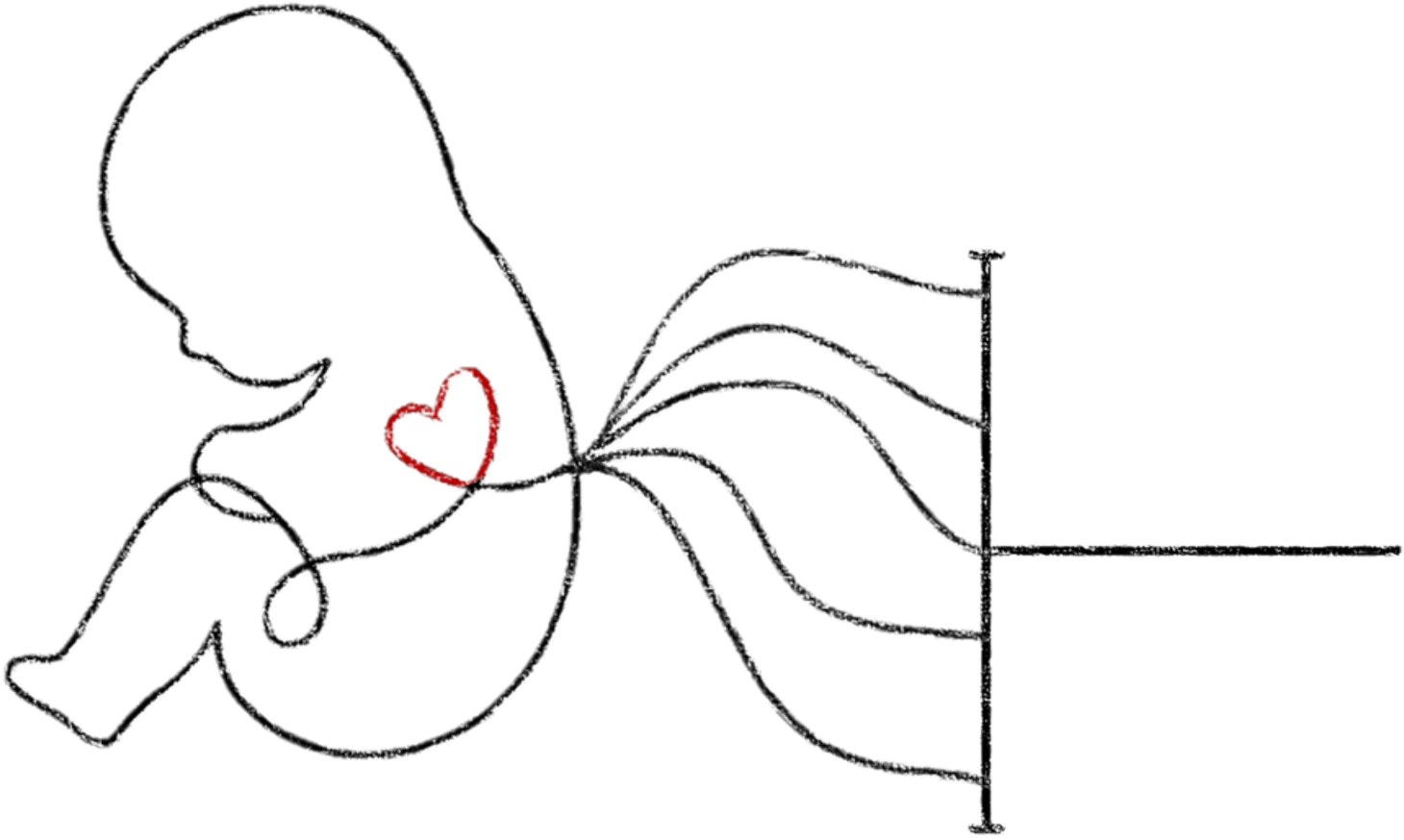
Early Warning of Haemodynamic Instability Using Machine Learning in Critical Congenital Heart Disease Patients



Harmen Schmidt
Master Thesis
Technical Medicine

Early Warning of Haemodynamic Instability Using Machine Learning in Critical Congenital Heart Disease Patients

Harmen Schmidt
Student number: 4712595
Date: 25 August 2025

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 Pediatric Intensive Care,
Erasmus MC Sophia Children's Hospital
24 February 2025 – 9 September 2025

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
Eris van Twist
Brian van Winden

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

Universiteit Leiden

TU Delft Delft University of Technology

Erasmus Universiteit Rotterdam

Preface

After 8 years, this thesis marks the end of my time as a student. It's been a long and windy road during which I often doubted whether I was on the right track. Before starting, I hesitated for a long time about whether I wanted to study medicine or physics. I decided on the BSc Applied Physics, but after a full year I made the switch to the BSc Technical Medicine as the perfect middle road between the two programs. After the BSc, I found myself again reconsidering my study choice and contemplated the MSc Biomedical Engineering, looking for a more technical challenge. In the end, I continued with the MSc Technical Medicine, and although the next destination is yet unknown, I am confident I am on the right path.

When deciding where to graduate, I was looking for a department with enthusiastic supervisors and an interesting project. After glowing reviews from some of my fellow students and a positive—albeit brief—interaction during my BSc thesis, I decided to contact the Paediatric Intensive Care department of the Erasmus MC Sophia Children's Hospital. After our first meeting, I knew I had to look no further.

I want to thank all of you: Rogier, Jan Willem, David, Eris, and Brian. This project was not always sunshine and rainbows for me, but I can confidently say that every meeting I entered with one of you, I exited with more enthusiasm and motivation than when I entered. Thank you, Eris and Brian, for your daily supervision and can-do attitude, as well as always having your door open (as long as I don't forget my employee card). Thank you, Rogier and Jan Willem, for our meetings. You often pretended not to follow along with the more technical talk, but in the end you managed to provide valuable insight after all. Thank you, David, for our Monday meetings and starting my week with a boost of inspiration. Lastly, thank you, Joppe, for taking the time to read and evaluate my work and for being part of my thesis committee.

I'd also like to thank all of the friends working on their thesis at the Erasmus MC during this same period. Having lunch and coffee breaks, and sharing this experience, made it much better. I'd also like to thank my roommates for providing much-needed distractions (and dinner). And, of course, my family for supporting me throughout my studies.

Harmen Schmidt,
Rotterdam, August 2025

Abstract

Background: Patients suffering from critical congenital heart disease (cCHD) require cardiac intervention within the first year of life. During the postoperative period, patients are at risk of haemodynamic instability resulting in insufficient organ perfusion and subsequent organ failure. To prevent this, patients are placed in the paediatric intensive care unit (PICU) where various vital parameters can be monitored. However, interpreting these continuous data streams can be challenging. Machine learning offers the potential to support clinical decision-making in this setting, but challenges remain, particularly in labelling haemodynamic instability and accounting for varying physiology of this patient demographic. The goal of this study was to improve the retrospective labelling of haemodynamic instability and evaluate the affect of age-stratified subpopulation on model performance.

Methods: This study used a retrospective dataset of continuously measured parameters (heart rate, respiratory rate, mean arterial pressure, central venous pressure, oxygen saturation, and perfusion index) collected from post-operative cCHD patients admitted to the PICU of Erasmus MC Sophia Children's Hospital, the Netherlands, between January 2016 and April 2025. A new scoring system was developed to quantify the haemodynamic support received by patients and to identify intervention times at which support was increased. These interventions were used to label haemodynamic instability in the a period dT prior to intervention. The resulting labelling was applied to train a random forest algorithm to predict haemodynamic instability, and the model was subsequently retrained on age-based subpopulations.

Results: A total of 425 patients were included for this study. The new labelling method resulted in 5.7% of the data being labelled as haemodynamically unstable. The random forest using the new labelling achieved an average (SD) area under precision-recall curve (AUCPR) of 0.233 (0.041) on the test set during cross-validation and final test AUCPR of 0.203. The largest age subpopulations were 0–30 days and 90–180 days. The class prior of instability was 9.2% in the 0–30 days subpopulation and the prediction model achieved an AUCPR of 0.244 (0.064). In the 90–180 day subpopulation the class prior was 4.9% and an AUCPR of 0.221 (0.105) was achieved.

Conclusion: This study proposed a new method of retrospectively labelling haemodynamic instability with the goal of training a predictive model to predict these instabilities. A random forest model trained using the new labelling showed limited improvement, with an AUCPR of 0.204 and an AUCROC of 0.762. Age-based subpopulation analysis indicated potential for reduced data variation, though larger cohorts are needed for better generalisation. Further refinements in the retrospective labelling of haemodynamic instability are required for an effective prediction model to be developed.

