Generalised Label-free Artefact Cleaning for Real-time Medical Pulsatile Time Series

Xuhang Chen, Ihsane Olakorede, Stefan Yu Bögli, Wenhao Xu, Erta Beqiri, Xuemeng Li, Chenyu Tang, Zeyu Gao, Shuo Gao, *Senior Member, IEEE*, Ari Ercole, Peter Smielewski

***Abstract*—Artefacts compromise clinical decision-making in the use of medical time series. Pulsatile waveforms offer opportunities for accurate artefact detection, yet most approaches rely on supervised manners and overlook patient-level distribution shifts. To address these issues, we introduce GenClean, a generalised label-free framework for real-time artefact cleaning, implemented within the ICM+ clinical research monitoring software. Leveraging an in-house dataset of 180,000 ten-second arterial blood pressure (ABP) samples for training, we first investigate patient-level generalisation, demonstrating robust performance under both intra- and inter-patient distribution shifts. As an initial exploration beyond the development cohort, we further validate its effectiveness for ABP through site-level generalisation on the MIMIC-III database. We also provided an extension of our method to photoplethysmography (PPG), highlighting its potential applicability to diverse medical pulsatile signals. The real-time integration and these generalisation studies collectively demonstrate the practical utility of our framework in continuous physiological monitoring and represent a promising step towards improving the reliability of high-resolution medical time series analysis.**

***Index Terms*— Artefact detection, Medical time series, Pulsatile signals, Real-time monitoring, Machine Learning.**

I. Introduction

Medical pulsatile time series, such as arterial blood pressure (ABP) and photoplethysmography (PPG) signals, provide vital insights into systemic physiology and clinical decision-making. These time series are continuously measured at high resolution either in hospital settings or ambulatory wearable healthcare applications, enabling the detailed analysis of dynamic waveforms and temporal trends to support the detection of clinical-relevant events[1]. However, these signals are often contaminated by artefacts from clinical interventions (e.g., blood sampling, line flushing) or patients' motion[2]. Such artefacts can distort signal characteristics, leading to misinterpretation of physiological events[3] or false alarms[4]. This risks alarm fatigue[3] and potentially inappropriate medical interventions at worst.

Over the past decades, various methods have been proposed to mitigate artefacts in medical pulsatile signals. Recent advances in machine learning artefact cleaning methods have shown superior performance compared to the traditional statistical and signal processing methods [5], [6], [7], particularly in addressing the challenges posed by non-stationary signals. Most contemporary methods are based on neural networks [8], [9], [10], [11], [12], [13], [14]. These approaches outperform traditional methods by leveraging their large number of parameters to perform end-to-end artefact cleaning on time series data.

Despite these advances, three major challenges (Fig. 1) that hinder the broader adoption of the current methods can be identified. First, generalisation issues at both the site and patient levels limit the usage and performance of current methods. Site-level discrepancies (i.e., inconsistent harmonisation) are well-recognised[15], including device variations, protocols, and sampling frequencies across collection sites, which restrict the model application to specific sites where the data is collected. However, patient-level distribution shifts remain underexplored, yet degrade the model performance[16]. Distribution shift refers to the differences in statistical properties (e.g., mean, variance, non-linear and high order statistics) of variables of interest. It is particularly pronounced in medical time series signals and

Manuscript received XXX; revised XXX; accepted XXX. Date of publication XXX; date of current version XXX. This research was supported in part by the National Natural Science Foundation of China (Grant No.: 62171014); in part by Beihang Ganwei project (Grant No.: KG16321901); in part by the Swiss National Science Foundation (Grant No.: 210839/225270) (*Corresponding author: Xuhang Chen, Shuo Gao).*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Committee of Cambridge University Hospitals, under ethical approval REC 23/YH/0085.

Peter Smielewski reports a relationship with Cambridge Enterprise Ltd, Cambridge, U.K. that includes consulting or advisory. Peter Smielewski has patent with royalties paid to Cambridge Enterprise Ltd, Cambridge, U.K. The remaining authors declare no competing interests.

Xuhang Chen, Ihsane Olakorede, Stefan Yu Bögli, Wenhao Xu, Erta Beqiri and Peter Smielewski are with the Brain Physics Laboratory, Division of Neurosurgery, Department of Clinical Neurosciences, University of Cambridge, Cambridge, CB2 0QQ, UK (e-mail: xc369@cam.ac.uk).

