and assessed as means to address the inter-patient variability. The best performing settings for the ML model including three parameters for the temporal aspect of the input data and the assigned labels are analysed.

3 Methods

3.1 Pipeline

Figure 3.1 gives an overview of the four sections needed for a successful model development. Two approaches can be recognised in the pipeline, one based on a population level and one on a patient level. In this thesis, these approaches are called the *between-patient* and *in-patient* approach, respectively.

The between-patient approach was developed with the goal to create one robust model that fits all patients. By using data of entire admissions of all patients within the train or test set, the model should not learn to recognise specific patients. In practise, this model could be used to just predict haemodynamic instability in newly admitted patients without any additional steps, improving implementation feasibility in a health care setting.

The in-patient approach was added to explore the prediction capabilities of the model when specifically trained within one patient, removing the inter-patient variability. Practically, this could mean that, when a patient is admitted to the PICU, the first few hours of data are used to

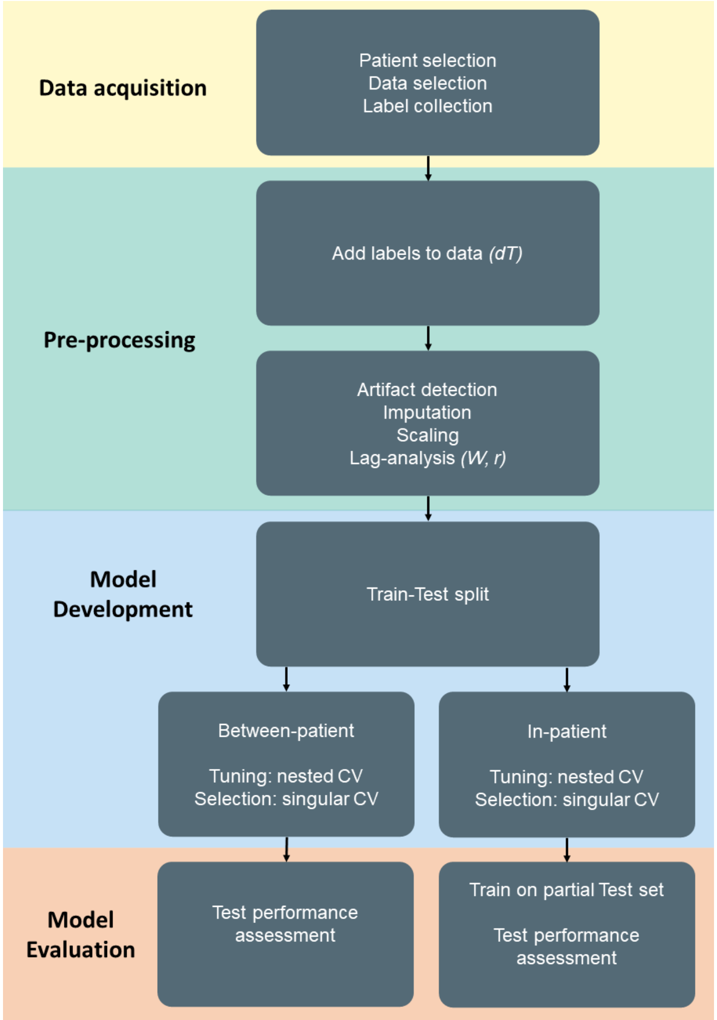


Figure 3.1: An overview of the pipeline that exists of four stages. Each stage is covered in the method section.
CV: cross-validation, dT: duration of labels haemodynamic instability, W: window-width, r: horizon

train the model. The remaining data collected during the admission could be used for the model to predict haemodynamic instability.
Importantly, to make the in-patient model clinically feasible, a minimum training time (and thus a minimum amount of training data) is desired to use the predicting capabilities as soon as possible. Therefore, an analysis for the train-test-split was added to find an equilibrium between the minimum required trainset size and prediction performance.

Both approaches share most of the sections from Figure 3.1, and are explained below.

3.2 Data Acquisition

3.2.1 Study Cohort

A retrospective study was conducted in patients with critical congenital heart disease (CCHD) admitted to the PICU at the Erasmus MC Rotterdam between January 2016 and April 2023. All subjects with at least 5 unique numeric measurements for each of the included parameters (HR, RR, SpO2, CVP and MAP) during the entire admission were eligible for inclusion. Exclusion criteria were an admission length less than 12 hours or a birth weight below 2500g. The exclusion criterion for birth weight was applied as a substitution to exclude pre-term infants, since the gestational age is not structurally recorded in our EHR. The gestational age or age of the patient in general, might influence the vital signs of the patient due to physiology immaturity (16). Once full term, the chance of the birth weight being below 2500 grams is estimated to be 2.5%, based on the mean and standard deviation (SD) provided by Chen et al. (17).

3.2.2 Data

All patients underwent continuous monitoring of vital signs (1 Hz) with data storage on a secure server (Dräger, Lübeck, Germany). The HR was calculated using the R-peak interval measured with at least three leads of the electrocardiogram (ECG, 3M St. Paul, Minnesota, United States), RR was measured using the impedance of the ECG electrodes, the other measurements recorded are the post-ductal SpO2 (Masimo, Irvine, California, United States), MAP (Becton Dickinson, Franklin Lakes, New Jersey, United States) and CVP (Becton Dickinson, Franklin Lakes, New Jersey, United States).
When the oxygen delivery to the main organs is failing, the body will activate mechanisms, such as increasing the cardiac output by increasing the heart rate, to compensate (18). A change in vital signs is expected when a patient is clinically deteriorating. The aforementioned parameters were selected based on the method of *Zoodsma et al.* (15). Due to availability, the cerebral regional oxygen saturation (rSO2) was replaced by the CVP, which with context of other parameters could be an indirect representation of the venous return (19).

To be able to predict haemodynamic instability in real-time, the input parameters should be frequently updated. When parameters that are recorded once a day are included, the type of model developed in this thesis is dependent on these variables that are not 'high-frequently' obtained, reducing the resolution of the prediction. During the literature review, continuous data (and prediction) was defined as: *Data must be gathered with at least 1 measurement per minute (1/60 Hz) AND the recording length must be at least 1 hour.* This definition is also used in this thesis, resulting in resampling to 1 measurement per minute.

Ethical approval has been granted for retrospective research contributing to dashboarding and ML model development using the continuously recorded vital signs that are obtained during routine care.
The data are accessible using a Digital Research Environment (myDRE, ANDREA B.V.), which is complying to safety norm ISO 27001 for storing patient data (20). Folder- and filenames of the

raw data stored in this location include the patient ID, which are pseudonymised using a study ID.

3.2.3 Labels

To be able to develop an algorithm to detect clinical deterioration, data must be labelled as haemodynamic stable or unstable. Due to lack of a clear definition in literature and for feasibility reasons, a straightforward labelling-system was created. Physicians were consulted to derive labels based on therapeutic interventions that are common in (increasingly) haemodynamically unstable patients and that are structurally registered in the electronic health record (EHR). Assuming physicians are acting correct and in the best interest of the patient, the interventions and their administration times were retrospectively collected and used to create windows of instability. Interventions performed in treating haemodynamic instability at the PICU in the Erasmus MC include inotropic and vasopressive medication administrations and/or fluid therapy. The following interventions including the used substance were retrieved from the EHR:

Table 3.1: The administration (intervention) types and their medication used to create the periods of haemodynamic instability.

Administrations	**Medicine/substance**
Inotropes	Epinephrine, milrinone, dobutamine, dopamine, phenylephrine, isoprenaline
Vasopressors	Vasopressin, desmopressin, norepinephrine
Pulmonary vasodilators	Nitric Oxide (NO), sildenafil, bosentan
Fluid challenges	NaCl, ringer's lactate, blood transfusion

Based on clinical experience alone, it is difficult to retrospectively define how long a patient was already haemodynamically deteriorating before the interventions mentioned in Table 3.1 were performed. Several options for the different lengths of instability before an intervention (dT) were included in developing the algorithm, to find whether a specific duration of instability exists where the model was better capable of predicting the instability.

The following dT durations were used in the analysis: 20, 60 and 120 minutes before administrating medication or fluid therapy. In Figure 3.2, a visual representation of the use of the dT parameter is added.

By introducing this method, in essence, three different datasets were created with different class distributions. Less timestamps are labelled as unstable in the *dT: 20 minutes* dataset than the *dT: 120 minutes* dataset. Therefore, the class priors (class distribution) were calculated to be able to give more context to the model evaluation between models using different settings for dT.

3.3 Pre-processing

The methods used for resampling, artefact detection, scaling, and imputation are elaborated in Appendix A.1.

3.3.1 Lag-analysis

For capturing all possible information in time-series data, the temporal dependency was incorporated. The method employed was lag analysis, which involves examining the impact of time delays, referred to as lags, of a variable on itself within a time-series dataset. The goal is to understand how past values of a variable influence its current or future values.

To get an indication of the importance of lags, the cumulative explained variance of a principal component analysis (PCA) was collected. By dividing the data in sections of 50 minutes, the number of components needed to explain 95% of the variance in the data was used to give a

rough estimate of the importance of the time lags.
Essentially, this method similar to a sliding window, as illustrated in Figure 3.2. When applying this sliding window, several settings are available to fine-tune the lags according to the algorithm's objectives:

- W: The window width, equals to the number of time stamps included in the lag-analysis. During algorithm development, widths of 5, 30, and 50 minutes were included.
- r: The horizon will determine how many time stamps in the feature the model should predict, based on the data within the window. During algorithm development, widths of 1, 5, 15, 30, 45, and 60 minutes were included.