Contents

List of Abbreviations

AUC	Area Under the Curve
AUCPR	Area Under the Precision–Recall Curve
AUCROC	Area Under the Receiver Operating Characteristic Curve
CHD	Congenital Heart Disease
cCHD	Critical Congenital Heart Disease
CV	Cross-Validation
CVP	Central Venous Pressure
ECMO	Extracorporeal Membrane Oxygenation
EHR	Electronic Health Record
HR	Heart Rate
IS	Intervention Score
MAP	Mean Arterial Pressure
MDI	Mean Decrease in Impurity
ML	Machine Learning
PICU	Paediatric Intensive Care Unit
PI	Perfusion Index
RR	Respiratory Rate
SD	Standard Deviation
SpO_2	Peripheral Oxygen Saturation
VIS	Vasoactive-Inotropic Score

Introduction

Approximately 8–9 in 1000 children are born with congenital heart disease (CHD), making it the most common birth defect[1–4]. Of these, around 26% present with critical congenital heart disease (cCHD), requiring cardiac intervention within the first year of life[5]. During the perioperative period, these infants are at risk of haemodynamic instability, a clinical state in which the body is unable to maintain adequate blood pressure, leading to insufficient blood flow and oxygen delivery, and consequently compromised organ perfusion.

To reduce this risk, patients are admitted to the paediatric intensive care unit (PICU), where multimodal monitoring provides continuous measurements of vital parameters such as heart rate (HR), respiration rate (RR), mean arterial pressure (MAP), central venous pressure (CVP), and oxygen saturation (SpO_2). Although multimodal monitoring helps track these physiological parameters, it is not feasible for hospital staff to continuously track all parameters for all patients. Subtle but clinically important changes can therefore be overlooked. Recent advances in machine learning (ML) offer the potential to support clinical decision-making by automating the analysis of the vital parameters. Supervised learning, which uses labelled data to train models that can classify or predict clinical outcomes, has already been widely applied in healthcare for tasks such as risk stratification and outcome prediction.

For supervised ML to be applied effectively, accurate labelling is necessary to distinguish stable from unstable periods. Retrospective labelling of instability remains challenging. MAP is often used in adult patients, with a predefined threshold below which instability is assumed, but this oversimplifies the complex developing cardiovascular physiology of young children[6]. An alternative approach is to define haemodynamic instability based on the interventions used to counteract it, such as the administration of fluids or vasopressors/inotropes[7–9].

One example of this intervention-based approach is the work of Van Winden et al., who defined instability using retrospective labelling based on such treatments[10]. Their algorithm generated predictions from five continuously monitored vital parameters: HR, RR, MAP, CVP, and SpO_2. They found that performance was highest when the model was trained on data from the same patient to whom it was later applied (*in-patient* training), but declined when trained on data from other patients in the same cohort (*inter-patient* training). Since the aim is to enable real-time prediction, requiring prior training on the specific patient would be impractical, especially during the early hours of admission when timely detection is most critical.

In their method, a period of time before an intervention was labelled unstable. Interventions were defined as the administration of inotropes, vasopressors, pulmonary vasodilators, or fluids. This definition did not account for dosage differences and treated all intervention changes as equivalent. As a result, labelling errors could occur. For example, errors arose when medication was re-registered in the electronic health record (EHR) despite no change in dose, or when the dosage was reduced. Such entries would still be labelled as interventions, even though the patient's haemodynamic state was stable or improving.

In addition to labelling inaccuracies the large difference between the *in-patient* and *inter-patient* performance is notable. This suggests that the heterogeneity within the patient population limits the ability of models trained across patients to generalise effectively. One strategy to address this issue is to identify subpopulations with lower variability in the patient physiology, where predictive models might perform more reliably. The normal range for vital parameters of children changes as they grow older[11,12]. By stratifying patients into age-based groups, models may better capture the unique physiological characteristics within each subgroup, potentially improving inter-patient prediction accuracy.

The aim of this study is to build upon the algorithm developed by van Winden et al., with the goal of enhancing the performance of the inter-patient algorithm by addressing the previously discussed labelling and patient variability issues. Before turning to these improvements, the Data Acquisition & Model Setup section will describe the original study cohort and how it was expanded for this study, as well as outline the model setup and training pipeline employed in the work by van Winden et al. Chapter 1 will focus on redefining the labelling of haemodynamic instability, while Chapter 2 will examine how model performance can be improved by training subpopulations based on age. In Chapter 3 some of the assumptions made during labelling will be assessed to gain further insight into the data and to inform future research directions.

Data Acquisition & Model Setup

The work of van Winden et al. serves as the foundation for this study, providing the original patient cohort, preprocessing, and modelling framework. In this study, the cohort was extended to include extra patients, applying the same inclusion and exclusion criteria. In addition, an extra vital parameter was added to the full cohort in the form of the perfusion index (PI), which reflects the ratio of pulsatile to non-pulsatile blood flow and serves as an indicator of peripheral perfusion. The following section describes the pipeline and model setup used by Van Winden et al. as well as the expanded cohort used for this study.

Method

Study Cohort

This single centre retrospective study consisted of data collected in patients with cCHD. All patients were admitted to the PICU in the Erasmus MC Sophia Children's Hospital, Rotterdam, the Netherlands, between January 2016 and April 2025. Patients of an age between 0 and 365 days were included if they were admitted for a minimum duration of 800 minutes (13.33 hours). Patients with a birth weight < 2500 g were excluded. This was used to exclude prematurely born patients, since gestational age was not registered consistently. The chance of a full-term infant being born with a birth weight of 2500 g is 2.5%[13]. Patients were excluded if not all of the following parameters were measured HR, RR, MAP, CVP, SpO_2, and PI. Patients were also excluded if fewer than five unique values were measured in one of the parameters as this was assumed to be the result of faulty measurements.

Data Collection & Preprocessing

HR and RR were measured using three-lead electrocardiogram (3M, St. Paul, MN, USA). The SpO2 and PI was measured using the Rainbow SET pulse oximeter (Masimo, Irvine, CA, USA). MAP and CVP were measured using an arterial line and central venous line respectively (Becton and Dickinson, Franklin Lakes, NJ, USA).

The exact preprocessing pipeline is detailed in Appendix A.1. To summarise, data was resampled to 1/60 Hz, scaled, and imputed for missing values. To balance the data across patients, a continuous segment of 800 minutes was selected for each patient. A binary indicator was added to mark imputed data points, which served as a seventh feature alongside the six vital parameters.

Model Design & Training

The prediction model consists of a random forest classifier that uses vital parameters measured within a window (W) as input features to predict future instability (Figure 1). The prediction horizon (H) dictates how many minutes into the future the model makes its predictions. The window moves along the signal in steps of one minute.

The model was trained using a nested cross-validation structure, with five-fold cross-validation applied in both the inner and outer loops (see Figure 2). The inner loop selects the optimal hyperparameters by training each fold with a different set of hyperparameter configurations. Appendix A.2 provides an explanation of hyperparameters and a full list of tuned hyperparameters. This process is repeated 20 times, and the configuration achieving the highest average precision on the validation set of the five inner folds is selected. These hyperparameters are then used to train the model on the corresponding outer fold and evaluated on its associated test sets. This procedure is

repeated for each outer fold. A final test set is kept separate from the training data during model development and is only used for the final evaluation of the model.

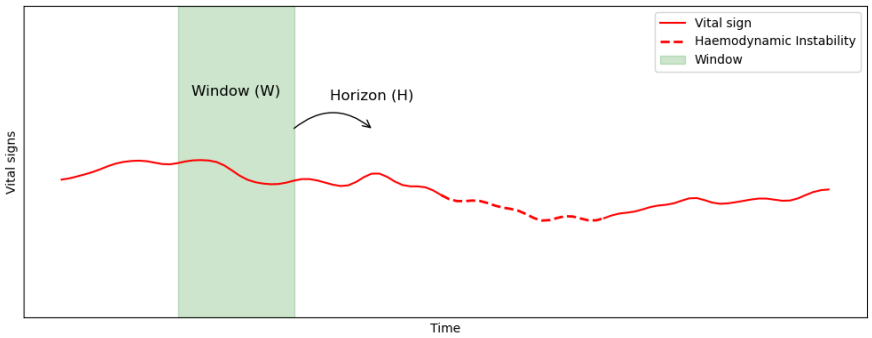


Figure 1: Illustration of the model input and labelling framework. Vital signs are observed in a moving time window (W), which is used to predict haemodynamic instability at a future time point. How many minutes into the future the prediction is made is dependant on the set prediction horizon (H).

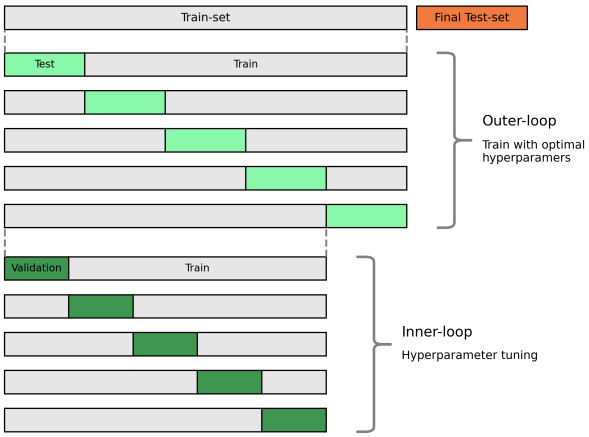


Figure 2: Nested cross-validation procedure. The outer loop splits the dataset into training and test folds to evaluate the model with optimal hyperparameters. Within each outer training fold, the inner loop further splits the data into training and validation folds for hyperparameter tuning. The final performance is assessed on the independent test set.