Zeyu Gao is with Department of Oncology and the CRUK Cambridge Centre, University of Cambridge, CB2 0QQ, UK.

Chenyu Tang is with the Department of Engineering, University of Cambridge, Cambridge, UK, CB3 0FA.

Shuo Gao is with Hangzhou International Innovation Institute, Beihang University, Hangzhou, China, and he and Xuemeng Li are with School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing, China. (e-mail: shuo_gao@buaa.edu.cn)

Ari Ercole is with the Division of Anaesthesia, Department of Medicine, University of Cambridge, Cambridge, CB2 0QQ, UK.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2026.3655095

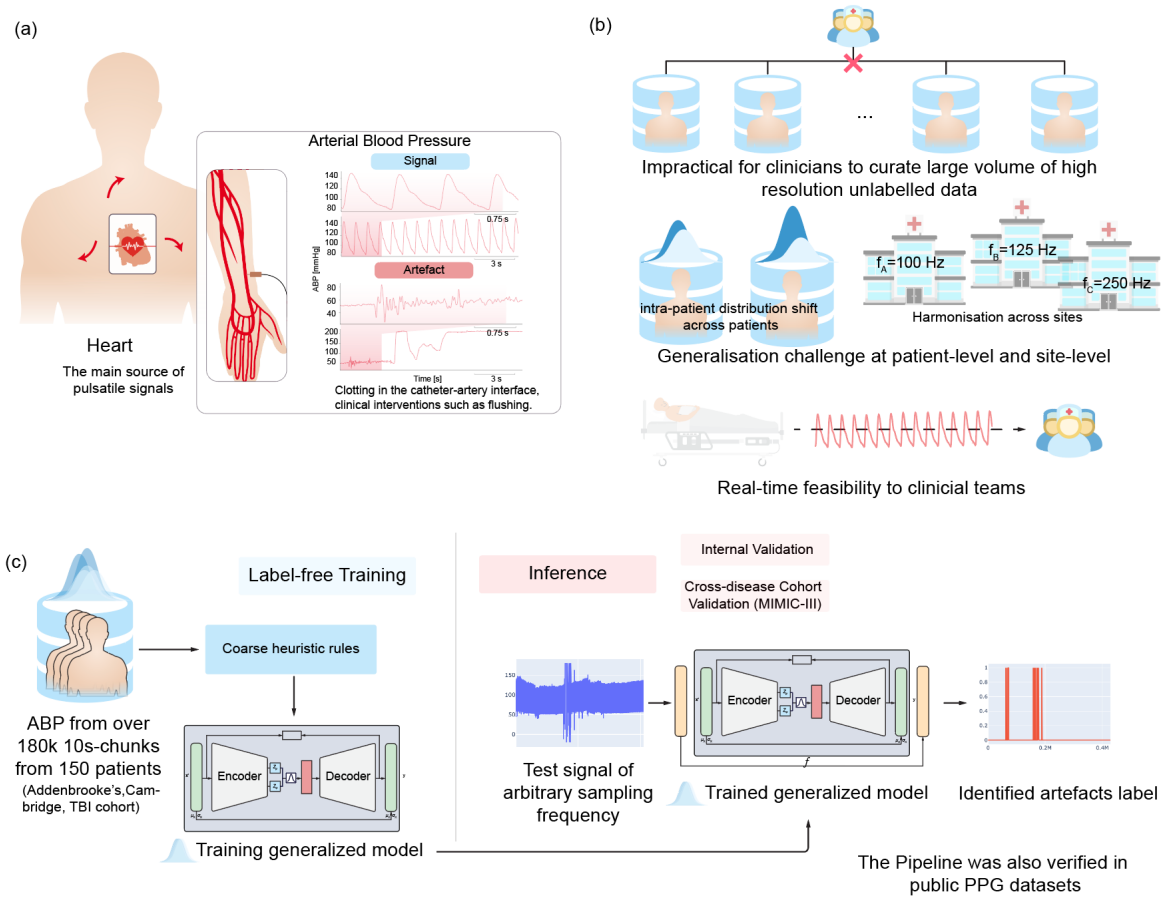

Fig. 1. Overview of the challenges and framework for label-free medical pulsatile signal artefact cleaning methods. a) Illustration of the origin of medical pulsatile signals and the arterial blood pressure (ABP) measurement. b) Key challenges in analysing medical pulsatile signals: (i) the scarcity of artefact-annotated data (ii) generalisation issues at the site- and patient-level (iii) lack of real-time feasibility as offline analysis is limited to retrospective analysis. c) The proposed framework for artefact detection. During training, data of 180,000 10-second samples from 150 patients were used to train a generalised model. Coarse heuristic rules are injected to filter out the extreme values to improve the input data quality of our label-free method. During inference, the trained model can process signals that include various generalisation challenges, such as frequency inconsistency and patient-level distribution shifts. Additionally, we verified our effectiveness in both the internal and cross-disease cohort MIMIC-III dataset. Also, this pipeline was verified in other medical pulsatile signals, such as PPG, to demonstrate the generalisability.

