Automated Early Warning of Acute Heart Failure in *ex vivo* Heart Evaluation Using Hemodynamic Flags

Hanna Nilsson[1], Erik Steen[1], Kristian Soltesz[1], Henry Pigot[2]

[1] Department of Automatic Control, Lund University, Lund, Sweden
[2] Artificial Intelligence and Bioinformatics in Cardiothoracic Sciences, Department of Translational Medicine, Lund University, Lund, Sweden

Abstract

In ex vivo *heart evaluation, an explanted heart is kept beating outside the body for functional assessment. Hearts may develop acute heart failure (AHF), requiring prompt flow reduction; detection currently relies on continuous visual monitoring. We aimed to automate AHF detection in a porcine* ex vivo *system. Data from 13 experiments were split chronologically into training (10 experiments, 29 events) and test (3, 5 events) sets. Of seven recorded signals, four became beat-by-beat features, each expressed as its short-term deviation from a causal baseline. Three control signals formed a composite flag. A detection triggered when the weighted sum of active flags reached a predefined count. Three variants—all-feature, reduced (heart rate and left atrial pressure only), and weighted (these two double-weighted)—were compared. The weighted model generalized best, with F_2 of 0.80 (training) and 0.65 (test), detecting 90% of training events with roughly one false positive per true positive. On extended datasets simulating continued operation, F_2 improved to 0.85 and 0.76, with all test events detected. Median detection latency was ~ 4 s before intervention. Automated early warning therefore appears feasible.*

1. Introduction

Globally, only about one-third of hearts from brain-dead donors are transplanted, despite the large number of patients on transplant waiting lists [2]. Some donor hearts are considered marginal and do not meet adopted inclusion criteria due to age or other risk factors. *Ex vivo* heart evaluation could assess the functional quality of such hearts before transplantation. Building on advances in non-ischemic heart preservation [3, 4], an *ex vivo* perfusion system is under development to evaluate marginal donor hearts and assess their suitability for transplantation (Fig. 1) [1]. A key step toward clinical implementation is automation.

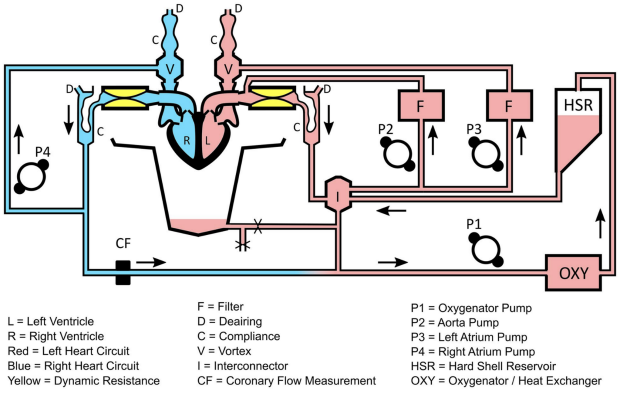

Figure 1. Schematic of the *ex vivo* perfusion system, reproduced from [1] under CC BY-NC 4.0. The left and right heart circuits are perfused separately; the pressures used in this study are measured on the left side.

During *ex vivo* evaluation, hearts may develop acute heart failure (AHF), which reduces pumping efficiency and causes backfilling in the perfusion system. To protect the heart and the system, the input flow is manually reduced until normal rhythm is restored. This intervention currently relies on continuous visual monitoring. An automated warning system would reduce this dependence on human oversight.

This work investigates whether AHF events can be detected—and ideally anticipated—automatically from hemodynamic signals available in the system. We extract beat-by-beat features motivated by cardiac physiology, build a simple flag-based detection model with three variants of increasing emphasis on the most informative signals, and evaluate them offline using causal processing. To our knowledge, no automated early warning for AHF during *ex vivo* cardiac perfusion has been reported. The study is limited to a porcine model and to signals from the left side of the heart, which dominates systemic circulation.

2. Methods

2.1. Data and pre-processing

Data from 13 porcine heart experiments were analyzed. All experiments were approved by the Ethics Committee for Animal Research at Lund University (approval Nos. 5.8.18-22454/2021 and 5.8.18-16949/2025) and conducted in accordance with Directive 2010/63/EU. Porcine hearts were used due to their anatomical and hemodynamic similarity to human hearts [5].