Model Evaluation

The area under the precision–recall curve (AUCPR) was selected as the primary metric for evaluating model performance due to the low prevalence of positive instability labels in the dataset. In imbalanced settings, metrics such as accuracy or the area under the receiver operating characteristic curve (AUCROC) can be misleading. ROC curves plot sensitivity against the false positive rate, which is strongly influenced by the large number of negative cases. As a result, a model that identifies very few positive cases can still achieve a high AUCROC if it predicts the majority class well. In contrast, AUCPR focuses on recall (synonymous to sensitivity) and precision, making it

more responsive to changes in positive class detection. This is especially important when the goal is to detect haemodynamic instability, where missing a positive case carries a greater cost than generating a false alarm.

When interpreting AUCPR, it is important to be aware that the baseline performance is directly tied to the proportion of positive labels in the dataset, referred to as the class prior. For example, if 10% of the samples are positive, a random classifier would result in an AUCPR of 0.10. This is in contrast to the more commonly used AUCROC which has a fixed baseline of 0.50 regardless of class distribution. As a result, AUCPR values are not directly comparable across datasets with different class distributions.

For all model the mean and standard deviation (SD) of the outer-fold CV performance metrics was given for the train set, the test set was only used for a final performance evaluation.

Results

Study Cohort

An overview of the study cohort characteristics and vital parameters is given in Table 1. The study cohort consisted of 651 eligible ICU patients aged between 0 and 1 year and with a minimum ICU stay of 800 minutes. Of these, 169 patients were excluded due to one or more vital parameter not having been recorded. An additional 43 patients were excluded for having a birth weight below 2,500 grams. A final 14 patients were removed because one or more of their vital parameters contained fewer than five unique values, resulting in a final cohort of 425 eligible patients.

Table 1: Summary of study population and vital sign characteristics for both the old cohort originally used by Van Winden et al. and the new extended cohort used for this study. The PI was not used in the old cohort.

Study Population	**Old Cohort (n=224)**	**New Cohort (n=425)**
Male population, n (%)	136 (61)	234 (55)
Age (days), median [Q1–Q3]	132 [85–182]	104 [43–169]
Vital Signs	**Median [Q1–Q3]**	**Median [Q1–Q3]**
HR (beats/min)	137 [121–175]	140 [125–154]
RR (breaths/min)	41 [32–54]	35 [28–42]
SpO_2 (%)	97 [94–99]	97 [95–99]
MAP (mmHg)	57 [34–68]	59 [50–69]
CVP (mmHg)	12 [8–18]	10 [6–13]
PI	–	1 [1–3]

HR, heart rate; RR, respiratory rate; SpO_2, peripheral oxygen saturation; MAP, mean arterial pressure; CVP, central venous pressure; PI, perfusion index

Chapter 1: Labelling Haemodynamic Instability

As mentioned, one of the constraints of the labelling method employed by van Winden et al. was that it did not account for medication dosage. This led to errors in the labelling, as the continuation or reduction of administered medication was registered the same as an increase, resulting in false labels as we are only interested in increases. In this chapter we propose a new method for labelling instability that quantifies the haemodynamic support received by the patient. The new labelling method is compared to the previous approach, and the effect of expanding the study cohort is examined.

1.1 Methods

One existing method of quantifying a patient's need for cardiovascular support is the vasoactive inotropic score (VIS)[14]. The VIS sums the dosage of various vasopressors and inotropes, weighting each according to its pharmacological potency, resulting in Formula 1.

$$\text{VIS} = \sum_{i=1}^{m} w_i \cdot d_i \tag{1}$$

Where w_i is the weighting factor for vasoactive/inotropic agent i, d_i is the dosage in μg/kg/min, and m is the total number of vasoactive or inotropic agents included in the score.

Since its introduction, several additional drugs have been incorporated into the formula, resulting in the complete list of weights shown in Table 2[15–18].

Table 2: Weight factor used in VIS score calculation for various drugs.

Drug	**Weight (w)**
Dopamine	1
Dobutamine	1
Enoximone	1
Phenylephrine	10
Milrinone	10
Olprinone	10
Levosimendan	50
Epinephrine	100
Norepinephrine	100
Vasopressin	10,000

Traditionally, the VIS score has been used to predict patient morbidity and mortality after cardiac surgery[19]. For our purpose of defining interventions, we expanded upon this method by adding fluid administrations and the application of extracorporeal membrane oxygenation (ECMO) in order to capture all methods of intervention in cases of haemodynamic instability. Fluid administration included sodium chloride, Ringer's lactate, and blood products, and was weighted using the administered volume by body weight in ml/kg. The application of ECMO was assigned a set weight of 150, as it is considered the last-resort intervention in cases of haemodynamic deterioration, and therefore a high weight was appropriate. This combined scoring resulted in what we referred to as the intervention score (IS), shown in Formula 2.

$$\text{IS} = \text{VIS} + \text{FA} + \text{ECMO} \tag{2}$$

Where VIS is the VIS score calculated according to Formula 1, FA is the administered fluid volume relative to body weight in ml/kg, and ECMO represents the administration of ECMO, assigned a fixed weight of 150. The IS serves as a comprehensive measure of all haemodynamic support a patient is receiving. An increase in the IS is interpreted as an intervention responding to haemodynamic instability. The dynamic and continuous nature of the IS enables distinguishing between different levels of intervention severity by applying a threshold to the IS increase considered a relevant intervention. The minimum increase in IS required for an intervention to be classified as indicative of instability is denoted as ΔIS_{min}. An increase in haemodynamic support serves as a response to instability. However, because instability is often not recognised immediately, it is assumed that a time period precedes the intervention during which the patient is already unstable. This delay between the onset of instability and the intervention is referred to as dT. An illustration of all algorithm and labelling parameters is provided in Figure 3.

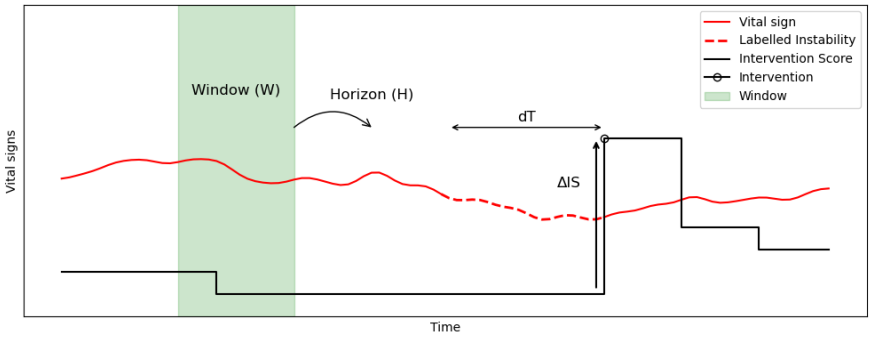


Figure 3: Illustration of the model input and labelling framework. Vital signs are observed in a time window (W), which is used to predict haemodynamic instability at a future time point. How many minutes into the future the prediction is made is dependant on the set prediction horizon (H). Instability is labelled retrospectively using increases in the intervention score (ΔIS). If ΔIS is larger than the set threshold ΔIS_{min} it is marked as an intervention ($\oplus$). The delay dT represents the assumed time between the onset of instability and the clinical response, during which the patient is labelled as unstable.

1.1.1 New Labelling

The new labelling method was compared to the original method used by Van Winden et al. The same random forest classifier and nested cross-validation framework were applied to both labelling methods. The model was trained on the old cohort of 224 patients using both the original and the new IS-based labels to compare the two labelling methods. The IS-based labelling was then applied to the new enlarged cohort of 425 patients. For these experiments, the following labelling and algorithm parameters were used: $W = 50$ min, $H = 45$ min, $dT = 120$ min, and $\Delta IS_{min} = 0$ (only for the IS-based labelling).

1.1.2 Varying dT and ΔIS$_{\text{min}}$

The IS-based labelling method allows flexibility in defining what constitutes a relevant intervention. Rather than treating every increase in IS as an intervention, the minimum required increase ΔIS_{min} can be raised to be more selective in what is considered a relevant intervention. To examine how this affects model performance, three values of ΔIS_{min} were tested: 0, 4, and 8. Appendix A.3 explains why these thresholds were chosen. For every ΔIS_{min}, dT was varied to reassess the effect of this parameter when we set higher thresholds for ΔIS_{min}.

To ensure that observed differences were due to the labelling strategy rather than other modelling choices, the prediction horizon, H, was fixed at 1 minute, which provided the best baseline performance. The window size, W, was kept constant at 50 minutes.

The baseline AUCPR is influenced by the class prior, meaning that changes in the labelling of instability—such as changing ΔIS_{min} and dT—alter the class balance and thus shift the baseline. This complicates comparisons between labelling strategies. To address this, negative samples were undersampled until the class distribution was balanced at 1:1 across all definitions of instability.

After this evaluation, the labels that produced the best performance were used to train a new model with a prediction horizon of 45 minutes.

Feature importance was assessed using the mean decrease in impurity (MDI), calculated as the total reduction in Gini impurity across all splits. This measure reflects the contribution of each feature to the model's predictions. For interpretability, feature importances were aggregated across all time points of a single vital parameter, so that cumulative importance values for each parameter were reported rather than importances for individual minutes.