manifests itself at the intra-patient (e.g., ABP amplitude variations in physiological events) or inter-patient (e.g., individualised baseline ABP waveform) level [17]. Second, the scarcity of artefact-annotated data significantly limits the applicability of supervised learning methods. The absence of artefact-annotated data also highlights the urgent need for robust label-free methods in this domain. Third, the lack of real-time artefact cleaning power is a critical limitation of the current approaches. While retrospective analyses of medical time series aid research, real-time processing is essential for timely clinical decision support and false alarm reduction. Without real-time methods, the utility of artefact cleaning in dynamic, high-stakes environments remains restricted.

In this work, we address these gaps by developing GenClean, a generalised label-free framework for real-time artefact cleaning in medical pulsatile time series and its deployment within routine clinical monitoring (Fig. 1). Leveraging our label-free training method, we utilise 180,000 10-second (500 hours) arterial blood pressure (ABP) samples, extracted from traumatic brain-injured patients treated in Addenbrooke's Hospital, Cambridge, to train our model. We then conduct quantitative experiments to examine patient-level generalisation, demonstrating performance under both intra- and inter-patient distribution shifts. As an initial exploration beyond the development cohort, we validate our method through cross-cohort experiments (i.e., train on one cohort and test on another) on ABP and photoplethysmography (PPG) signals, including a more challenging cross-disease cohort setting on MIMIC-III patients, to showcase its effectiveness for artefact cleaning. Finally, we integrate our framework into ICM+, a clinical research monitoring software (Cambridge Enterprise Ltd, Cambridge, UK), to verify and demonstrate its real-time feasibility in clinical workflows. Our primary contribution is therefore a practical, label-free framework for real-time artefact cleaning that can be deployed within existing clinical monitoring software, while our cross-cohort and multi-signal studies serve as initial explorations of its broader generalisation. Overall, our label-free artefact-cleaning framework provides a practical step towards generalisable, real-

time medical signal analysis across international clinical sites and varying harmonisation levels of medical signals.

II. Methods

A. Data Preparation

We used ABP data sourced from the Brain Physics Research Database as an internal training and test set and conducted a cross-disease cohort validation on the MIMIC-III datasets. The Brain Physics database ABP data (n=160) was measured by arterial line (Baxter Healthcare, Deerfield, Illinois) inserted into the radial or femoral artery and recorded at a frequency of 120 Hz using ICM+ software in patients with traumatic brain injury admitted into the Neurocritical Care Unit, Addenbrooke's Hospital, Cambridge, UK (2011-2019), requiring intracranial pressure monitoring. ABP was monitored through radial or femoral arterial lines connected to pressure transducers (Edward Lifesciences, Irvine, CA). For the training and validation set, each patient contributed 1,000 10-second samples and 200 samples to the validation set. This resulted in a total of 180,000 10-second samples. For the test set, each patient provided 100 balanced 10-second samples, amounting to a total of 1,000 samples (Table.1), ensuring sufficient representation of individual physiological variability in the dataset. The cross-disease cohort, MIMIC-III, was collected from critical care units of the Beth Israel Deaconess Medical Center in Boston, USA (2001-2012). ABP was collected by an HP CMS (Merlin) monitoring device at the sampling frequency of 125 Hz [18]. We randomly selected five patients with three-hour sections as a validation experiment due to the scarcity of patients with ABP data.

For additional medical pulsatile time series validation, we utilised a publicly available PPG dataset[10], commonly utilised in healthcare applications. The DaLiA dataset (n=15; 1,837 10-second samples) included recordings from 15 participants performing daily tasks such as walking, cycling and driving. The WESAD dataset (n=15; 8,664 10-second samples) included recordings from 15 participants in different emotional states such as neutral, stressed, and amused. The PPG signals were segmented into 10-second samples to align with the input shape of our model. Following the previous work from Chen et al.[10], we employed the DaLiA dataset for label-free training and the WESAD dataset for testing.