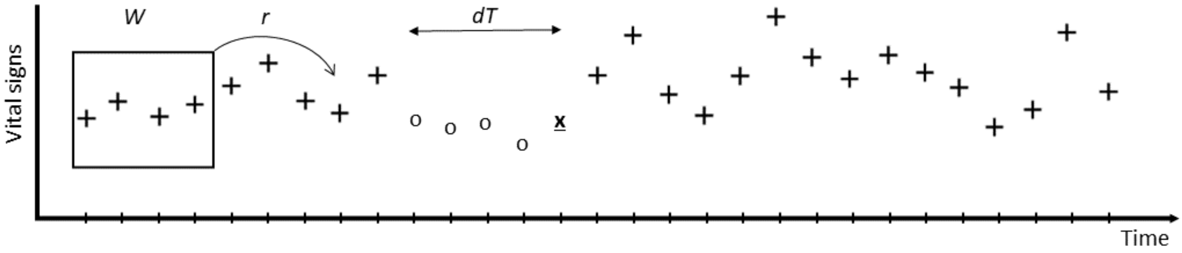


Figure 3.2: This figure illustrates the Lag-Analysis and all its settings. In this example, one fictional vital sign is shown, using a *'+'* symbol for the haemodynamically stable timestamps, a *'x'* symbol for the timestamps with an intervention and an *'o'* symbol for the timestamps labelled as haemodynamic instability.
The total number of timestamps, before and during an intervention, labelled as haemodynamically unstable is decided by the dT setting, which is set to 5 timestamps (or minutes) in this example.
In this example, the window-width (W) is set to 4 timestamps (or minutes), which is the input data for the model to predict the timestamp after the horizon(r), which is set to 4 timestamps (or minutes).

Horizon analysis

An additional analysis was performed to further explore the effect of increasing the horizon further than 60 minutes. The in-patient approach was used for this analysis, with a W of 30 minutes and a dT of 60 minutes. The horizon settings used were: 1, 5, 15, 30, 45, 60, 120, 240, 360, 480 minutes. The number of settings were limited to keep the computational time manageable.

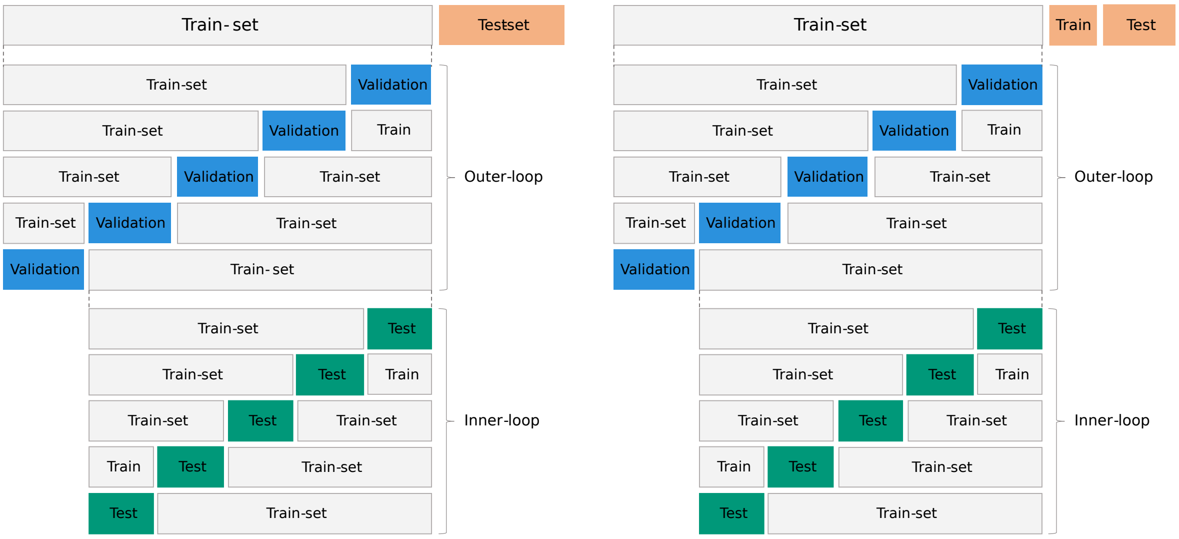
3.4 Model development

3.4.1 Cross-validation structure

In Figure 3.3, an overview of the two CV structures is illustrated.

This independent test set (orange in Figure 3.3) was used for the final performance of the model, using data that had not been included in the development of the model in both approaches. The distribution of patients was 80% and 20% for the train and test set, respectively. The train-test-split should also distribute the total number of minutes of data according to the specified split. Since admission lengths between patients vary, a difference of 5% with the specified split was deemed acceptable.

Zoodsma et al. suggest differences between cyanotic and non-cyanotic CCHD (SpO_2 <90% and >90%, respectively) are influencing the algorithms performance (15). This was accounted for by stratifying these subjects over the train and test set.



(a) Nested CV for the between-patient approach.

(b) Nested CV for the in-patient approach.

Figure 3.3: Visual overview of the nested cross-validation (CV) structures. The inner-loop was used for hyper-parameter tuning. The outer-loop was used for a performance assessment of the best performing classifier from the inner-loop.
The in-patient approach needs to be trained first on a new subset of patient data, therefore, the final test set was again split in a train and test section.

Traditional CV involves splitting the dataset into train and test sets, fitting the model on the train set, and evaluating its performance on the test set. This process was repeated multiple times to ensure robustness of the performance estimate. However, when hyper-parameters need to be tuned, traditional cross-validation may lead to an overestimation of performance due to the risk of overfitting to the test set. Nested CV addresses this issue by introducing an outer-loop and an inner-loop. The outer-loop used in these models split the dataset into five train and validation sets (blue in Figure 3.3, similar to traditional cross-validation. Within each iteration of the outer-loop, an inner-loop of cross-validation was performed on the train set to select the best hyper-parameters for the model. To find the best performing hyper-parameters, twenty unique combinations of hyper-parameters are randomly selected, using a random grid search from SkLearn(21). The best performing model was selected, using the mean performance of all test sets in the inner-loop (green in Figure 3.3). This model was then trained on the entire train set of the outer-fold and was evaluated on the corresponding validation set (blue in Figure 3.3). This process was repeated for each iteration of the outer-loop.

Ideally, when a clear relation is found in the data, hyper-parameters of the five highest scoring models are comparable. If this is the case, the model is expected to generalise well.

Once generally well performing models were created using the nested CV setup, a singular CV setup was used to develop one final model. This process is comparable to one inner-loop of the nested CV, meaning new sets of hyper-parameters were searched and tested during this CV method. For the between-patient approach, this meant using the whole train set and individual test set (orange in Figure 3.3a) for a final performance assessment of the between-patient approach. For the in-patient approach, the best performing hyper-parameters of the singular CV were used for training on the small train set (orange in Figure 3.3b) and the final performance was assessed using the small test set (also orange in Figure 3.3b, simulating a new group of admitted patients.

3.4.2 Classifier

The main component of a ML algorithm is the classifier that is tuned using the aforementioned nested CV. For this iteration of the model, a random forest classifier was used.

A random forest classifier is an ensembled ML technique, which combines multiple decision trees for a collaborative prediction/classification model. Each decision tree was trained on a subset of the training data. This introduces randomness in the data and the important features, managing against overfitting and improving the generalisation of the model. A decision tree in a random forest grows by splitting the feature space of the input data into subsets based on the values of the features.

By combining multiple decision trees, one classifier was created that can handle non-linear relationships between data and labels.

The classifier is shaped by hyper-parameters, which were tuned during the inner-loop of the nested CV. In Table 3.2 an overview of the used hyper-parameters and a short explanation is illustrated.

Table 3.2: The hyper-parameters of the random forest classifier that were tuned in the nested CV, are briefly explained.

Hyper-parameter	**Explanation**
Class weight	Corrects for class imbalances on a level of all data, or on the level of a sub-set of data used for one decision tree.
Loss criterion:	Determines the formula used to calculate the quality of a new split, the best split is selected for growing a decision tree.
Maximum depth:	Sets the maximum number of consecutive decisions a tree can make before coming to a final classification.
Maximum features:	Sets the number of features from the input data can be considered to use in the decision trees.
Maximum leaf nodes	Limits the maximum number of leaves (in other words, results of the decision tree) on the decision tree, and indirectly determines the maximum number of splits that can be made. This parameter simplifies the model, generally reducing overfitting.
Minimum samples per split	This hyper-parameter prevents overfitting, by setting a minimum of samples that need to be involved in one of the categories when a split is made.
Number of estimators	This hyper-parameter limits the number of decision trees that can be inside the forest. More decision trees usually improve a model's robustness when trained properly using, for example, nested CV.

3.5 Model Evaluation

The model performances were assessed using the area under the precision recall curve (AUCPR score), mostly due to class imbalances. By selecting the AUCPR, the importance of the positive labels (haemodynamically unstable) was emphasised.

While performing the nested cross-validation, the mean AUCPR scores over the inner-loops were used for hyper-parameter selection. Furthermore, the SD was computed for all inner-loops to evaluate the stability of performance across various data splits. For the final performance assessment of the singular on the test set, performance metrics were added to provide a comprehensive evaluation, including AUCPR, AUCROC, accuracy, balanced accuracy, and the F1-score.

4 Results

4.1 Data acquisition

4.1.1 Study cohort

Out of 637 admissions collected, 229 admissions from 218 different patients were included in the analysis. 76 (12%) admission were excluded due to a total admission length of less than 12h. Another 90 (14%) admissions were excluded due to the patient's birth weight being below 2.5 kg. Lastly, 242 (38%) admissions did not have measurements in all vital signs and were also excluded.

The baseline characteristics and the diagnoses of the included patients are available in Table 4.1

Table 4.1: The baseline characteristics of the study cohort, including the diagnoses.