1.1.3 Impact of Population Size on Model Performance

The influence of population size on model performance was assessed by creating subsets of varying sizes. Subset sizes started at 50 patients and increased in steps of 50 up to the full cohort size of 425. For each subset size, patients were randomly drawn from the full population, and the model was retrained on each subset. This process was repeated five times for every subset size to account for variability due to random selection. The labelling settings with the best performance resulting from Section 1.1.2 were used for this evaluation.

1.2 Results

1.2.1 New Labelling

A full overview of the results of the new labelling and extended cohort is shown in Table 3. The IS-based labelling resulted in a much lower class prior of 6.6% compared to the old labelling class prior of 17.5%. The IS-based labelling also resulted in a lower AUCPR compared to the old method, with values of 0.189 (0.073) and 0.423 (0.074), respectively. The AUCROC of the old and new labelling was 0.766 (0.050) and 0.745 (0.074), respectively. When the new extended cohort was used, the class prior was 5.7% and the performance increased to an AUCPR of 0.233 (0.041).

Table 3: Performance of the random forest classifier with the original labels on the old cohort, with IS-based labels on the old cohort, and with IS-based labels on the new cohort.

	Old Cohort, Old Label *(n=224)*	**Old Cohort, New Label** *(n=224)*	**New Cohort, New Label** *(n=425)*
Class Prior	17.5%	6.6%	5.7%
AUCPR	0.423 (0.074)	0.189 (0.073)	0.233 (0.041)
AUCROC	0.766 (0.050)	0.745 (0.074)	0.833 (0.029)

Reported performance metrics are the mean (SD) of the outer CV folds. SD, standard deviation; CV, cross-validation; AUCPR, area under the precision-recall curve; AUCROC, area under the receiver operator characteristic curve.

1.2.2 Varying dT and ΔIS$_{min}$

The AUCPR of the models trained at different ΔIS_{min} with varying dT are shown in Figure 4. Increasing dT with $\Delta IS_{min} = 0$ led to a gradual increase in AUCPR, with the best performance of 0.606 (0.054) at dT = 120 minutes. The best overall result was achieved with $\Delta IS_{min} = 4$ and dT = 20 minutes, with an AUCPR of 0.624 (0.073). $\Delta IS_{min} = 8$ performed best with dT = 40 minutes, achieving an AUCPR of 0.605 (0.038).

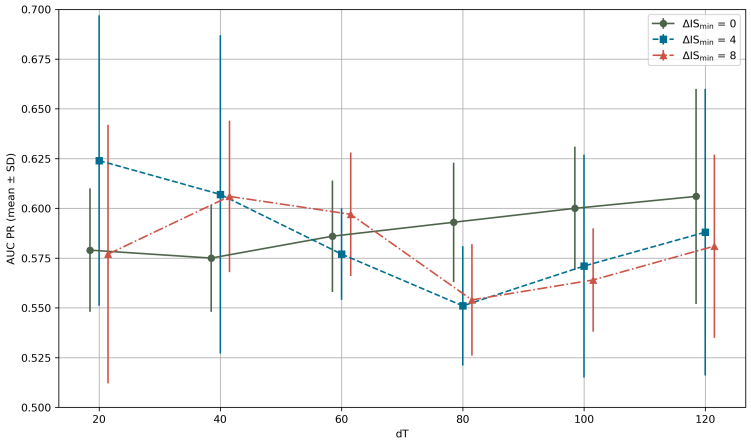


Figure 4: AUCPR as a function of evaluation window length (dT) for different thresholds of minimum required increase in instability score (ΔIS_{min}). Each curve shows the mean model performance across CV folds for ΔIS_{min} values of 0, 4, or 8, with the SD as error bars. To enable fair comparison across thresholds, the class balance was fixed by undersampling negative samples until a 1:1 ratio was achieved. AUCPR, area under the precision–recall curve; SD, standard deviation; CV, cross-validation.

The two best performing labelling techniques, $\Delta IS_{min} = 0$ with dT = 120 and $\Delta IS_{min} = 4$ with dT = 20, were used to train two models for the final test set. This resulted in AUCPR values of 0.204 and 0.043 and AUCROC values of 0.762 and 0.880, respectively. The class prior for ΔIS_{min} = 0 with dT = 120 was 4.19% in the test set, and at $\Delta IS_{min} = 4$ with dT = 20, the class prior was 0.81

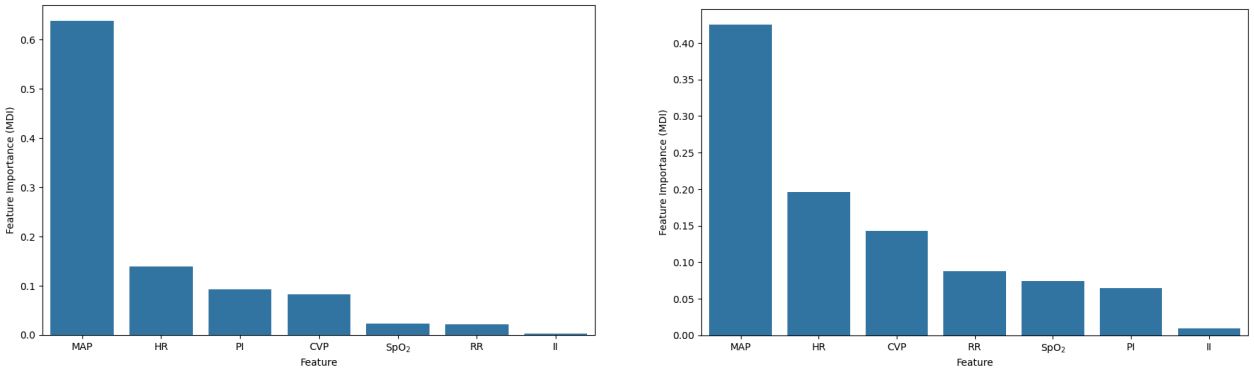
The cumulative feature importance for both models is shown in Figure 5. Both models scored MAP highest, followed by HR, while the imputation indicator scored lowest in both.

1.2.3 Impact of Population Size on Model Performance

The model performance across various population sizes is shown in Figure 6. $\Delta IS_{min} = 0$ and dT = 120 were used as labelling settings for the model training as they achieved the best performance in Section 1.2.2.

1.3 Interpretation of Results

The new labelling resulted in a much lower class prior compared to the old labelling, suggesting that the old method contained a lot of false labels. The results show that the AUCPR of our trained random forest using the new labelling method was lower compared to the old method. However, the drop in AUCPR can be explained by a lower class prior under the new IS-based labelling method, which set a lower baseline AUCPR. After new patients were added to the dataset, the average CV-



(a) $\Delta IS_{min} = 0$, dT = 120 **(b)** $\Delta IS_{min} = 4$, dT = 20

Figure 5: Cumulative feature importance score for the final model trained using different parameters for the labelling of haemodynamic instability. (a) Feature importance in the model trained using $\Delta IS_{min} = 0$, dT = 120. (b) Feature importance in the model trained using $\Delta IS_{min} = 4$, dT = 20. MAP, mean arterial pressure; HR, heart rate; PI, perfusion index; CVP, central venous pressure; SpO_2, peripheral oxygen saturation; RR, respiratory rate; II, imputation indicator; MDI, mean decrease in impurity.

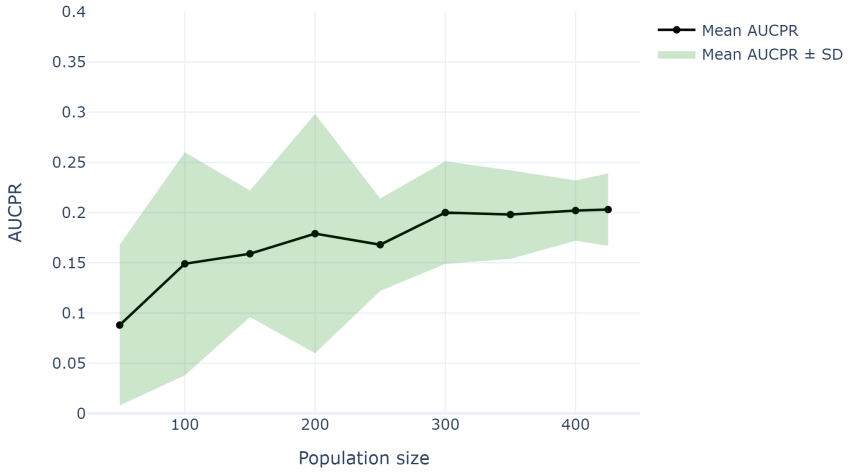


Figure 6: Model performance across population sizes. The mean AUCPR and standard deviation are shown for population subsets ranging from 50 to 425 patients. $\Delta IS_{min} = 0$, dT = 120, were used as labelling parameters. AUCPR, area under the precision–recall curve

fold performance increased and the SD of fold performance decreased, indicating that the enlarged cohort helped the model generalise better and reduced overfitting.