B. Heuristic Rules

Medical pulsatile time-series signals, such as ABP and PPG, provide rich physiological information and exhibit high interpretability compared to other biomedical signals (e.g., EEG). We applied heuristic rules based on domain knowledge to filter out extreme artefactual or non-physiological signals (e.g., negative values, flat lines) that are unsuitable for clinical decision-making. The ABP signal is clipped to a physiological range of 0 – 300 mmHg, with the pulse amplitude above 15 mmHg [8].

A frequency filter is applied to PPG signals to suppress the frequencies outside of the 0.5 Hz – 3 Hz range [19]. Segments where more than 30% of the total signal power falls outside this range were classified as artefacts.

TABLE I
Details of In-Hospital ABP Datasets

Split	Patients	Per-patient sampling	Label
Training	150	1000	Unlabelled
Validation	150	200	Unlabelled
Test	10 (held-out)	100	Balanced labelled (positive:negative=1:1)

$$\rho = \frac{\int_{0.5}^{3} P(f)df}{\int_{0}^{f_s} P(f)df} \; and \; \text{flag artefacts if } \; \rho < 0.7 \qquad (1)$$

These criteria can enhance the quality of data for training label-free variational autoencoders to learn physiology while minimising the impact of extreme artefacts.

C. Artefact Cleaning Framework

To address the challenges posed by diverse datasets in artefact cleaning, we propose a unified framework combining the frequency adaptation, variational autoencoder (VAE)-based backbone, and adaptive norm module. This design ensures compatibility across varying sampling rates for site-level generalisation, robust representation of clean waveforms, and effective handling of patient-level distribution shifts.

We apply the variational autoencoder to learn the probabilistic representation of a clean waveform which constrains the learned normal waveform, leveraging a one-dimensional convolutional structure in the encoder and decoder. This architecture focuses on local signal features, aligning with clinical insights into pulsatile signal characteristics. This design considers the trade-off between the feasibility of real-time implementation, physiological domain knowledge, and algorithm robustness. The loss for training is selected as the standard evidence lower bound objective loss.

$$\mathcal{L} = \mathbb{E}_{q_\varphi(z|x)}[log\ p_\theta(x|z)] - D_{KL}(q_\varphi(z|x)\|p(z)) \qquad (2)$$

$$\mathbb{E}_{q_\varphi(z|x)}[log\ p_\theta(x|z)] = \frac{1}{N}\|x - \hat{x}_\theta(z)\|_2^2 \qquad (3)$$

$$D_{KL}\big(q_\varphi(z|x)\|p(z)\big) = \frac{1}{2}\sum_{i=1}^{d}(\mu_i^2 + \sigma_i^2 - log\sigma_i^2 - 1) \qquad (4)$$

where $q_\varphi(z|x)$ is the encoder, $p_\theta(x|z)$ is the decoder, x_i is the input, $\hat{x_i}$ is the decoder output, μ_i and σ_i are the mean and standard deviation of the latent variable z.

The Frequency Adapter is designed to manage high-resolution data from multiple sites (e.g., 125 Hz for the MIMIC-III vs. 120 Hz for the training set). Input signals are resampled to the target frequency the model requires, and outputs are resampled back to the original frequency for downstream analysis. This reversible process preserves the original temporal structure while ensuring compatibility with the model's training pipeline.

In our study, the sampling frequency is provided in the HDF5 data stream from ICM+ software, and a tool for extracting it is available in our code. Building upon this adapter structure and estimated frequency design, our module can handle a direct one-dimensional signal input and return the exact same length