Signals were sampled at 250 Hz (200 Hz in some experiments; all analysis was time-based) and included aortic pressure (AP), coronary flow (CF), left ventricular pressure (LVP), and left atrial pressure (LAP). Control signals—left pump flow (LP), and left diastolic and systolic control pressures (LDCP, LSCP)—were recorded at 1 Hz. Each recording was trimmed to exclude startup and shutdown phases, long missing segments were removed, remaining gaps were linearly interpolated, and LVP, LAP, AP, and CF were zero-phase low-pass filtered (second-order Butterworth, 30 Hz cutoff).

The data were split chronologically: the first 10 experiments (29 events, 17 h 40 min) for training and the 3 most recent (5 events, 6 h 55 min) for testing. After each event, the subsequent 60 s were removed. This mirrors how the warning system would operate in practice: once a detection is made, the system is paused while the failure is handled, and restarted once the heart is stable.

2.2. Beat-by-beat features

A beat-to-beat representation was constructed by detecting systolic peaks in LVP, with peaks arising from the dicrotic notch discarded to retain one peak per beat. For each beat we extracted left ventricular systolic and end-diastolic pressures (LVSP, LVEDP), the maximum rate of ventricular pressure rise ($dP/dt_{\max}$), LAP, aortic systolic and end-diastolic pressures (ASP, AEDP), and CF. Heart rate (HR) was derived from the interval between consecutive beats. Two composite features, aortic pulse pressure (APP $=$ ASP $-$ AEDP) and coronary perfusion pressure (CPP $=$ AEDP $-$ LVEDP), were also formed. These features reflect mechanisms associated with AHF: elevated filling pressures and HR, and reduced contractility and coronary perfusion [6–10].

Each feature was expressed as the difference between a short-term local estimate and a longer-term baseline estimate, both computed with causal median filters so that only past and current samples were used. The sign was chosen so that every feature increases during an event (e.g. contractility markers were inverted).

2.3. Labeling and feature selection

Events were labeled retrospectively from manual pump-flow reductions: a sharp drop in LP was flagged, and each candidate was retained only if the surrounding LVP showed irregularity indicative of failure, following consultation with experienced operators. Because some LVP-derived features were also used for detection, this introduces a potential optimistic bias. A point label (L_1) marked the time of intervention; an extended label (L_2) additionally marked the 10 s preceding each event, so that features could be evaluated without setting thresholds.

For feature selection, the areas under the ROC curve (AUROC) and under the precision–recall curve (AUPRC) were computed against L_2 over a grid of local and baseline window lengths, on event-centered excerpts spanning up to 120 s before each event. The selected windows were 1–5 s (local) and 45–120 s (baseline). HR and LAP were the most informative signals (AUPRC up to 0.48 and 0.59, against a positive-class prevalence of 0.12 in these excerpts), consistent with the known roles of tachycardia and elevated filling pressure in AHF. AEDP was dropped as a standalone feature, as no threshold achieved usable recall, but is retained within APP and CPP.

2.4. Flag-based detection model

Each feature was assigned a threshold. A flag activated when the feature exceeded its threshold for at least two consecutive beats, and remained active for 10 s (extended if the threshold was re-exceeded). A detection triggered when the number—or weighted sum—of simultaneously active flags reached a count N_{flags}; a 12 s persistence window merged temporally close detections (Fig. 2). The three control signals were combined by a logical OR into a single composite flag with a longer (120 s) duration, motivated by the observation that many events follow system adjustments.

Three models were compared. The *all-feature* model used all informative signals (HR, LAP, LVEDP, LVSP, $dP/dt_{\max}$, ASP, APP, CF, CPP) plus the control flag, each weighted equally. Feature thresholds were set automatically to reach a target recall on the training data. The *reduced* model used only HR and LAP, plus the control flag, with manually tuned thresholds. The *weighted* model used the same features as the all-feature model but double-weighted the two most informative, HR and LAP.

2.5. Datasets and evaluation

Because the data were truncated at intervention, an *extended* dataset was also created by extrapolating each feature and label 30 s beyond every event using a decayed linear trend, simulating continued operation without inter-