When comparing different ΔIS_{min} and varying dT, the effect on performance differed. With a higher threshold for ΔIS_{min}, performance increased when a shorter dT was used, contrary to what was observed at $\Delta IS_{min} = 0$, where performance was best at the highest dT setting. This may be attributed to more severe interventions, and thus larger increases in VIS score, being associated

with shorter, more acute episodes of haemodynamic deterioration. In contrast, smaller medication changes may have reflected longer, more gradual periods of deterioration. For $\Delta IS_{min} = 8$, the behaviour appeared to follow a similar trend to $\Delta IS_{min} = 4$, with the best performance at shorter dT, peaking at dT = 40. However, at dT = 20 performance decreased again, which could have been caused by the stricter instability parameters resulting in fewer positive labels and thus a smaller training dataset. It should be noted however that all differences in performance were very small and overall performance was poor. If anything the results demonstrate how it is difficult to pick a single dT which accurately captures all instability events.

The final test performance using $\Delta IS_{min} = 4$ and dT = 20 as labelling parameters achieved a higher AUCROC but a poor AUCPR. This is likely the result of the low class prior of the positive class hindering proper model training and leading to biased AUCROC results. The final $\Delta IS_{min} = 0$ and $dT = 120$ model performed as expected based on the nested CV results. Noteworthy is the model's heavy reliance on MAP for predictions, although not unexpected as blood pressure is the main contributor to adequate organ perfusion and often the driving factor when it comes to increasing haemodynamic support. Most of the interventions included in the IS also aim to increase blood pressure.

Reducing the population size resulted in decreasing performance. As population size increased, the performance gains showed diminishing returns, although the variance continued to decrease.

Chapter 2: Age-stratifed Subpopulations

In the experiments conducted by van Winden et al., both in-patient and inter-patient performance of the model were evaluated. Inter-patient refers to training the model on one set of patients and applying it to a separate set, while in-patient refers to training (partially) on data from a patient and then applying the model to new data from the same patient. The in-patient models achieved much better results, indicating high inter-patient variability. To address this, this chapter aims to reduce inter-patient variability by stratifying patients by age and asses how this affects model performance.

2.1 Method

2.1.1 Data Leakage Assessment

As an initial test to assess the impact of patient-specific patterns on performance using our IS-based labelling approach, we conducted an experiment in which cross-validation splits were no longer stratified by patient, thus allowing data from the same patient to *leak* between the training and test sets. This permitted segments of a single patient's data to appear in both sets. These tests were performed on both the original cohort used by Van Winden et al. and the new complete dataset. The following model parameters were used in this analysis: $W = 50$ min, $H = 45$ min, $dT = 120$ min, and $\Delta IS_{min} = 0$.

2.1.2 Age-stratified Models

Subgroup analysis was performed by dividing the population into quarter-year age groups, with neonates (0–30 days) as a separate age group resulting in five age groups: 0–30 days, 30–90 days, 90–180 days, 180–270 days, and 270–365 days. Neonates were used as a separate age group due to the high prevalence of this age group within our cohort and the rapid development of children during this early period. Based on the results in Figure 6 we chose to require a minimum of 100 patients per age group for model training. The models were trained using the following model parameters: $W = 50$ min, $H = 45$ min, $dT = 120$ min, and $\Delta IS_{min} = 0$.

2.2 Results

2.2.1 Data Leakage Assessment

The full results of the models trained with data leakage are shown in Table 4. The models resulted in a mean (SD) AUCPR of 0.824 (0.095) on the old data cohort, and an AUCPR of 0.681 (0.136).

2.2.2 Age-stratified Subpopulation

An overview of the results of the age-based subpopulation models be found in Table 5. Two of the defined age groups contained enough patients for model training, with 105 patients in the 0–30 day old group and 168 patients in the 90–180 day old group. The occurrence of instability differed between the two age populations. The class prior of instability was 9.2% for the 0–30 day group and 4.9% for the 90–180 day group. The 0–30 day subpopulation achieved an AUCPR of 0.244 (0.064), and the 90–180 day subpopulation achieved an AUCPR of 0.221 (0.105).

Table 4: Performance metrics of the old and new study cohort with model trained with non-patient stratified train and test splits.

	Old Cohort *(n=224)*	**New Cohort** *(n=425)*
Class Prior	6.6%	5.7%
AUCPR	0.824 (0.095)	0.681 (0.136)
AUCROC	0.983 (0.011)	0.968 (0.015)

Reported performance metrics are the mean (SD) of the outer CV folds. SD, standard deviation; CV, cross-validation; AUCPR, area under the precision-recall curve; AUCROC, area under the receiver operator characteristic curve.

Table 5: Performance metrics for models trained on age-based subpopulation fo 0–30 day and 90–180 days old.

	0–30 Days *(n=105)*	**90–180 Days** *(n=168)*
Class Prior	9.2%	4.9%
AUCPR	0.244 (0.064)	0.221 (0.105)
AUCROC	0.796 (0.106)	0.794 (0.077)

Reported performance metrics are the mean (SD) of the outer CV folds. SD, standard deviation; CV, cross-validation; AUCPR, area under the precision-recall curve; AUCROC, area under the receiver operator characteristic curve.

2.3 Interpretation of Results

Models trained with data leakage show drastic improvements in predictive performance, suggesting substantial variability in patient-specific patterns related to haemodynamic instability. The decrease in performance observed when using the full cohort may indicate that the larger dataset introduces greater variance in the population, also reflected in our patient characteristics in Table 5. However, a caveat of the method used for this test is that—for smaller populations—a larger proportion of the training data comes from the same patient to whom the model is being applied. For example, if a model is trained on 200 patients, then for any given test window roughly 1/200 of the training data comes from the same patient. If the cohort doubles to 400 patients, this overlap is reduced to 1/400. As a result, the performance drop seen in the full cohort may not reflect higher variance in the dataset but rather a reduction in patient-level leakage. The increase performance of the old cohort can thus be a result of the methodology used.

The two models trained on age-based subpopulations do not achieve better results than the model trained on the full study cohort. Both also exhibit high variance across the CV folds, likely due to the smaller population size within each subpopulation, limiting model generalisation. At this scale, there is no advantage to training models on age-based subpopulations. However, compared with non-age-stratified populations of a similar size, as shown in Figure 6, these models perform better than the AUCPR of 0.149 (0.111) and 0.159 (0.064) achieved with a sample size of 100 and 150 patients, respectively. With larger cohorts within the age categories, performance may improve and generalise if the trend of Figure 6 holds.

Chapter 3: Labelling Evaluation

In Chapter 1 the best labelling parameters were assessed. The method used there applied the same algorithmic setup as for the prediction of instability, except the prediction horizon, H, was set to one minute. The prediction algorithm relied on prior time windows as input data, meaning that—particularly at the start of an instability period—the model input consisted of time points outside of the labelled instability period. In order to asses the labelled period itself, we want to compare individual minutes within the labelled instability period rather than prior time windows, allowing us to directly evaluate whether unstable minutes can be distinguished from stable minutes using our labelling. Repeating this experiment for different labelling parameters for dT and ΔIS_{min} allows for a reevaluation of the labelling parameters.

Another aspect of the labelling requiring reevaluation is the assumption that an instability period ends immediately after an intervention. It is physiologically improbable that the patient will become stable immediately after an intervention yet the current labelling assumes so.

In this chapter, we tested whether unstable minutes can be distinguished from stable minutes when compared directly. The current labelling method was further examined by assessing both the distinction between stable and unstable minutes under different labelling settings and the assumption that instability ends immediately after intervention.

3.1 Method

3.1.1 Stable vs. Unstable

A classifier was trained to distinguish between stable and unstable samples. Each sample consisted of a single minute of data containing the measured vital parameters. Class balance was achieved by undersampling stable samples to obtain a 1:1 ratio. Training was repeated for different values of dT and ΔIS_{min}. dT was varied from 20 to 240 minutes in steps of 20 minutes, and for each dT setting, ΔIS_{min} values of 0, 4, and 8 were tested. A second-degree polynomial regression classifier was used, selected for its simplicity and its ability to capture basic non-linear relationships.

3.1.2 Pre- vs. Post-Intervention

To evaluate whether our assumption that instability is immediately over after intervention is reflected in the data a similar method to Section 3.1.1 was used. Instead of using randomly selected stable minutes we compare the minutes immediately post-intervention to the minutes pre-intervention. For each intervention ending a labelled period of instability, 120 minutes of data before and after were selected. The method of positive and negative samples selection ensured balanced classes. ΔIS_{min} was set at 0 for this experiment. A second-degree polynomial regression classifier.

3.2 Results

3.2.1 Stable vs. Unstable

Figure 7 shows the performance of the model across different dT and ΔIS_{min} settings. The performance The best performance was achieved using $dT = 180$ minutes and $\Delta IS_{min} = 8$ with an AUCPR of 0.789 (0.003).

3.2.2 Pre- vs. Post-Intervention

The classifier trained on pre- and post-intervention data resulted in an AUCROC of 0.571 ($\pm$0.006).