Characteristics	**Value**
Male gender, n (%)	136 (61)
Admission Age (days), median [Q1-Q3]	132 [85-182]
Vital Signs	**Median[Q1-Q3]**
HR, beats per minute	137 [121-175]
RR, breaths per minute	41 [32-54]
SpO2, (%)	97 [94-99]
CVP, mmHg	12 [8-18]
MAP, mmHg	57 [34, 68]
Diagnosis	**No. of patients**
Ventricular septal defect	62
Tetralogy of Fallot	47
Coarcation of Aorta / aortic anomalies	28
Atrioventricular septal defect	25
Atrial septal defect	9
Complete transposition of great vessels	9
Double outlet right ventricle	8
Hypoplastic left heart syndrome	8
Total anomalous pulmonary venous return (TAPVR)	7
Hypoplastic right heart syndrome	1
Truncus arteriosus	4
Right ventricular outflow tract obstruction	3
Others (Ebstein, tricuspid atresia, etc.)	7

4.2 Pre-processing

4.2.1 Labels

When using three different settings for dT, three different data set with different class priors were created. The class distribution increased when an increased dT setting was used. In Table 4.2, the exact percentages of samples labelled as unstable are displayed for each dT setting. These values should be considered when comparing the AUCPR scores between models trained with different dT settings.

dT	Class prior (%)
20	5
60	12
120	18

Table 4.2: Class prior, which might slightly deviate from the values showed here, based on the different window-width (W) sizes.

4.2.2 Lag-analysis

The cumulative explained variance of the PCA is shown in Figure 4.1, to get an indication of the contribution of including time-lags. The data was segmented into non-overlapping windows of 50 minutes (W), thereby eliminating redundancy in the input data provided to the model. With 27 components, approximately 95% of the variance of the dataset can be explained. These cannot be directly translated to number of lags, however, it does indicate that not all variance in the data is determined by 1 or 2 components.

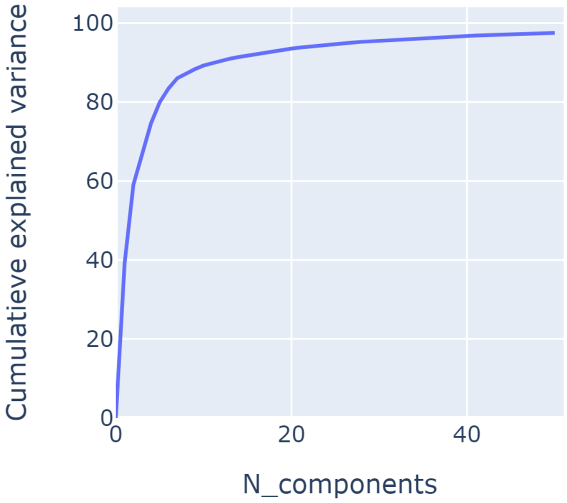


Figure 4.1: Cumulative explained variance from a principal component analysis, with the number of components on the x-axis.

4.3 Model development

4.3.1 Between-patient approach

Hyper-parameter tuning

The model's hyper-parameter tuning was repeated for all combinations of settings for the parameters W, r, and dT.

In Figure 4.2, the training and validation performance (AUCPR) is shown for each of these combinations. Each sub-plot contains three lines, coloured according to the colours of the train, test, and validation set in the CV-structures of Figure 3.3. Only the best performing model of the nested CV is illustrated for each combination of W, r and dT settings. The best performing model was defined as the model with the highest validation score, also shown with the blue line. The SD of the train (grey) and test performance (green) of the inner-loop are added as an indication of the stability of the selected model.

When exploring the results of Figure 4.2, it is noticeable that the performance gradually

improves with increasing dT parameters.
When comparing different window-widths (W), the performance on the train set seems to gradually improve with increasing window-width, however, the test performance generally stays the same.
The performance of the validation set (blue) appears to gradually decrease as the horizon parameter increases, regardless of the W and dT settings. The performance of the train and test set do not show a clear change with increased settings for the horizon.

Based on this figure, the settings for W, r and dT were selected to train one final classifier for the between-patient approach. Firstly, W was selected. The mean test score and validation score were generally speaking more uniform in the W = 50 minutes apart from the first column (dT = 20 minutes). When the optimal dT setting was selected, a uniform increase was seen in the performance with increased values for dT, regardless of W or r, therefore 120 minutes was selected. Finally, the optimal r was selected. Since the model seemed to overfit on the train set in some parameter combinations with a horizon of 60 minutes, the horizon was set to 45 minutes, still maintaining the medical benefits of a large horizon.

The results are shown in Table 4.3.

Table 4.3: The selected temporal settings for training the final model for both approaches.

Parameter	W	dT	r
Between-patients	50	120	45
In-patients	50	120	45

Final performance

When training on the entire train set, the best performing hyper-parameters were selected by the singular CV and are available in Appendix A.1. The model was then evaluated using the performance on the test set (orange in Figure 3.3a).

In Table 4.4 the final evaluation of the performance of between-patient approach is shown. Since a dT setting of 120 minutes was selected, the class prior was 18%, which would be de AUCPR score if the model was fully dependent on chance.

Table 4.4: Final performance of the between-patient approach. The training performance are the mean train and validation scores of the singular CV. The testing performance was obtained using the test set (orange in Figure 3.3a).

Training performance	
Train AUCPR (mean (SD))	0.852 (0.031)
Validation AUCPR (mean (SD))	0.450 (0.126)
Testing performance	
AUCPR	0.402
AUCROC	0.714
Accuracy	0.745
Balanced accuracy	0.317
F1-score	0.410

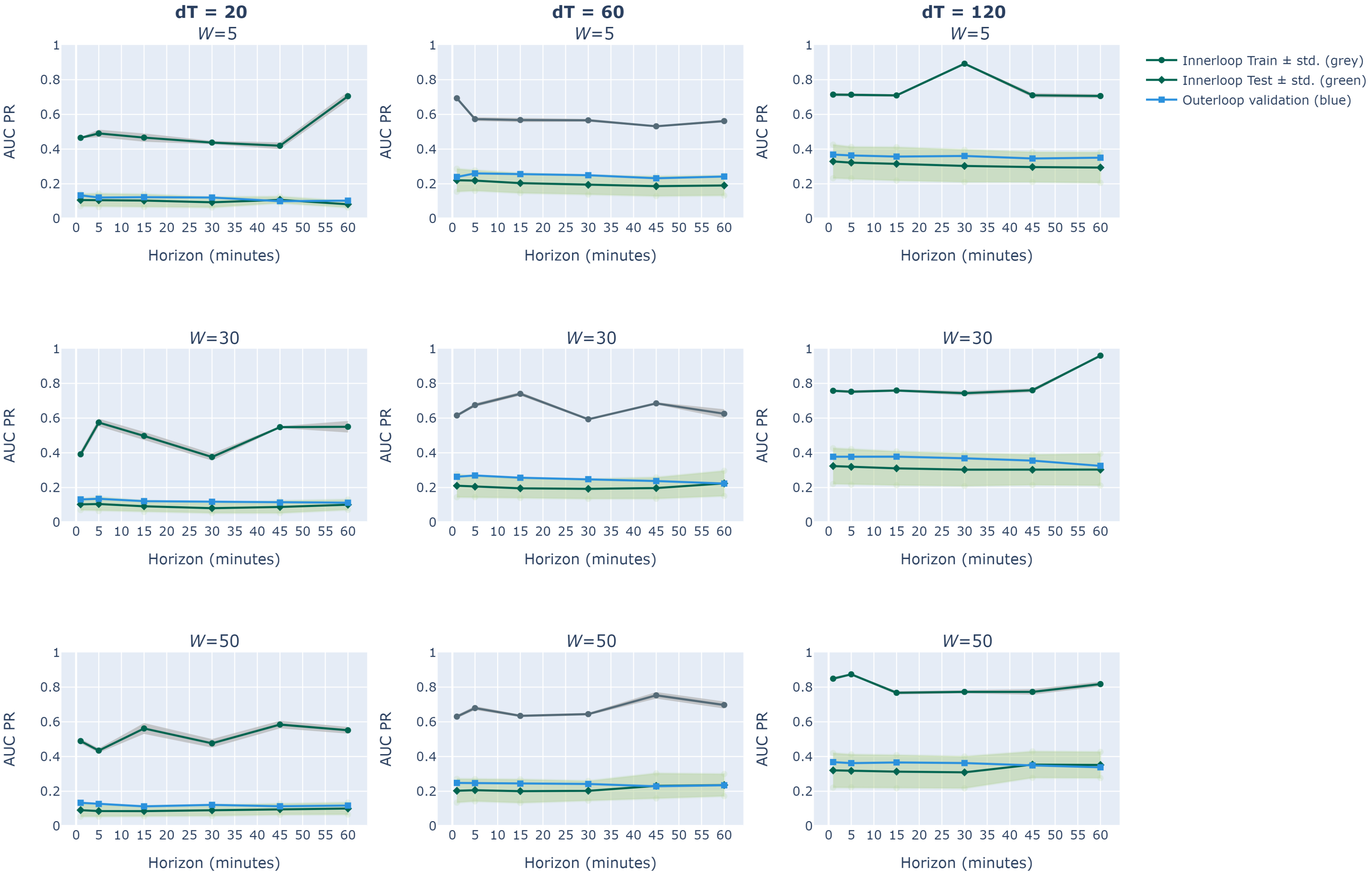


Figure 4.2: Training performance of the between-patient approach and additional analysis for the window-width (W), duration of instability (dT), and the horizon (r). Each column represents another dT setting and each row represents a different W setting.

In each sub-figure, three lines are plotted. The grey line represents the mean (SD) of the inner-loop training performance. The green line represents the mean (SD) of the inner-loop test performance. Lastly, the blue line represents the validation performance. All colours are synchronised with the CV-structure in Figure 3.3

4.3.2 In-patient approach

Hyper-parameter tuning

The same hyper-parameters and additional settings were tuned for the in-patient approach. The training-results of the in-patient approach are shown in Figure 4.4, where the y-axis again represents the AUCPR score for all combination of settings. The three lines within each subplot represent the mean (SD) AUCPR score of the inner-loop train and test sets and the AUCPR score of the outer-loop validation set. The best model is again defined by the model with the best performance on the validation set.