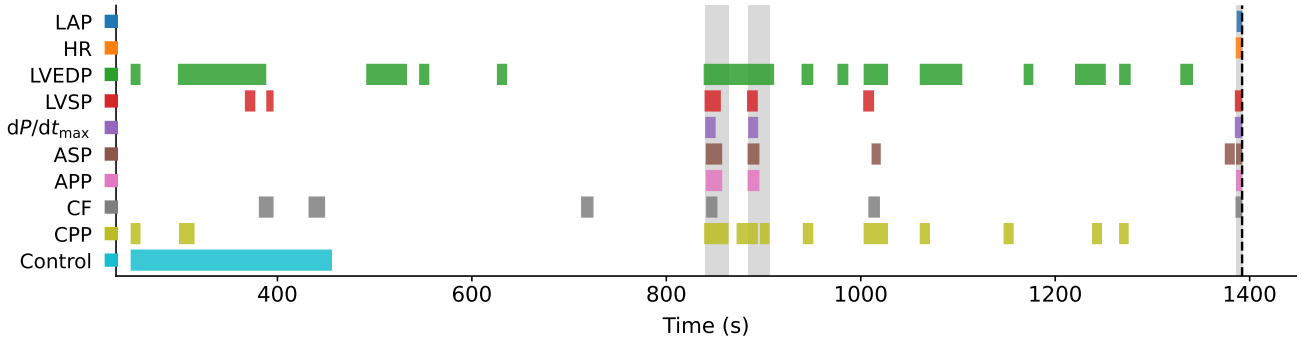

Figure 2. Example of the flagging mechanism. Each feature raises a flag when it exceeds its threshold (colored bars); a detection (grey) is triggered when the (weighted) number of simultaneously active flags reaches N_{flags}, here 6 with nine features plus the control flag. The dashed line marks the manual intervention; the two detections near 850 s are false positives.

Table 1. F_2 score for each model on the full and extended datasets. Training columns give the maximum over the $(N_{\text{flags}}, \alpha)$ grid; test columns give F_2 at the train-optimal hyperparameters. Best test value in bold.

	Full		Extended	
Model	Train	Test	Train	Test
All-feature	0.74	0.59	0.81	0.74
Reduced	0.79	0.36	0.85	0.67
Weighted	0.80	**0.65**	0.85	**0.76**

vention. Models were evaluated on the full and extended datasets, for both training and test.

Detections were scored by overlap with L_1: a true positive (TP) required overlap with an event, a false negative (FN) was a missed event, and a false positive (FP) was a detection outside any event. The F_β score,

$$F_\beta = (1+\beta^2)\,\frac{\text{Precision}\cdot\text{Recall}}{\beta^2\,\text{Precision}+\text{Recall}},$$

with $\beta = 2$, was the primary metric. It weights recall above precision because a missed event may damage the heart, whereas a false alarm mainly reduces usability [11, 12]. For each model, F_2 was computed over a grid of N_{flags} and a uniform threshold scaling factor α. Test sets were evaluated using the hyperparameters optimal on the corresponding training set. Detection latency, the time between the onset of L_1 and the detection, was analyzed for true positives.

3. Results

On the full training data all three models detected 86–90% of AHF events. In Table 1 the reduced and weighted models slightly outperformed the all-feature model, so adding features did not necessarily help. The weighted

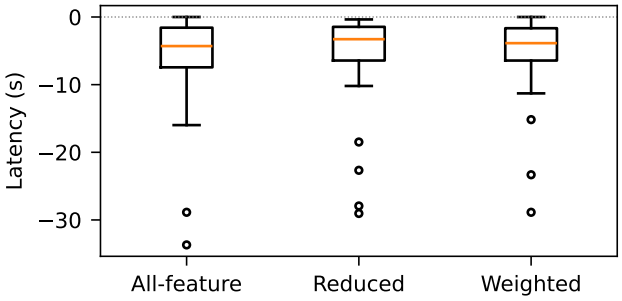

Figure 3. Detection latency relative to manual intervention on the full training dataset. Negative values denote detection before intervention; medians are near −4 s.

model achieved the highest F_2 on both test sets. On the training data it detected 90% of events (26 of 29) with roughly one false positive per true positive, about 1.2 false alarms per hour over the 17 h 40 min recording. On the full test set it detected 4 of 5 events with 7 false positives (1.0 per hour), as did the all-feature model with 10 false positives, whereas the reduced model detected 2.

The reduced model generalized poorly: F_2 dropped from 0.79 (train) to 0.36 (test) on the full data. Restricting the model to the two best-training features therefore appears to encourage overfitting. On the extended datasets all models improved and the weighted model detected all five test events; these numbers are optimistic, since the extrapolation lets features keep evolving and thereby eases detection.

Latency analysis (Fig. 3) showed median detections $\sim$ 4 s before manual intervention, occasionally more than 30 s early, with only small differences between models. On the extended datasets, where later detection was also possible, most latencies remained negative: events were typically anticipated rather than merely detected.