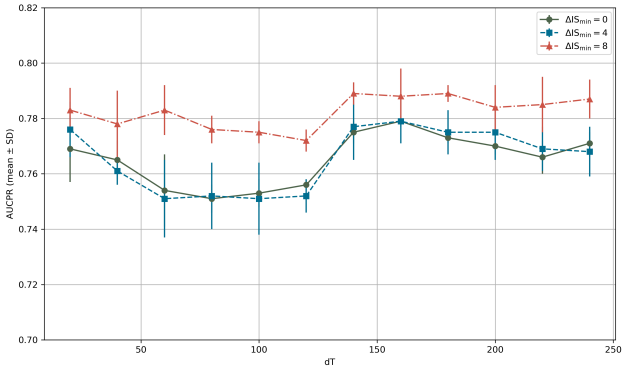


Figure 7: AUCPR (mean ± SD) of a second-degree polynomial regression classifier distinguishing stable from unstable minutes across different dT settings. Results are shown for ΔIS_{min} values of 0, 4, and 8.

3.3 Interpretation of Results

The algorithm was able to consistently achieve reasonable performance across different dT values, demonstrating we can distinguish between the unstable and stable labelled data. The model was still able to distinguish unstable from stable minutes even at longer dT, with performance maintained up to $dT = 240$. An unexplained but noticeable improvement in performance occurred between $dT = 120$ and $dT = 140$ across all ΔIS_{min} thresholds. Increasing ΔIS_{min} to 8 yielded clear improvements compared with the lower thresholds, which aligns with expectations that more substantial interventions correspond to more pronounced periods of haemodynamic instability, making them easier to distinguish.

The model showed limited ability to differentiate between pre- and post-intervention samples. This suggests that, although interventions stabilise patients, there is likely a period of adjustment during which medication takes effect. The onset and impact of medication also vary across the different medication included in the VIS. When compared with the $\Delta IS_{min} = 0$, $dT = 120$ results from Figure 7, performance was significantly higher in the stable vs. unstable experiment at 0.756 (0.002). Both experiments used the same setup, with the exception of the stable samples. In the stable vs. unstable experiment, stable samples were randomly selected from all available stable data and could therefore originate from periods without recent interventions. This supports the hypothesis that the period immediately after intervention cannot yet be considered stable.

Discussion & Conclusion

This study aimed to improve upon the prediction algorithm developed by van Winden et al. by improving the labelling of instability and test model performance on age-stratified subpopulation. A new labelling method was introduced, based on an intervention scoring system designed to capture all forms of haemodynamic support. The new IS-based labelling resulted in a significantly lower prevalence of instability compared to the old labelling suggesting the inaccuracies of the old labelling caused a lot of false labels now no longer present. The IS-based labelling performed best with parameters $\Delta IS_{min} = 0$ and $dT = 120$, yielding an AUCPR of 0.204 and an AUCROC of 0.880 on the test set of the new cohort. Retraining models on age-stratified subpopulations resulted in mean (SD) AUCPR values of 0.244 (0.064) and 0.221 (0.105)for the age groups 0–30 days and 90–180 days, respectively, and AUCROC values of 0.796 (0.106) and 0.794 (0.077). Although these results were comparable to those from the full cohort, the subpopulation models showed poorer generalisation, likely due to the smaller sample sizes.

Comparison to Literature

Existing research on the prediction of haemodynamic deterioration in the paediatric patients remains limited. The available studies differ substantially in their methodological approaches. In the following sections, key areas of divergence are examined and compared to the approach of this study, including the definition of haemodynamic instability, the selected features, and the choice of prediction horizon.

Most studies predicting deterioration in the PICU use more severe outcomes for labelling, such as mortality, cardiac arrest, or unplanned intubation[20–23]. Others adopt definitions similar to ours, relying on indicators such as fluid administration and medication changes[9,24]. Studies predicting clearly defined events—mortality, cardiac arrest, and unplanned intubation—have achieved good performance, with all achieving an AUCROC of ≥ 0.94. Potes et al. used intervention-based labelling and achieved an AUCROC of 0.81 when predicting interventions within the next hour[9]. Stein et al. also used intervention-based labelling similar to ours, achieving an AUCPR of 0.55 and an AUCROC of 0.95, although they used an even broader prediction window of 12 hours[24].

Our model was designed using a 50 minute data window making predictions 45 minutes into the future. Our models makes per minute prediction. In order to improve the current results using a broader prediction window can help, attempting to predict instability at some point within a time window instead of predicting instability exactly minute-by-minute. Stein et al. instead of using a moving window to make minute-by-minute predictions, designed their algorithm to predict deterioration at any point within the next 12 hours, based on data from the preceding 6 hours[24]. This provides a much broader time frame for prediction and bypasses some of the challenges we have faced in accurately labelling instability on a minute-by-minute basis. Although this approach may be less clinically relevant in (P)ICU wards where patients are closely monitored, it can still be valuable in hospital wards with less intensive care and a lower staff-to-patient ratio. In these settings, it helps identify high-risk patients who should be monitored closely over the next 12 hours.

Beyond the differences in prediction window Stein et al. also utilised a broader range of feature variables. They incorporated ventilator settings, the COMFORT score, and various laboratory values[24]. Their analysis identified bilirubin, creatinine, ion gap, and the COMFORT score, in addition to blood pressure metrics, as the most influential features for the prediction of instability. Potes et al. used 36 ICU measurements, with the Shock Index, pH, mean airway pressure, and normalised urine output ranking highest in their feature importance analysis[9]. Limiting their model to vital signs only reduced AUCROC from 0.77 to 0.71, illustrating the added value of

laboratory data and mechanical ventilation settings. Both studies outperformed our model, though methodological differences such as the broader prediction windows used limit direct comparison. Nevertheless, their results suggest that incorporating a broader range of features can be of added value to prediction performance.

Limitations

The biggest limitation of this study is the need for retrospective labelling of instability. The reliance on interventions without further insight into the status of the patient makes distinguishing clinically relevant periods of haemodynamic instability challenging. There are many considerations that go into the clinical decision-making surrounding an intervention, and it will not always be the result of haemodynamic instability. The retrospective labelling also required an estimation of the duration of instability preceding intervention, where a single duration was attributed to all periods of instability. This use of a fixed duration is a simplification, as the true length of instability events is likely to vary considerably between cases.

Another limiting factor arises from faults in the EHR registration. The analysis of subpopulations was restricted to age-based groups because many patients did not have clearly recorded diagnoses. With more than 50% of the database lacking a registered diagnosis, it was not possible to create diagnosis-based subpopulations for model training. The registration of administered interventions also showed inconsistencies, with limited clarity on how certain interventions were administered. This may have resulted in some false labels.

Finally, the use of a random forest introduced a limitation in eligible patients, as random forests cannot handle missing features. This reduced the patient population by 25%. Excluding patients with missing parameters also introduced bias into the dataset. For example, difficulty with arterial line placement is associated with a worse patient condition[25].

Future Directions

In order to improve labelling accuracy, it would be valuable to prospectively record haemodynamic instability events of interest as they occur. Having access to a couple of clearly defined instability events would provide clearer insight into how such events present themselves. Analysing common characteristics of these instability periods, such as specific medication types used as intervention, the duration of instability, or increases in IS, could help define more selective and clinically meaningful labelling criteria to be used for retrospective labelling in the future.

For further labelling improvements, it is also important to consider how current labelling choices influence model training. In Chapter 1 and Chapter 3 two different approaches were used to assess the best dT. In both cases, results were similar across all values of dT. This shows that there is likely a large variation between the duration of instability events and applying one value for dT does not capture all accurately. The current algorithm punishes itself during training as a result of the incorrect labelling due to having to pick a single dT. For example an instability with a duration of 90 minutes is labelled as having a 120 minute duration. The algorithm is now penalised during the first 30 minutes if it predicts no instability, even though it is correct in reality, because our labelling falsely says the instability has already started. To avoid this the definition of a true prediction can be adapted by considering any positive prediction during a consecutively labelled instability period as a correct prediction for that entire instability period. This way the algorithm does not penalise itself during training for late predictions. From a clinical perspective a late prediction is still valuable, even if only the last minute within a labelling period is predicted positive, the prediction horizon still gives the algorithm a 45-minute lead time. A post hoc analysis (Appendix

A.5), in which any positive prediction within a labelled period was counted as correctly identifying the entire event, improved AUCPR to 0.32 and AUCROC to 0.79 without retraining the model.

Similarly, a positive prediction before the labelled instability period is currently considered a false positive, even though it may represent correct early detection of an instability which lasted longer than the assigned dT. Our current method also assumes that patients stabilise immediately after intervention. In reality, the onset of action varies between interventions, and there is likely a stabilisation period while medication takes effect. In Chapter 3 we have shown that it is hard to distinguish between unstable time points and time points immediately after intervention, indicating that the assumption that this period can be considered stable is false.

Ghorbani et al. developed Proximity-Aware Time Series Anomaly Evaluation (PATE)[26], a metric that addresses such issues by introducing buffer zones before and after anomalies (in our data an anomaly would correspond to a period of instability) and calculating AUC metrics using weighted true and false positives based on their temporal proximity to the anomaly. Considering the post-intervention period as neither stable or unstable during evaluation is also an option. Adopting such evaluation strategies and retraining could yield better performing prediction algorithms while still aligning with clinical application.