When examining these plots, the first notable observation is that the mean scores are higher compared to those obtained with the between-patient approach. When increasing the duration of instability (dT), the performance gradually increased regardless of other settings.
Another noteworthy observation is the effect of the window width (W). Specifically, the performance gap between the train, test, and validation sets appeared to decrease as W increases."
Lastly, the horizon (r) seems to marginally improve the model's performance, regardless of the other settings.

Based on this figure, the final settings for W, r and dT were selected to train one final classifier for the in-patient approach. Firstly, 50 minutes for the setting W was selected. The difference between the train, test and validation set were minimal when using the 50-minute window indicating less overfitting. For the dT settings, a duration of 120 minutes was selected. This duration had the best performance, regardless of the other settings. Lastly, the horizon (r) was selected. The behaviour of r was almost equivalent in all tested settings. Since the performance was marginally higher at 45 minutes and additionally to keep the model similar to the between-patient approach, the horizon was set to 45 minutes. The selected settings are also summarised in Table 4.3.

Final performance

Firstly, the singular CV was performed on the entire train set. The best hyper-parameters were selected based on the best validation AUCPR score. The hyper-parameters are available in Appendix A.2. The in-patient approach is considered quite a complex model, with a lot of decision trees with a lot of specific classification conditions (max. leaf nodes.)

To be able to evaluate the final performance on the test set, the model was trained on 80% of the test set (illustrated as train in orange in Figure 3.3b). These train scores were not retrieved and are displayed as 'NA' in Table 4.5. The final performance was evaluated by the remaining 20% of the test set (called as test in orange in Figure 3.3b) and can be seen in the second 80% column of Table 4.5.
Lastly, it should be noted that the test performances (AUCPR) are higher than the mean train or validation performance of the singular CV.

Train-test-split analysis

Apart from the final performance trained on 80% of the test set, the test performance of the model using reduced train sizes are also shown in Table 4.5. Since the final model was trained on the small train set (orange in Figure 3.3b for this analysis, no mean train performances from the nested CV are available and are displayed with 'NA' in the Table 4.5.

Table 4.5: Training and testing performance of the In-patient approach. In addition to the 80-20 train-test-split, the results of different train-test-splits are displayed. The training performance is not available, since no CV was performed for the additional analysis.

Train percentage:	**80%**	**80%**	**60%**	**40%**	**20%**	**10%**	**5%**
Training performance							
Train AUCPR, mean (SD)	0.984 (0.001)	NA	NA	NA	NA	NA	NA
Validation AUCPR, mean (SD)	0.959 (0.002)	NA	NA	NA	NA	NA	NA
Testing performance							
AUCPR	NA	0.998	0.997	0.994	0.964	0.904	0.817
AUCROC	NA	0.999	0.999	0.998	0.992	0.976	0.949
Accuracy	NA	0.992	0.985	0.977	0.954	0.930	0.913
Balanced accuracy	NA	0.954	0.916	0.869	0.749	0.622	0.561
F1-score	NA	0.975	0.955	0.927	0.847	0.752	0.692

For the 10% train-test-split, the label distribution remains approximately 17% in both the train and test set. The total duration of the train set is 27 hours, which translates to approximately 80 minutes per patient.

4.3.3 Horizon analysis

In Figure 4.3 the AUCPR is plotted against the different settings for the horizon (r). The performance of the models with a higher horizon setting had higher AUCPR scores than those with a lower horizon setting. No other parameters were tuned during this analysis.

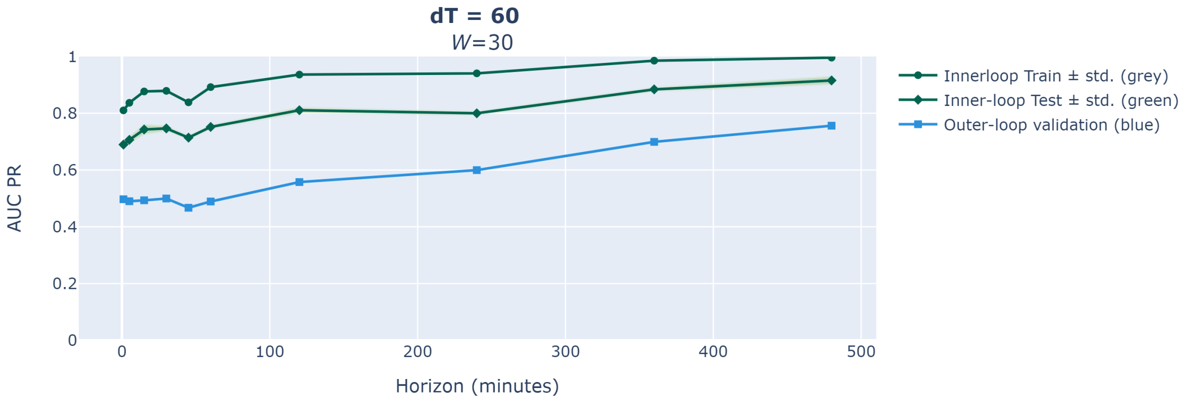


Figure 4.3: The AUCPR is displayed for different horizon (r) settings to explore the effect of increasing the horizon past 60 minutes. The grey line represents the mean (SD) of the inner-loop training performance. The green line represents the mean (SD) of the inner-loop test performance. Lastly, the blue line represents the validation performance. All colours are synchronised with the CV-structure in Figure 3.3

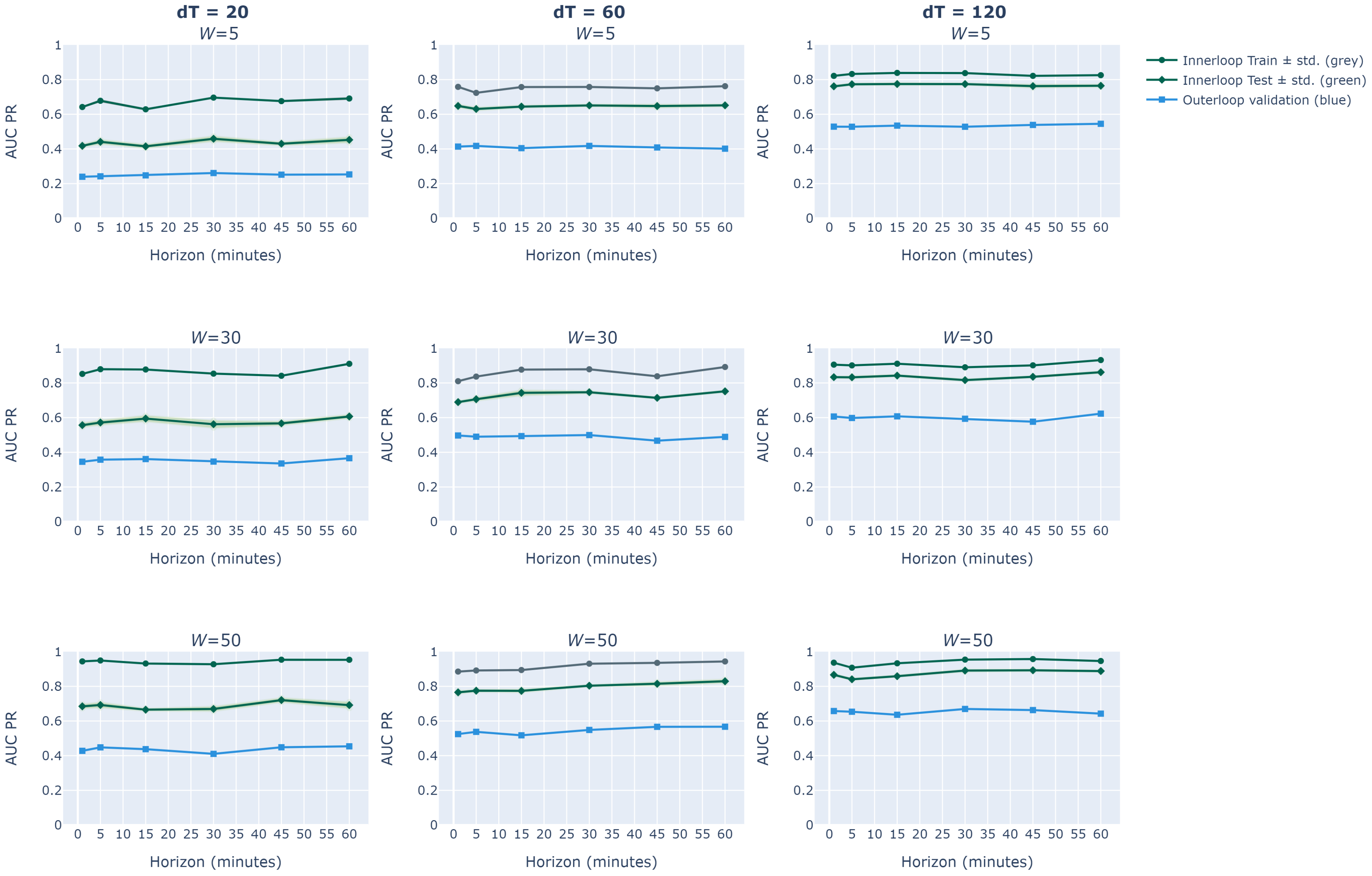


Figure 4.4: Training performance of the in-patient approach and additional analysis for the window-width (W), duration of instability (dT), and the horizon (r). Each column represents another dT setting and each row represents a different W setting.

In each sub-figure, three lines are plotted. The grey line represents the mean (SD) of the inner-loop training performance. The green line represents the mean (SD) of the inner-loop test performance. Lastly, the blue line represents the validation performance. All colours are synchronised with the CV-structure in Figure 3.3

5 Discussion

Two random forest classifiers were trained using a nested CV structure to detect haemodynamic instability in CCHD patients admitted to the PICU, using common clinical interventions as labels for haemodynamic instability. After developing multiple models with different temporal settings (W, r and dT), this study shows that an increased window-width (W, 50 minutes) and an increased duration of instability (dT, 120 minutes) have positive influence on the model performance. However, the model's performance appears unaffected by the horizon (r) setting, as comparable test performances were observed when using a r of 5 or 60 minutes.