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2026.3655095

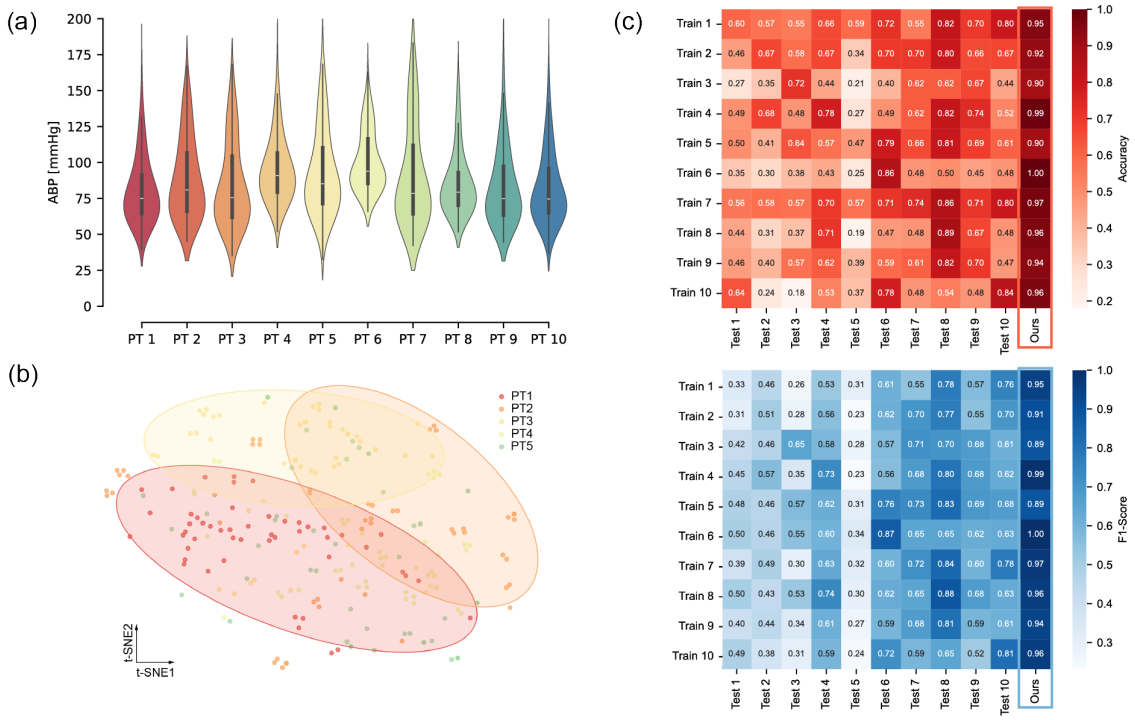

Fig. 2. Visualisations of data distribution, generalisation, and model performance. (a) Violin plots of non-artefactual ABP signal values for 10 patients, illustrating the distribution of high resolution arterial blood pressures. Clear inter-patient variability is observed, reflecting physiological and pathological differences among patients, while the varying locations and shapes of the density peaks across patients suggest inter-patient differences (unique physiological characteristics in each patient). (b) t-SNE visualisation of the first five patients, selected to enhance clarity and reduce visual clutter. The distinct clusters show three patients with clear distribution shifts, while the other two exhibit overlap with these clusters, indicating potential patient-specific morphological variability. (c) Generalisation matrices comparing accuracy (left) and F1-score (right) for cross-patient training and testing (train on patient i, test on patient j) in the previous label-free method. The last column is the performance of our generalised label-free method. The result highlights the model's generalisation capability, with better performance seen in patients with consistent data distributions.

decoded results and corresponding label, enabling convenient usage.

Furthermore, to address the patient-level distribution shift issue, we applied reversible instance normalisation [20] within a large group of patients. Considering a one-dimensional signal sample $x_k \in R^C$, ($x_k^{(i)}$ is the real value at the time stamp i) the mean and variance are computed at the instance-level,

$$\mathbb{E}[x_k] = \frac{1}{C}\sum_{i=1}^{C} x_k^{(i)}, Var[x_k] = \frac{1}{C}\sum_{i=1}^{C}(x_k^{(i)} - \mathbb{E}[x_k])^2 \quad (5)$$

Denote the original self-supervised network as $M_\omega: R^C \rightarrow R^C$. The input becomes

$$\hat{x}_k = \frac{\gamma}{\sqrt{Var[x_k] + \epsilon}}(x_k - \mathbb{E}[x_k]) + \beta \quad (6)$$

Where γ and β are trainable parameters, ϵ is a small value to avoid division by zero. To restore the original scale after processing, the output of the model is transformed back as

$$\hat{y}_k = \sqrt{Var[x_k] + \epsilon}\left(\frac{M_\omega(\hat{x}_k) - \beta}{\gamma}\right) + \mathbb{E}[x_k] \quad (7)$$

The symmetric structure of this normalisation layer within the network ensures the scale stability.

Finally, artefacts are identified based on reconstruction error between the input signal and the decoder output. We empirically used the mean squared error (MSE) as a metric, and the artefact threshold was empirically set as the 90th percentile of our validation set. Segments exceeding this threshold are classified by the algorithm as artefact contaminated. This yielded stable performance in our cohorts. To evaluate the performance of artefact identification, based on the artefactual sample and our experts-labelled data, we employed a comprehensive set of metrics, including accuracy, sensitivity, specificity, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC).