Beyond labelling and evaluation, feature selection additional features can aid in improving model performance. Our feature importance analysis in Chapter 1 showed that MAP provided greater discriminatory value than all other features combined, suggesting some features may be obsolete. Features currently not included in our algorithm such as laboratory measurements and ventilatory settings have been shown by other studies to improve prediction of haemodynamic instability. Incorporating these features will provide a more comprehensive view of the patient's condition and improve predictive performance.

Adding more features, particularly laboratory values and mechanical ventilation settings, will introduce additional missing data into the dataset, as some patients will not be receiving mechanical ventilation and certain laboratory tests are not routinely performed on all patients. To avoid discarding patients without a complete feature set or resorting to extensive imputation, using an algorithm capable of handling missing data should be considered. XGBoost is a decision tree–based algorithm which can natively handle missing data and exploit patterns in the missingness to improve predictions[27]. In the study by Stein et al., XGBoost also achieved the best performance, outperforming recurrent neural networks and logistic regression in predicting interventions[24].

As discussed in the literature comparison an entirely different application of the predictive model would be to shift the focus from minute-by-minute forecasts to identifying patients at elevated risk of haemodynamic instability within a broader time frame. A post hoc analysis was performed (Appendix A.6) in which we used a positive prediction within the first 2 hours of a patient admission to label patients as unstable or stable for the next 10 hours. This resulted in an AUCPR of 0.621 and an AUCROC of 0.800. A caveat of this approach that some of the vital parameters used for our model are not measured continuously in medium care wards, so adaptation of the features used to train the model is necessary.

Expanding the study cohort may not substantially improve performance for the overall patient population, but it would provide larger age-based subpopulations, which could enhance both their predictive performance and generalisation.

Conclusion

This study proposes a new method of labelling haemodynamic instability based on interventions using the developed IS. However, the random forest model trained with the IS-based labelling, aimed at predicting haemodynamic instability, did not show substantial improvement, achieving an AUCPR of 0.204 and an AUCROC of 0.762.

Analysis using age-based subpopulations with the goal of reducing variation in the data showed promising results, although a larger population is needed to improve performance and enhance generalisation.

Further analysis of the labelling revealed that some assumptions about the start and end of periods of instability need to be revisited in order to improve labelling in the future before a viable prediction model can be developed.

References

[1] van der Bom T, Zomer AC, Zwinderman AH, Meijboom FJ, Bouma BJ, Mulder BJM. The changing epidemiology of congenital heart disease. Nature Reviews Cardiology. 2010;8(1):50-60. Available from: https://doi.org/10.1038/nrcardio.2010.166.

[2] Reller MD, Strickland MJ, Riehle-Colarusso T, Mahle WT, Correa A. Prevalence of Congenital Heart Defects in Metropolitan Atlanta, 1998-2005. The Journal of Pediatrics. 2008;153(6):807-13. Available from: https://doi.org/10.1016/j.jpeds.2008.05.059.

[3] van der Linde D, Konings EEM, Slager MA, Witsenburg M, Helbing WA, Takkenberg JJM, et al. Birth Prevalence of Congenital Heart Disease Worldwide. Journal of the American College of Cardiology. 2011;58(21):2241-7. Available from: https://doi.org/10.1016/j.jacc.2011.08.025.

[4] Dastgiri S. Prevalence and secular trend of congenital anomalies in Glasgow, UK. Archives of Disease in Childhood. 2002;86(4):257-63. Available from: https://doi.org/10.1136/adc.86.4.257.

[5] Oster ME, Lee KA, Honein MA, Riehle-Colarusso T, Shin M, Correa A. Temporal Trends in Survival Among Infants With Critical Congenital Heart Defects. Pediatrics. 2013;131(5):e1502-8. Available from: https://doi.org/10.1542/peds.2012-3435.

[6] Giesinger RE, McNamara PJ. Hemodynamic instability in the critically ill neonate: An approach to cardiovascular support based on disease pathophysiology. Seminars in Perinatology. 2016;40(3):174-88. Available from: https://doi.org/10.1053/j.semperi.2015.12.005.

[7] Dung-Hung C, Cong T, Zeyu J, Yu-Shan OY, Yung-Yan L. External validation of a machine learning model to predict hemodynamic instability in intensive care unit. Critical Care. 2022;26(1). Available from: https://doi.org/10.1186/s13054-022-04088-9.

[8] Rahman A, Chang Y, Dong J, Conroy B, Natarajan A, Kinoshita T, et al. Early prediction of hemodynamic interventions in the intensive care unit using machine learning. Critical Care. 2021;25(1). Available from: https://doi.org/10.1186/s13054-021-03808-x.

[9] Potes C, Conroy B, Xu-Wilson M, Newth C, Inwald D, Frassica J. A clinical prediction model to identify patients at high risk of hemodynamic instability in the pediatric intensive care unit. Critical Care. 2017;21(1). Available from: https://doi.org/10.1186/s13054-017-1874-z.

[10] van Winden B. Predicting Haemodynamic Instability in Critical Congenital Heart Disease Patients: A Proof of Concept; 2024. Unpublished work.

[11] Fleming S, Thompson M, Stevens R, Heneghan C, Plüddemann A, Maconochie I, et al. Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies. The Lancet. 2011;377(9770):1011-8. Available from: https://doi.org/10.1016/s0140-6736(10)62226-x.

[12] Haque IU, Zaritsky AL. Analysis of the evidence for the lower limit of systolic and mean arterial pressure in children. Pediatric Critical Care Medicine. 2007;8(2):138-44. Available from: https://doi.org/10.1097/01.pcc.0000257039.32593.dc.

[13] Chen Y, Wu L, Zou L, Li G, Zhang W. Update on the birth weight standard and its diagnostic value in small for gestational age (SGA) infants in China. The Journal of Maternal-Fetal & Neonatal Medicine. 2016;30(7):801-7. Available from: https://doi.org/10.1080/14767058.2016.1186636.

[14] Gaies MG, Gurney JG, Yen AH, Napoli ML, Gajarski RJ, Ohye RG, et al. Vasoactive–inotropic score as a predictor of morbidity and mortality in infants after cardiopulmonary bypass*. Pediatric Critical Care Medicine. 2010;11(2):234-8. Available from: https://doi.org/10.1097/pcc.0b013e3181b806fc.

[15] Favia I, Vitale V, Ricci Z. The Vasoactive-Inotropic Score and Levosimendan: Time for LVIS? Journal of Cardiothoracic and Vascular Anesthesia. 2013;27(2):e15-6. Available from: https://doi.org/10.1053/j.jvca.2012.11.009.

[16] Nguyen HV, Havalad V, Aponte-Patel L, Murata AY, Wang DY, Rusanov A, et al. Temporary biventricular pacing decreases the vasoactive-inotropic score after cardiac surgery: A substudy of a randomized clinical trial. The Journal of Thoracic and Cardiovascular Surgery. 2012;146(2):296-301. Available from: https://doi.org/10.1016/j.jtcvs.2012.07.020.

[17] Landoni G, Lomivorotov VV, Alvaro G, Lobreglio R, Pisano A, Guarracino F, et al. Levosimendan for Hemodynamic Support after Cardiac Surgery. New England Journal of Medicine. 2017;376(21):2021-31. Available from: https://doi.org/10.1056/nejmoa1616325.

[18] Yamazaki Y, Oba K, Matsui Y, Morimoto Y. Vasoactive-inotropic score as a predictor of morbidity and mortality in adults after cardiac surgery with cardiopulmonary bypass. Journal of Anesthesia. 2018;32(2):167-73. Available from: https://doi.org/10.1007/s00540-018-2447-2.

[19] Koponen T, Karttunen J, Musialowicz T, Pietiläinen L, Uusaro A, Lahtinen P. Vasoactive-inotropic score and the prediction of morbidity and mortality after cardiac surgery. British Journal of Anaesthesia. 2019;122(4):428-36. Available from: https://doi.org/10.1016/j.bja.2018.12.019.

[20] Aczon MD, Ledbetter DR, Laksana E, Ho LV, Wetzel RC. Continuous Prediction of Mortality in the PICU: A Recurrent Neural Network Model in a Single-Center Dataset*. Pediatric Critical Care Medicine. 2021;22(6):519-29. Available from: https://doi.org/10.1097/pcc.0000000000002682.

[21] Lee B, Kim K, Hwang H, Kim YS, Chung EH, Yoon JS, et al. Development of a machine learning model for predicting pediatric mortality in the early stages of intensive care unit admission. Scientific Reports. 2021;11(1). Available from: https://doi.org/10.1038/s41598-020-80474-z.

[22] Kim SY, Kim S, Cho J, Kim YS, Sol IS, Sung Y, et al. A deep learning model for real-time mortality prediction in critically ill children. Critical Care. 2019;23(1). Available from: https://doi.org/10.1186/s13054-019-2561-z.

[23] Rusin CG, Acosta SI, Vu EL, Ahmed M, Brady KM, Penny DJ. Automated Prediction of Cardiorespiratory Deterioration in Patients With Single Ventricle. Journal of the American College of Cardiology. 2021;77(25):3184-92. Available from: https://doi.org/10.1016/j.jacc.2021.04.072.