This study has shown that the model's generalisation at the population level (i.e. between-patient approach) is poor, currently, with the used imputation and generalisation methods. This conclusion is supported by the notable differences between the mean train (85%, AUCPR) and test performance (40%, AUCPR).

When considering the model at the patient level (i.e. in-patient approach), generalisation issues were mitigated by training and testing on each patient individually. This approach yielded mean train performance of 98% (AUCPR) and a test performance of 99% (AUCPR), when applied to a new study cohort.

These scores were obtained, using 80% of the final test set for training. Practically, lowering the amount of training data (e.g. 10%) is desired for clinical feasibility, yielding an AUCPR score of 91%. While the in-patient approach does introduce new challenges for implementation, it demonstrates the potential to use supervised ML for the detection of haemodynamic instability at the bedside of critically ill CHD patients admitted to the PICU. However, achieving accurate predictions that are clinically relevant require objective data labeling and substantial pre-processing. The insights gained from this process will be used for future improvements of the model.

5.1 Interpretation of results

Lag analysis

The cumulative explained variance plot can be seen in Figure 4.1. To accumulate 95% of the variance within the data, approximately 27 components were needed. This indicates that more than a few lags were contributing to the variance within the input parameters. However, it did not add information about what time lags are contributing.

During hyper-parameter tuning one of the settings was selecting the number of features used for training each decision tree. While increasing W increased the availability of features to the model, only a random sub-selection of hyper-parameters was used. Each decision tree in the random forest gets assigned a new random sub-selection of features. No analysis was performed to compare the important features between the different window-widths.

5.1.1 Between-patients

While the between-patients would be ideal for clinical implementation, the present study has shown that the model does not generalise well when predicting haemodynamic instability for new patients.

While the training-scores in Figure 4.2 are inconsistent, they are significantly higher than the test and validation score, indicating overfitting on the training data. This overfitting in this approach is most likely due to the inter patient variability. However, the testing and validation performance are comparable, indicating that the model is not overfitting to the test data within the inner-loop of the nested CV. With a comparable performance of the validation set to the test set, the model shows consistency when introduced to new data.

Although the performance was consistent, low scores were observed. The CV splits of both the inner- and outer-loop were repeated exactly the same for each training iteration. By using nested CV, every admission is at least once part of the test set. The difference in performance based on the split for the between-patient approach is clearly visible by the wide SD of the test score in Figure 4.2 and is not because the performance was evaluated using a significantly different split. Hence, the relatively low test and validation scores are most likely not caused by the data splits. Additionally, since the best performing models are displayed in Figure 3.3a, it makes sense that the validation scores are more comparable to the upper-range of the SD-interval, instead of the mean test scores. If the validation set was significantly different from the test sets in the CV, the validation scores are expected to be outside the SD range.

It is more likely that the inter-patient variability is too substantial making it difficult to establish a correlation between the data and the labels. Moreover, the relationship between the data and the current version of the labels may contribute to the poor performance.

Intuitively, the performance is expected to decrease when increasing the horizon, since it is less likely that a correlation exists with a large time difference between the data and the moment of prediction. However, during the analysis, the horizon did not seem to influence the performance of the model. Therefore, the consistent test and validation scores, might indicate the inherent baseline performance of the model. In other words, most of these settings did not seem to matter, because the model will always predict the same scores. However, since the AUCPR scores were higher than class prior for each dT variant, it is assumed that the model is contributing to the performance score.

Another possible explanation could be incorrect implementation of the methodology, resulting in the observed consistent performance. The methodology of the nested CV was validated, using the methods available in Appendix A.4. When overfitting to one patient, the model was capable of approaching a perfect train, test, and validation score, suggesting that the methodology was implemented correctly.

Finally, it might be possible that the model is capable of predicting haemodynamic instability with similar performance when horizons of, for example, 5 or 60 minutes were used. The model might select different features from the lag-analysis to split the decision trees, which are more predictive further in the future.

More research regarding the behaviour of the current between-patient approach is needed before this method could be implemented in practise.

5.1.2 In-patient

When comparing the training performance of the in-patient approach, shown in Figure 4.4, to the between-patient approach, the scores were significantly higher. Similarly, the mean test performance of the inner-loop was significantly higher than the validation performance of the outer-loop. Intuitively, since the data were split within patients, such a significant deviation should not occur. The reason the validation performance was lower than the inner-loop test performance is unclear. It could be caused by a methodological error that does not correctly split the train and validation set using data from the same patients. However, the validation scores are expected to be comparable to those of the between-patient approach if completely unseen data is included in the validation set. No additional analysis was performed to validate the in-patient train-test-split within the nested CV structure.

As mentioned in the result section, when using a train size of 80%, the model had a higher AUCPR score on the test set compared to the mean AUCPR score of the train set. These results should not be directly compared, since a new model was trained and tested with the in-patient approach for the final performance. The singular CV was only used to select the hyper-parameters.

Nonetheless, the observed test performance of the model achieved with a train set size of 80%, were approaching a perfect test score. Since the sample-size was lower in the test set (the entire orange test set from Figure 3.3), it was easier for the model to 'overfit' on these patients, resulting in high test scores. Overfitting is used in the in-patient approach and pleads for a successful translation to a new group of patients.

Train-test-split analysis

It is clinically relevant to be able to train the in-patient algorithm with as little data as possible. The model can start predicting the patient's clinical status sooner if less data is needed for the model development.

Although the performance of the model using 5% of the data for training decreased compared to other train sizes in the in-patient approach, the model was still outperforming the between-patient approach. This 5% translates to approximately 40 minutes of data per patient.
The performance was comparable to larger train set, until the train set was set at 40%. The balanced accuracy and the F1-score seemed to slowly decrease when the train set percentage was further reduced. The dataset is imbalanced with a class prior of approximately 18%, which is likely contributing to the lower performance in the F1-score and balanced accuracy. These measures are derived from sensitivity and specificity, emphasizing different aspects of label classification than performance metrics derived from the positive predictive value. The clinical implementation dictates the weight of the false positives or false negatives. Since the model is trying to predict haemodynamic instability, the false negatives should be avoided. Therefore, the positive predictive value (i.e. precision) and sensitivity (i.e. recall) are considered more important, which makes the AUCPR more relevant for this specific algorithm.
Based on this performance, a train set threshold between 10 and 20% seems feasible without considering other practical aspects of implementing an in-patient ML model in practise.

Importantly, the data were stratified for labels in this analysis. When assessing the clinical feasibility, the performance should be tested within one patient. One of the requirements of this method, is that both labels must be available when training the algorithm. Due to this limitation, the training time might require more data in a real admission compared to the minimum required amount of data determined by this train-test-split analysis. During this analysis, the performance of a patient specific model using only one patient (simulating a new admission at the PICU) was not yet explored, making it difficult to determine the feasibility of this model in clinical practise.

5.1.3 Horizon analysis

The horizon was included to be able to predict haemodynamic instability in a timely manner. Since the performance remained stable when increasing the horizon, this additional analysis was performed. When the horizon was increased to 480 minutes, the performance started rising instantly after increasing the horizon parameter past 60 minutes. Figure 4.3 shows the performance of different horizon parameter while using window-width of 30 and a dT of 120 minutes. The nested CV methodology was validated, using two sub-analyses, elaborated in Appendix A.4. The methodology worked as expected using a small number of patients, suggesting that this was not the cause of the rising performance of the increased horizon.
As mentioned in Appendix A.1.2, during the pre-processing, a sub-selection of 800 timestamps was collected for each admission. The horizon parameter was set after creating this sub-selecting, decreasing the amount of data for training the algorithm. The theory is, when increasing the horizon parameter, too much data gets lost in the process, resulting in an overestimation of reality. In Appendix A.2, an illustration of the theory is supplemented. To validate this theory, future horizon analyses are needed to explore the effect of different horizon settings with an equal amount of data.

Apart from pre-processing related causes of the behaviour of the horizon, it is most likely that the horizon is underperforming due to the lack of a clear correlation between the input data and the assigned labels. This argument is stressing the importance of a correct labelling method and feature selection for a successful model development.

To be certain the horizon is contributing to predicting haemodynamic instability, multiple aspects of the model (i.e. labels and horizon) must be further improved to become more reliable and safer before this model can be implemented as a bedside warning system for critically ill CCHD patients.

5.1.4 Labels

In addition to the horizon being performing as expected, the used labels also require additional analyses. The current labelling system includes subjectivity and the timing of true instability onset and treatment are not always obvious.

Assuming physicians are treating these principles correctly, their decisions while treating a patient are considered the golden standard. In this iteration of the model, the labels were based on the interventions mentioned in Table 3.1. These were selected using expert opinion and availability in the EHR.
Due to personal experience, treatment plans may vary for patients depending on the attending physician. In extreme cases this variability may even threaten the continuity of care (18). These 'hunches' from physicians are not retrospectively retrievable and cannot be included in the labelling system.
Moreover, a delay between symptoms' onset and treatment exists, of which the duration may also vary between physicians or nurses. To counteract this phenomenon, the analysis for finding an optimal duration of labeling haemodynamic instability (dT) was conducted.

Limitations of the labelling method also hide in the data acquisition. Firstly, the time of administration in the EHR might not always concur with the true admission time, introducing more inconsistency when assigning labels. Secondly, the interventions collected from the EHR also include changes in the infusion speed of the infusion pump, which can be increased or decreased. A decrease of infusion speed is also considered an intervention, possibly over-presenting haemodynamically unstable periods. Moreover, in this iteration all interventions were considered equally important. However, in reality, it is unlikely that vasoactive medication is initiated as frequently as fluid therapy. Further research is required to better understand the contribution of various interventions in defining objective labels for haemodynamic instability.