4. Discussion

The feature ranking aligned with known AHF physiology: HR and LAP, which rise during failure, were most informative, while contractility-related features ($dP/dt_{\max}$, LVSP) decreased as expected. AHF here therefore resembles *in vivo* failure in the signals considered, although the system provides no ECG or systemic information of the kind used by most clinical detection methods [13].

Among the variants, the weighted model generalized most consistently: up-weighting the most informative features appears to balance sensitivity and robustness better than discarding the others, as the reduced model does. The reduced model was also tuned differently (manual thresholds), so its weaker generalization cannot be attributed to the feature set alone. The latency results indicate that the approach can provide early warning, not only detection.

Several limitations temper these findings. The dataset is small, especially the test set (5 events), so model-ranking differences should be read as tentative. Labels were derived retrospectively from a surrogate (pump flow) and partly from LVP, which is also a feature source; this may introduce optimistic bias. Modeling was offline; although all processing was causal, online behavior is not guaranteed. Results on the extended datasets are optimistic by construction. The findings are therefore exploratory. The most important next step is online validation, ideally of both the reduced and weighted models, toward clinical *ex vivo* evaluation of human donor hearts.

5. Conclusions

A simple flag-based model detects AHF events in an *ex vivo* perfusion system before manual intervention, from beat-by-beat hemodynamic features derived from left-heart pressures and flow. The weighted model, which double-weights heart rate and left atrial pressure, generalized most consistently and anticipated the operator in most cases. The evidence is limited by a small test set and by surrogate labels; online validation is the necessary next step toward automated early warning in *ex vivo* heart evaluation.

Acknowledgments

The authors thank Igelösa Life Science AB for hosting this work and providing the experimental data. Funding came from the Hans-Gabriel and Alice Trolle-Wachtmeister, Hjelm Family, Mats Paulsson and LMK foundations, and from the Knut and Alice Wallenberg Foundation through the Wallenberg AI, Autonomous Systems and Software Program (WASP). ES and KS are members of the ELLIIT Strategic Research Area.

References

[1] Steen S, Paskevicius A, Liao Q, Steen E. Ex vivo evaluation of the whole heart function allowing selective investigation of the right and left heart. Scandinavian Cardiovascular Journal 2024;58(1):2418084.

[2] Global Observatory on Donation and Transplantation. International report on organ donation and transplantation activities 2024. Technical report, Global Observatory on Donation and Transplantation, 2025.

[3] Nilsson J, Jernryd V, Qin G, Paskevicius A, Metzsch C, Sjöberg T, Steen S. A nonrandomized open-label phase 2 trial of nonischemic heart preservation for human heart transplantation. Nature Communications 2020;11(1):2976.

[4] McGiffin DC, Kure CE, Macdonald PS, et al. Hypothermic oxygenated perfusion (HOPE) safely and effectively extends acceptable donor heart preservation times: results of the Australian and New Zealand trial. The Journal of Heart and Lung Transplantation 2024;43(3):485–495.

[5] Lelovas PP, Kostomitsopoulos NG, Xanthos TT. A comparative anatomic and physiologic overview of the porcine heart. Journal of the American Association for Laboratory Animal Science 2014;53(5):432–438.

[6] Arrigo M, Jessup M, Mullens W, Reza N, Shah AM, Sliwa K, Mebazaa A. Acute heart failure. Nature Reviews Disease Primers 2020;6(1):16.

[7] McDonagh TA, Metra M, Adamo M, et al. 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. European Heart Journal 2021;42(36):3599–3726.

[8] Naar J, Mlcek M, Kruger A, et al. Acute severe heart failure reduces heart rate variability: an experimental study in a porcine model. International Journal of Molecular Sciences 2023;24(1):493.

[9] Nair R, Lamaa N. Pulmonary capillary wedge pressure. In StatPearls. Treasure Island, FL: StatPearls Publishing, 2025; .

[10] Duncker DJ, Koller A, Merkus D, Canty JM. Regulation of coronary blood flow in health and ischemic heart disease. Progress in Cardiovascular Diseases 2015;57(5):409–422.

[11] Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2008.

[12] Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning. 2006; 233–240.

[13] Lee SM, Kim T, Shin M, Choi JO, Chung MJ, Kim D. Deep learning algorithm for detection of acute heart failure using standard ECG waveforms. European Heart Journal Digital Health 2026;7(2):ztaf132.

Address for correspondence:

Henry Pigot
Department of Translational Medicine
Lund University, Box 118
221 00 Lund, Sweden
henry.pigot@med.lu.se