[24] Stein DF, Carter MJ, Booth J, Peters MJ, Ray S, Sebire NJ, et al. Predicting Cardiovascular deterioration in a paediatric intensive care unit (PicEWS): a machine learning modelling study of routinely collected health-care data. eClinicalMedicine. 2025;85:103255. Available from: https://doi.org/10.1016/j.eclinm.2025.103255.

[25] Yanko FM, Rivera A, Cheon EC, Mitchell JD, Ballard HA. Patient and Technical Factors Associated with Difficult Arterial Access and Ultrasound Use in the Operating Room. Children. 2023;11(1):21. Available from: https://doi.org/10.3390/children11010021.

[26] Ghorbani R, Reinders M, Tax D. PATE: Proximity-Aware Time Series Anomaly Evaluation. In: KDD '24. United States: ACM; 2024. p. 872-83. 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024 ; Conference date: 25-08-2024 Through 29-08-2024. Available from: https://kdd2024.kdd.org/.

[27] Chen T, Guestrin C. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016;11:785-94. Available from: http://dx.doi.org/10.1145/2939672.2939785.

Appendix

A.1 Preprocessing

Artefact removal was performed by identifying and excluding physiologically impossible values. Specifically, values equal to zero were replaced with missing values across all parameters except for heart rate, where a value of zero may represent a valid reading during asystole. Additionally, any negative values in pressure measurements were considered artefactual and were similarly replaced with missing values.

Next, last observation carried forward imputation was applied to fill in missing data in the measurements. A binary indicator was added to track whether any of the vital parameters had been imputed.

The vital parameters were recorded at varying frequencies. To standardise the frequency across parameters and patients—and to reduce computational demand—all parameters were downsampled to one per minute. The mean was used during downsampling for all parameters except the binary imputation indicator, for which the median was more appropriate due to its binary nature.

The duration of stay varies between patients. To ensure the algorithm is trained on equal length data from each patient, we select 800 consecutive minutes for each patient. In order to use as much real data as possible the 800 minutes are selected with the least imputations. The three data sections at the start, middle, and end of the admission are selected and the imputation indicator is used to select the period with the most real data.

The variability between patients extends beyond variability in length of stay. The baseline value of the vital parameters in our algorithm also varies based on age and pathophysiology. To correct for this, we scale each parameter using Scikit-learn's RobustScaler function. RobustScaler removes the median and scales the data according to the interquartile range, which makes it less sensitive to outliers compared with methods such as standardisation or min–max scaling. This ensures that extreme values do not disproportionately influence the distribution of the scaled features. The data scaler is created based on the first hour of the train and test set and applied to those two respectively.

A.2 Explanation of Hyperparameters

Hyperparameters are model parameters for na machine learning model which are not learned from the data itself. They must be chosen before training and can strongly influence both predictive performance and computational efficiency. For Random Forests, hyperparameters regulate aspects such as the number of trees, the maximum depth of each tree, how many features are considered at each split, and how splits are evaluated. Table A.2.1 shows the hyperparameters tuned during our nested cross-validation setup with a brief explanation of the parameter.

Table A.2.1: Overview of random forest hyperparameters tuned during nested cross-validation.

Hyperparameter	**Explanation**
`n_estimators`	Number of trees in the forest. More trees reduce variance at the cost of training time.
`max_depth`	Maximum depth of each tree. Controls model complexity; shallow trees prevent overfitting, while deeper trees capture more structure.
`max_features`	Number of features considered when looking for the best split. Smaller values increase tree diversity, larger values reduce bias.
`min_samples_split`	Minimum number of samples required to split an internal node. Higher values make trees more conservative and reduce overfitting.
`max_leaf_nodes`	Maximum number of leaf nodes per tree. Restricts growth and can improve generalisation.
`criterion`	Function used to measure the quality of a split. `gini` is slightly faster, `entropy` may be more informative.
`class_weight`	Adjusts weights inversely to class frequencies to handle imbalance.

A.3 ΔIS Threshold Choice

To determine suitable thresholds for labelling significant increases in IS, the distribution of all IS increases across all patients was examined. Figure A.3.1 shows a histogram of the frequency of observed IS increases across all data points. The distribution is highly skewed, with the majority of increases being small and only a few large jumps occurring. Based on this distribution, two thresholds were selected to define meaningful IS increases: a lower threshold of $\Delta IS = 4$ to capture more frequent but still notable changes, and a higher threshold of $\Delta IS = 8$ which ignores the bulk of the smaller increases we observe.

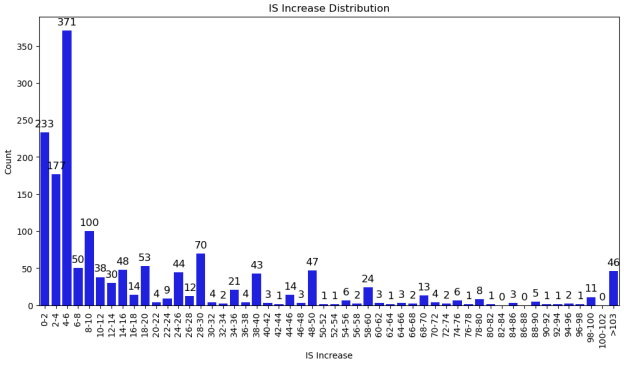


Figure A.3.1: Histogram showing the distribution of IS increase values across the dataset. This figure was used to inform the choice of thresholds for identifying significant increases in IS. IS, Intervention Score

A.4 Table of Full Varying dT and $\Delta IS_{\min}$ Results

The complete results for all combinations of dT and $\Delta IS_{\min}$ are provided in Table A.4.1. These values correspond to the experiments described in Section 1.2.2, where the effect of varying dT under different $\Delta IS_{\min}$ thresholds was analysed. Table A.4.1 reports the mean AUCPR and SD across CV folds for every setting.

Table A.4.1: Mean (SD) AUCPR across CV folds for $\Delta IS_{\min}$ values of 0, 4, or 8, whilst varying dT.

dT	$\Delta IS_{\min} = 0$	$\Delta IS_{\min} = 4$	$\Delta IS_{\min} = 8$
20	0.579 (0.031)	0.624 (0.073)	0.577 (0.065)
40	0.575 (0.027)	0.607 (0.080)	0.606 (0.038)
60	0.586 (0.028)	0.577 (0.023)	0.597 (0.031)
80	0.593 (0.030)	0.551 (0.030)	0.554 (0.028)
100	0.600 (0.031)	0.571 (0.056)	0.564 (0.026)
120	0.606 (0.054)	0.588 (0.072)	0.581 (0.046)

A.5 Event Level Detection Performance

We conducted a post hoc event-level analysis to complement per-minute evaluation metrics. y_{true} denotes the ground-truth labels, assigned using IS-based labelling. Correspondingly, y_{pred} denotes the binary predictions of the model and y_{prob} the associated continuous prediction scores. A deterioration event was defined as a continuous block of $y_{true} = 1$, which corresponds to the pre-intervention interval. Each such block was treated as a single event regardless of its length.

An event was considered correctly predicted, and thus a true positive, if $y_{pred} = 1$ occurred at least once within the first 120 minutes of the event block. If no positive prediction was present in this interval, the event was classified as a false negative. Outside of event blocks, continuous sequences of $y_{pred} = 1$ were grouped together and counted as a single false positive, thereby avoiding inflation of false alarms due to successive positive predictions. Negative periods without any positive prediction were considered true negatives. From these definitions, we obtained event-level counts of true positives, false negatives, false positives, and true negatives, and used these to compute sensitivity, specificity, precision, and the F1-score.

To assess threshold-independent performance, event-level AUCPR and AUCROC were also calculated using y_{prob}. For each positive event, the maximum value of y_{prob} within the 120-minute detection window was taken as the representative event score. For each negative period that contained at least one positive prediction, the maximum y_{prob} within that block was used. These event scores, combined with their corresponding binary event labels, provided the basis for calculating AUCPR and AUCROC at the event level.

We applied this method to the final test set using $\Delta IS_{min} = 0$, $dT = 120$ as labelling settings. The model was not retrained for this analysis only the performance metrics were recalculated. This resulted in an AUCPR of 0.36.

A.6 Patient Level Instability Prediction

Ad a post hoc patient-level analysis was conducted to see whether we can label stable or unstable patients based on only the start of 12 hours of data. y_{true} denotes the ground-truth labels, assigned using IS-based labelling. Correspondingly, y_{pred} denotes the binary predictions of the model and y_{prob} the associated continuous prediction scores. A patient was considered unstable if $y_{true} = 1$ occurred at any point during the full observation period, and stable if no positives occurred.

To make an early classification, we restricted evaluation to the first two hours of each patient's data. The patient-level prediction was defined as positive if $y_{pred} = 1$ occurred at least once within this early window. For the purpose of threshold-independent evaluation, the maximum value of y_{prob} within the same early window was taken as the representative patient-level score. These patient-level scores and labels were then used to calculate AUCPR and AUCROC.

We applied this method to the final test set using $\Delta IS_{min} = 0$, $dT = 120$ as labelling settings. The model was not retrained for this analysis only the performance metrics were recalculated. This resulted in an AUCPR of 0.621 and an AUCPR of 0.800.