Despite this labelling method attempting to objectify haemodynamic deterioration, it still involves a lot of subjectivity. This challenge is stressing the importance of a correct translation between the medical world and the world of ML algorithms with the strict requirements for correct labels. Due to time restrictions no additional analysis was performed to explore whether correlation exists between false negative predictions and a specific type of intervention. It is suggested to explore the incorrect classifications for medication/fluid type, dose and perhaps manually checking the reason of administration.
In this thesis, multiple settings of dT were explored. All these different values were prior to an intervention, however, not all interventions were directly effective, and most patients were not stable after only one of these interventions. Therefore, the period of instability should be extended until after the intervention.
Lastly, the utilised method of classifying haemodynamic instability is binary, which is not representative with the reality. Patients can also clinically deteriorate in a graduate manner, with the degree of instability increasing slowly. Labelling with a continuous scale would be preferred, so possibilities should be explored.
All in all, the quality of data and labels play an important role in a model's success or even feasibility.

5.1.5 Performance metrics

Originally, the AUCROC was used for the assessment of the CV method. Due to the nature of the AUCROC, it is likely to overestimate the performance of a model trained with an imbalanced dataset, because it is less sensitive to changes in the true negative rate (TNR) (22). This principal is also illustrated in the results of the train-test-split analysis in Table 4.5. While decreasing the train set size, the AUCROC is not decreasing. The AUCPR is better suited for imbalanced datasets, where the negative labels outweigh the positive samples or in situations where the correct classification of positive samples is considered more important. When comparing the results of the train-test-split analysis in Table 4.5, the type of performance metrics emphasises the importance of the class imbalance. While the AUCROC stays above 90% for all splits, the balanced accuracy (which is more sensitive to false negatives), is decreased from 95% to 56%, stressing the importance of a performance metric fitting the goal of the prediction model.

5.2 Comparison to literature

The combination of using labels as defined in this thesis and using only vital signs that are high-frequently available is not seen before in literature.

Potes et al. have attempted to create a risk score for haemodynamic instability for patients admitted to the PICU with an age between 1 month and 20 years (23). To account for the data variability introduced by age, five aged-based sub-selections were created. The start of a period of instability was defined by the first intervention of an admission, these interventions being either vasoactive medications or fluid therapy. The model will create a risk score up to 12 hours before the first haemodynamic unstable episode of the admission. This approach neglects the delay between instability onset and treatment initiation, posing a risk of creating a risk score for intervention. Their algorithm, also consisting of an ensembled decision tree method (Adaboost), was trained using 21 input parameters, including laboratory results and ventilator settings. Using intermittent input parameters, such as laboratory results, introduce predictions with different levels of confidence due to the number of missing variables. The model requires at least a heart rate and age to be able to generate a risk score. Additional sample rates requirements were established to be able to include these intermittent variables.
The model had an overall test AUCROC score of 81% and was externally validated by using data from another hospital, resulting in a validation AUCROC score of 81%. When only considering vital signs as input parameters over all age groups, the AUCROC was 71%. Class priors or class imbalances were not discussed, however, comprehensive analysis to potential bias were conducted.

When comparing the vital-signs only version of the model to the between-patient approach of this study, their approach does vary from this study. First of all, the study cohorts were not age restricted, the input parameters had no minimal required frequency and the model was only working for the first episode of haemodynamic instability, since the result is a risk score for developing haemodynamic instability.

When comparing test results of the vital signs only approach with the between-patient approach, the AUCROC scores are identical with 71%. They did not mention other performance metrics, so unfortunately it is not possible to further interpret and compare the results. As mentioned before the AUCROC is most likely an overestimation for predicting haemodynamic instability, however, this cannot be assessed based on the information given in their study.

On the other end of the spectrum are *Zoodsma et al.* (15), they only use high frequency vital signs in CCHD patients admitted post-operatively to the PICU. However, they did not define objective labels using data from the EHR. Their model exists of three sub-models, one to classify sensor dysfunction, on to classify the input parameters of one patient, and the last sub-model to follow a patient's clinical status through time.

The classifier used in sub-model 2 was a one-class support-vector-machine (OCSVM) and was trained using the normal distribution of the data, eliminating the need for labels during training. They mention that 29 hours are unstable in a total recording length of 210 hours collected from 10 patients included in the validation process. This translates to a class prior of approximately 14%, which would be comparable to a dT somewhere between 60 and 120 minutes in our study. The model was evaluated by two intensivists, retrospectively validating the prediction with the use of the input data, resulting in a balanced accuracy of 85%. The balanced accuracy of the between-patient approach is only 32%, making the performance of the model of Zoodsma et al. significantly higher. However, the validation process was subjective and not standardised. Making it hard to truly compare these performances or to reproduce this result.

Rahman et al. attempted to create an ensemble of decision trees (abstain boost) to predict a real-time risk score to predict interventions with a horizon of one hour (24). They included over 200.000 admissions, with a prevalence of 18%. Although using continuous and intermittent data collected from adults, their data was also labelled using fluid therapy or vasoactive medication as onset of haemodynamic instability. The period of instability was maintained until nothing was administrated for 12 hours. Their approach was more modular than the approach of *Potes et al.*, making it possible for the patient to have both stable and unstable periods during an admission. Data collected one hour before the intervention period was labelled as unstable, capturing the onset-treatment delay. They evaluated their model using 4 subset of input parameters, with AUCROC scores ranging from 72 to 82%. Additionally, the PPV, the specificity and sensitivity were reported with different cutoff values. Since this model was trained in adults, the results are not compared, since the physiology differs significantly. However, the methodology used is very robust and this approach should be taken into account when considering future steps in this project.

5.3 Strengths and limitations of the study

The strengths of this study involve a robust methodology with consistent high-frequently available data.

First of all, compared to most ML papers in health care, the sample size of 218 patients is substantial. Since 242 patients were excluded due to the absence of data, the number of included patients can be doubled when the model is better able to handle missing data. Additionally, the sampling rate of the available data was 1Hz, enabling the development of a high frequency/real-time prediction model for haemodynamic instability, with an equally high prediction resolution. Additionally, down-sampling of the data reduced the size of the dataset, lowering the computational power needed for developing the algorithm.
Furthermore, this study was the first study that attempts creating a high frequency prediction model for CCHD patients admitted to the PICU, with objectively assigned labels. Although, the performance of the model looks dependent on the quality of these labels, a first reproducible approach was created.

However, some limitations need to be addressed. First of all, although we attempted to create a homogenous study cohort, there may still be considerable heterogeneity as diagnoses and reasons for admission are not routinely recorded (in a standardised manner) when a patient is admitted to the PICU of the Erasmus MC Sophia children's hospital. Although a large variety of admission types may create a generalisable model, it can also introduce noise for a ML model to predict correctly. Not all admissions were guaranteed post-surgical or due to the CCHD.

Furthermore, including only patients with measurements of all input parameters introduced selection bias. For example, continuously measuring the MAP or CVP is currently not included in standard care. Nonetheless, an attempt at reducing the selection bias was made by not excluding data based on data quantity and choosing for imputation alternatives.

Lastly, the retrospective nature of the study is most likely the biggest limiting factor. For example, the labels cannot be properly generated, since the treating physician had specific considerations that are not retrievable from an EHR. In addition, data quality cannot be guaranteed, for example, the absence of an arterial catheter does not have to be due to improvement of the patient, perhaps the catheter failed and a new one cannot be inserted. Moreover noise cannot be retrieved with certainty from retrospective data, what could be annotated in a retrospective study, although very difficult and most likely unfeasible.

5.4 Future recommendations

When preparing the data, it became clear that not all parameters were available during the entire admission. The current classifier was not optimised for handling missing data. Therefore it is suggested to improve the current classifier or explore more classifiers that are capable of handling this type of data. This grants the opportunity to add a 'reliability' score to predictions not using all input parameters. Conroy et al. compared different versions of AdaBoost with abstaining that can handle missing data (24, 25). Their approach should be considered as the next step when more classifiers are being explored.

Another option would be exploring the 'black box' approach and try to train a deep learning model for this data, for example the model used by Ruiz et al. (26). However, the labels should be better defined before it is likely that a deep-learning approach is successful.

Before developing a model with a new classifier, the current model can be further improved as well. Currently, the model does not generalise well as can be seen in the between-patient approach. The data was scaled based on the data from entire population, using the mean and standard deviation. Since the in-patient performance generalises better than the between-patient approach, it looks like the inter-patient variability is too prominent. Firstly, the inter-patient variability might be too high due to the included diagnoses or due to the age difference of the study cohort, indicating that a revision of the inclusion criteria could improve the data homogeneity. Alternatively, the data could be scaled per patient, which comes with a similar drawback as the in-patient approach, i.e. a part of data should be scaled before the prediction can be performed. Compared to the in-patient approach, training the scaler is expected to require less data and the labels do not have to be considered. However, it is important that all parameters that are going to be used during the admission, must be available when the scaler is fitted.

When considering the clinical feasibility of the in-patient approach, the method training on new patients should be improved. The random forest classifier is capable of something that is called a 'warm start'. This method will use a model trained on, for example, the dataset included in this thesis. By increasing the parameter 'No. of estimators' and training on the first few hours of a new patient, new decision trees are added to the existing model for this specific patient, instead of generating a new classifier. Conceptually, this method would lower the impact of only training with one label. More research is needed to see whether adding the pre-trained model is adding robustness to the model only trained within one patient and whether training on the few first few hours of a new admission is adding to the existing model.

5.5 Conclusion

In conclusion, this study has attempted to create a supervised ML algorithm to detect haemodynamic instability in CCHD patients admitted to the PICU, using common clinical interventions as labels for haemodynamic instability. Despite this labelling method attempting to objectify haemodynamic instability, it still involves a lot of subjectivity of treating physicians. This challenge is stressing the importance of a correct translation between the medical world and the world of ML algorithms with the strict requirements for correct labels. This study has shown that the model's generalisation at the population level (i.e. between-patient) approach is poor with the used imputation and generalisation methods. It also showed that the patient level (i.e. in-patient) approach is correctly utilising its capability to train and predict within patients. However, a significant challenge lies in the high inter-patient variability, influencing both approaches. Additional analyses have indicated that the prediction model proposed in this study, which combines high frequency vital signs, labels, and temporal settings (W, r, dT), requires additional refinement before it can be considered clinically feasible to implement this model as a reliable bedside tool for predicting haemodynamic instability.

Bibliography

[1] D. van der Linde, E. E. Konings, M. A. Slager, M. Witsenburg, W. A. Helbing, J. J. Takkenberg, and J. W. Roos-Hesselink, “Birth Prevalence of Congenital Heart Disease Worldwide,” Journal of the American College of Cardiology, vol. 58, pp. 2241–2247, 11 2011.

[2] T. Van Der Bom, A. C. Zomer, A. H. Zwinderman, F. J. Meijboom, B. J. Bouma, and B. J. Mulder, “The changing epidemiology of congenital heart disease,” Nature Reviews Cardiology 2011 8:1, vol. 8, pp. 50–60, 11 2010.

[3] M. E. Oster, K. A. Lee, M. A. Honein, T. Riehle-Colarusso, M. Shin, and A. Correa, “Temporal trends in survival among infants with critical congenital heart defects,” Pediatrics, vol. 131, 5 2013.

[4] D. Bonnet, A. Coltri, G. Butera, L. Fermont, J. Le Bidois, J. Kachaner, and D. Sidi, “Detection of transposition of the great arteries in fetuses reduces neonatal morbidity and mortality,” Circulation, vol. 99, no. 7, 1999.

[5] J. Soongswang, I. Adatia, C. Newman, J. F. Smallhorn, W. G. Williams, and R. M. Freedom, “Mortality in potential arterial switch candidates with transposition of the great arteries,” Journal of the American College of Cardiology, vol. 32, pp. 753–757, 9 1998.

[6] G. Gutierrez, “Artificial Intelligence in the Intensive Care Unit,” Critical Care, vol. 24, 3 2020.

[7] D. van de Sande, M. E. van Genderen, J. Huiskens, D. Gommers, and J. van Bommel, “Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit,” Intensive care medicine, vol. 47, pp. 750–760, 7 2021.

[8] N. K. Adhikari and G. D. Rubenfeld, “Worldwide demand for critical care,” Current Opinion in Critical Care, vol. 17, pp. 620–625, 12 2011.

[9] S. S. Khairat, A. Dukkipati, H. A. Lauria, T. Bice, D. Travers, and S. S. Carson, “The Impact of Visualization Dashboards on Quality of Care and Clinician Satisfaction: Integrative Literature Review,” JMIR Human Factors, vol. 5, 4 2018.

[10] K. Y. Ngiam and I. W. Khor, “Big data and machine learning algorithms for health-care delivery,” The Lancet Oncology, vol. 20, pp. e262–e273, 5 2019.

[11] H. Habehh and S. Gohel, “Machine Learning in Healthcare,” Current Genomics, vol. 22, no. 4, 2021.

[12] T. Panch, P. Szolovits, and R. Atun, “Artificial intelligence, machine learning and health systems,” Journal of Global Health, vol. 8, no. 2, 2018.

[13] M. W. Sorensen, I. Sadiq, G. D. Clifford, K. O. Maher, and M. E. Oster, “Using pulse oximetry waveforms to detect coarctation of the aorta,” BioMedical Engineering Online, vol. 19, pp. 1–12, 5 2020.

[14] D. Castiñeira, K. R. Schlosser, A. Geva, A. R. Rahmani, G. Fiore, B. K. Walsh, C. D. Smallwood, J. H. Arnold, and M. Santillana, “Adding continuous vital sign information to static clinical data improves the prediction of length of stay after intubation: A data-driven machine learning approach,” Respiratory Care, vol. 65, no. 9, 2020.

[15] R. S. Zoodsma, R. Bosch, T. Alderliesten, C. W. Bollen, T. H. Kappen, E. Koomen, A. Siebes, and J. Nijman, “Continuous Data-Driven Monitoring in Critical Congenital Heart Disease: Clinical Deterioration Model Development,” JMIR Cardio, vol. 7, p. e45190, 2023.

[16] M. Paliwoda, F. Bogossian, M. W. Davies, E. Ballard, and K. New, “Physiological vital sign differences between well newborns greater than 34 weeks gestation: A pilot study,” Journal of Neonatal Nursing, vol. 26, pp. 226–231, 8 2020.

[17] Y. Chen, L. Wu, L. Zou, G. Li, and W. Zhang, “Update on the birth weight standard and its diagnostic value in small for gestational age (SGA) infants in China,” http://dx.doi.org/10.1080/14767058.2016.1186636, vol. 30, pp. 801–807, 4 2016.

[18] M. Gulliford, S. Naithani, and M. Morgan, “What is ’continuity of care’?,” http://dx.doi.org/10.1258/135581906778476490, vol. 11, pp. 248–250, 10 2006.

[19] L. M. Bigatello and E. George, “Hemodynamic monitoring.,” Minerva anestesiologica, vol. 68, pp. 219–25, 4 2002.

[20] “ANDREA - Digital Research Environment,” 1 2023.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay, “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825–2830, 2011.

[22] F. S. Nahm, “Receiver operating characteristic curve: overview and practical use for clinicians,” Korean Journal of Anesthesiology, vol. 75, p. 25, 2 2022.

[23] C. Potes, B. Conroy, M. Xu-Wilson, C. Newth, D. Inwald, and J. Frassica, “A clinical prediction model to identify patients at high risk of hemodynamic instability in the pediatric intensive care unit,” Critical Care, vol. 21, no. 1, 2017.

[24] A. Rahman, Y. Chang, J. Dong, B. Conroy, A. Natarajan, T. Kinoshita, F. Vicario, J. Frassica, and M. Xu-Wilson, “Early prediction of hemodynamic interventions in the intensive care unit using machine learning,” Critical Care, vol. 25, pp. 1–9, 12 2021.

[25] B. Conroy, L. Eshelman, C. Potes, and M. Xu-Wilson, “A dynamic ensemble approach to robust classification in the presence of missing data,” Machine Learning, vol. 102, pp. 443–463, 3 2016.

[26] V. M. Ruiz, M. P. Goldsmith, L. Shi, A. F. Simpao, J. A. Gálvez, M. Y. Naim, V. Nadkarni, J. W. Gaynor, and F. R. Tsui, “Early prediction of clinical deterioration using data-driven machine-learning modeling of electronic health records,” The Journal of thoracic and cardiovascular surgery, vol. 164, pp. 211–222, 7 2022.

[27] T. pandas development team, “pandas-dev/pandas: Pandas,” 1 2024.

A Appendix

A.1 Pre-processing

A.1.1 Artefact detection

Before training a model on high frequency time-series data, it is important filter out all artefacts. A limited form of artefact detection was applied, consisting of the following rules: every zero was replaced by a NaN-value, except for zeros in the HR parameter, since an asystole could potentially occur in these patients.
Moreover, all negative pressures for the CVP and MAP were replaced by NaN-values. Negative pressures in blood vessels are not naturally occurring and are due to sensor calibration or sensor dysfunction. All timestamps with missing data were imputed.

A.1.2 Imputation

Whenever data was missing or removed during the artefact detection, the simple imputation method forward filling was used as a starting point for this model (27). When using forward-filling, periods of missing data were filled with the last known value that is not null. To indicate whether a timestamp contains an imputed data point, an imputation parameter was created for each patient. Whenever one (or more) of the five parameters were imputed, the corresponding timestamp was annotated as 'imputed' with the value '1' in the imputation parameter. This parameter only contained values of '0' or '1' for the entire duration of the dataset and was included in developing the model, adding the opportunity for the model to correct for a correlation between 'real' and 'imputed' data.

In case of a missing value at the start of an admission, the method forward filling was not applicable Vital signs from these admissions were included from the first timestamp where all parameters had a first non-null value. After removing the timestamps at the beginning of an admission, the remaining data was reviewed again for admission length to comply with the inclusion criteria.

To create a ML model that performs best under ideal circumstances, it is crucial to avoid relying on imputed data whenever possible. Apart from some classifiers assuming complete data (e.g. Logistic Regression), ignoring missing values may also lead to a loss of (important) information, creating a false representation of reality. To prevent training on a database by half of the parameters being repetitions of the last known value, three subsets of 800 timestamps (approx. 13 hours) from the available data were extracted. The section with the fewest number of imputations was used for a more favourable model development and expectantly for a better overall performance.

A.1.3 Resampling

Continuous high frequency vital signs can alternate quickly based on true variation, measurement error, or signal noise. To reduce the effect of these fluctuations and to reduce the required computational capacity, the data was downsampled to one value per minute (1/60Hz) by calculating the mean. Only the 'imputed' parameter was downsampled using the median, to maintain binarity.

A.1.4 Scaling

Data from different patients were included, all with slightly different baselines and normal ranges for each parameter. These differences can occur due to age or the underlying disease. To correct for these differences, data was normalised using a `RobustScaler` from SkLearn (21). This scaler is capable of handling outliers and skewed dataset, by utilising the median and the interquartile range of all measurements. The first hour of each admission within the train set is extracted to create the scaler. In Appendix A.1 the power spectral density plots are shown for each parameter used for this decision. It illustrates the different data contributions per parameter and the outliers in the raw data. The distributions of SpO2 and CVP, in particular, are not Gaussian distributed. Furthermore, outliers are clearly visible for the HR, CVP, and ART M (MAP) vital signs. These data characteristics are arguments to use the RobustScaler.

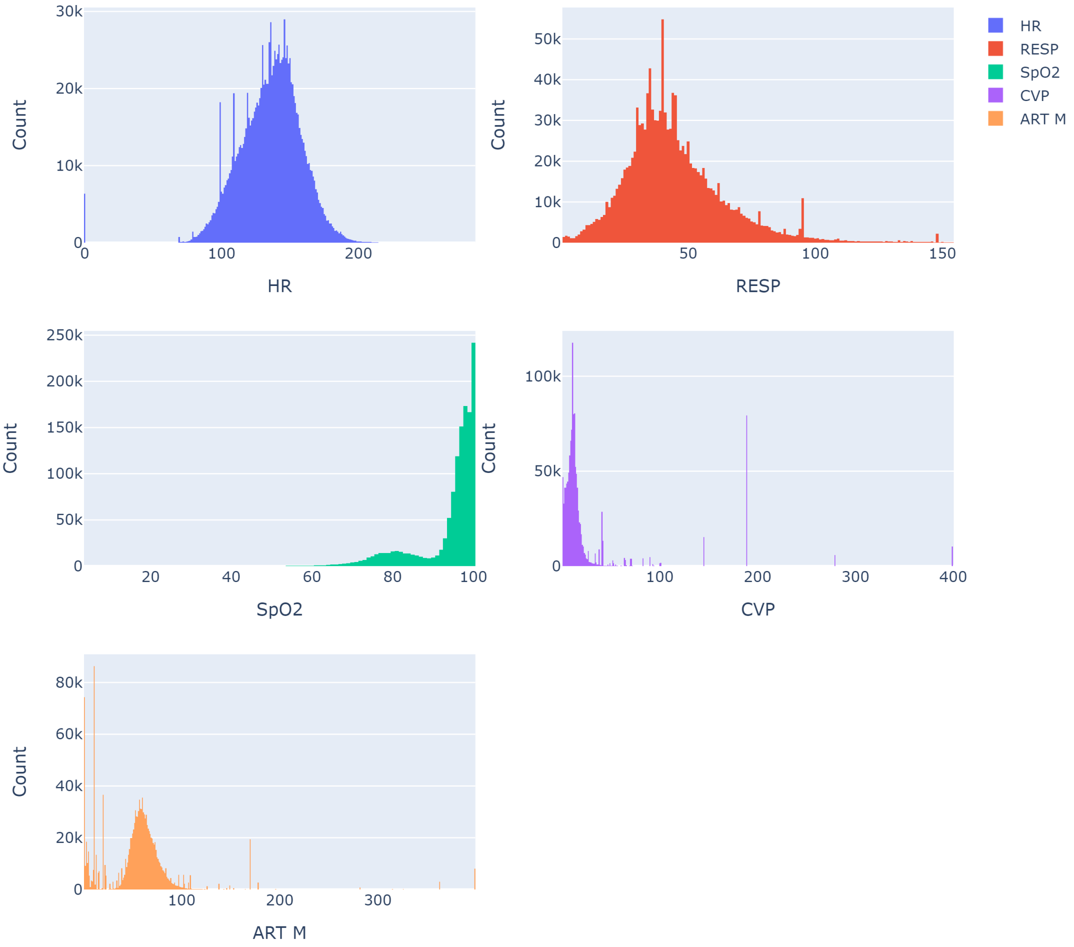


Figure A.1: Power spectral density plot of all raw input parameters.
HR: heart rate, RESP: respiratory rate (RR), SpO2: peripheral oxygen saturation, CVP: central venous pressure, ART M: mean arterial pressure (MAP)

A.2 Horizon analysis

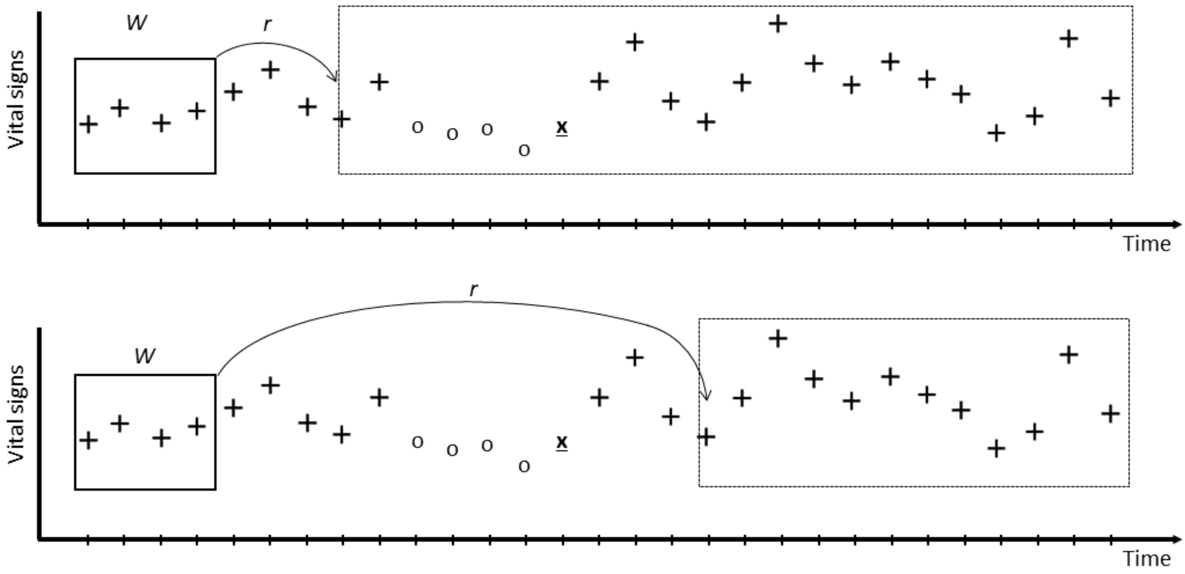


Figure A.2: Explanation of influence window-width (W) and horizon (r) on data size. This figure shows that the less data remains available for prediction (indicated by the box with the dotted line) with a high window-width and a high horizon.

A.3 Final hyper-parameters

Parameter	Value
Class weight	Balanced sub-sample
Criterion	gini
Max. depth	10
Max. features	10
Max. leaf nodes	500
Min. samples split	2
No. of estimators	100

Table A.1: Selected hyper-parameters for the between-patient approach.

When shortly summarized, the between-patient approach is a relatively simple model, apart from the ‘max. leaf nodes’ and the ‘No. of estimators’ hyper-parameters. The random forest classifier is allowed to create 100 decision trees, which all can have 500 final classification conditions. Apart from that, it can only use up to ten features and have a maximum of ten consecutive splits.

Parameter	Value
Class weight	Balanced sub-sample
Criterion	entropy
Max. depth	40
Max. features	50
Max. leaf nodes	600
Min. samples split	2
No. of estimators	200

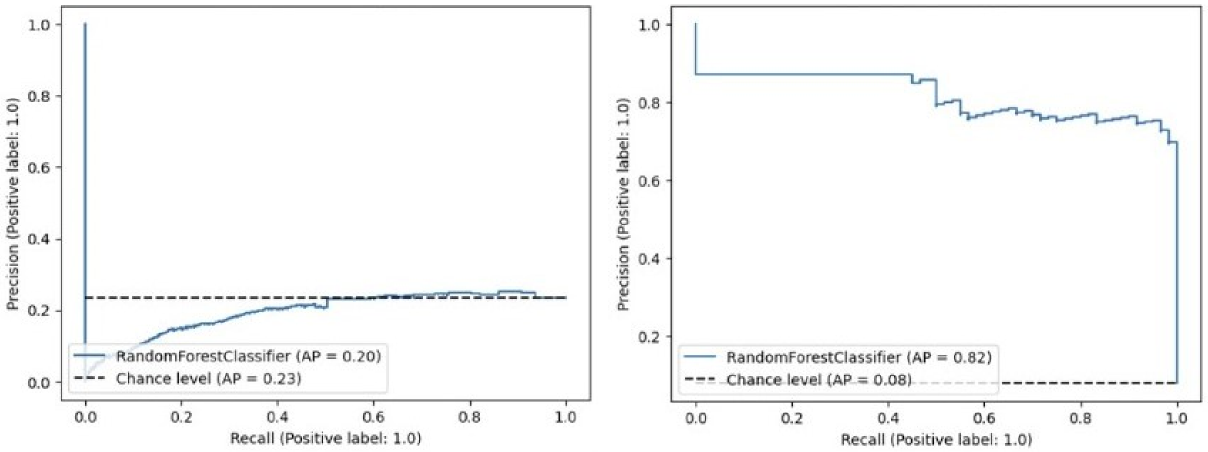
Table A.2: Selected hyper-parameters for the in-patient approach.

When shortly summarized, the in-patient approach has a complex model random forest classifier. It is allowed to use 200 decision trees, which all can have 600 classification conditions. The model can select up to 50 features and have up to 40 consecutive splits. Making it possible to really overfit on the train data.

A.4 Methodological validation

To ensure that the CV setup is working correctly and the results considering the horizon are not due to a methodological error, two sub-analyses are performed. The first approach is overfitting the model using data of only one patient. The data one random patient was copied five times after which a fivefold CV was started stratified for studyID. The patient was four time present in the train set and one time present in the test set, for all folds. This resulted in five perfectly overfitted models with an AUCPR train and test score of 1.00. This process was repeated for multiple horizon settings, all showing a similar result. This analysis indicates that the nested-CV for the between-patient is working as expected, validating the low test scores (green and blue) of Figure 4.2.

Another test performed concerning the cross-validation robustness, was overfitting on 1 patient and testing on another. The score is expected to be close to the class prior, when the patients' data is not comparable. The model was trained using a 3-fold CV, with data from patient A and B. The data of patient A was copied, making the data distribution between A and B 2:1. The PR-curve is added, see Figure A.3 for two situations: trained on patient A and tested on patient B (Figure A.3a) and trained on patient A and B (Figure and tested on patient A A.3b).



(a) Precision recall curve when trained on patient A and tested on patient B

(b) Precision recall curve when trained on patient A and patient B and tested on patient B

Figure A.3: Visual overview of the nested cross-validation (CV) structures. The inner-loop is used for hyper-parameter tuning. The outer-loop is used for a performance assessment of the best performing classifier from the inner-loop.
The in-patient approach needs to be trained first on a new subset of patient data, therefore, the final test set is again split in a train and test section.

These illustrations show that the model is behaving as expected, namely, performing close to the class prior when a new 'unfamiliar' patient is used for testing and performing well when a patient is used for training and testing.