D. Real-Time Implementation

The real-time implementation of our algorithm was integrated into the ICM+ Microsoft Windows-based software platform, running on an Intel Xeon CPU E5-2620 v3@ 2.40 GHz. The algorithm was first packaged as a plugin within the software, enabling activation for real-time performance evaluation. Using a simulated data stream, we measured the performance of the computer both with and without the plugin. For measuring the processing time, the plugin recorded the computation time required for a 10-second data sample. Simultaneously, we monitored resource utilisation, including CPU and memory usage, to evaluate the algorithm's computational overhead. This ensured that the algorithm met real-time requirements without compromising the responsiveness of the host computer.

III. Experimental Results

A. Datasets and Settings

We used both ABP and PPG datasets to evaluate our model. The primary ABP dataset (120 Hz) comprised data from patients (n=160; > 18 years) with traumatic brain injury (TBI), sourced from Brain Physics Database, Addenbrooke's Hospital, Cambridge, under ethical approval REC 23/YH/0085. Of these, data from 150 patients were allocated to the training and validation set. And data from the remaining 10 patients were reserved as an internal held-out set. We segmented both datasets into 10-second samples to manage the data efficiently, a duration that minimises computational and data-loading burdens while maintaining alignment with clinically relevant metrics, such as the pressure reactivity index (PRx)[21], a key metric used for continuous cerebral autoregulation monitoring in TBI.

To evaluate the generalisability of our model to a cross-disease cohort, we incorporated ABP signals from five patients retrieved from the MIMIC-III Waveform Database[18] as a cross-disease cohort validation set (see Supplementary Materials S4 for details). This cohort encompassed a broader range of patients from intensive care units (ICUs). Moreover, photoplethysmography (PPG) datasets, including DaLiA (n=15; 1,837 10-second samples) and WESAD (n=15; 8,664 10-second samples)[22], [23], were used to demonstrate further the validity of our model on other types of medical pulsatile signals.

Moreover, two experienced physiological-signal analysts (each with > 2 years of experience in ABP interpretation) independently annotated the ABP datasets. An artefact was defined as a non-physiological waveform, including segments containing saturation, flat-line, or motion-induced distortion. A total of 179 artefacts were identified in the five patients from the MIMIC-III dataset. The inter-rater agreement was high (Cohen's $\kappa = 0.96$), and disagreements were resolved by consensus. For the PPG dataset, we used the official labels. A full description of our methods and the training details is provided in the Supplementary Materials S1 and S2.

B. Distribution Shift Exploration

We investigated inter and intra patient distribution shift issues in the non-artefactual part of the internal held-out set. Inter patient distribution shifts can be observed via differences in statistical summaries (e.g., median, interquartile range, Fig.2a). All ten patients showed different median and quartile values. For example, the median of PT 6 was close to 90 mmHg and that of PT 3 was nearly 75 mmHg. This was confirmed using a two-sample Kolmogorov-Smirnov test with Bonferroni correction for each pair of patients, which showed a significant difference between all pairs of patients ($p<0.05$). Intra patient distribution shifts can be reflected by peaks and tails in the violin shape of each patient, particularly the multi-peak patterns. Each peak represents a distinct range of frequently occurring ABP values. These peaks may correspond to different physiological states experienced by the patient over time. For instance, the four-peak pattern in PT 10 suggests that the patient transitioned through multiple physiological states or events during monitoring, showing the intra patient distributional variability. Fig.2b employs t-distributed stochastic neighbour embedding (t-SNE) to project 10-second signal samples into two dimensions, visualising qualitative distribution shifts at the sample level for selected patients. PT 1, 2 and 3 formed distinct groupings, suggesting distributional differences among these patients, while data from PT 4 and PT 5 were scattered across multiple clusters, suggesting overlapping distribution characteristics. These clustering results reveal the intricate distribution patterns within the ABP signal. Overall, these observations demonstrate distribution deviations in patient ABP data.

Building on these observations, we evaluated the performance of our previously proposed label-free method8 in the face of the distribution shift challenge on the test set. In Fig.2c, the 10×11 generalisation matrix visualises the metrics (accuracy and F1 score) of training on patient i and testing on patient j where each row indicates training on the same patient, while the last column shows the results from our generalised label-free method trained on our training set. The diagonal line of both matrices manifests training and testing on the same patient, while the other area indicates cross-patient validation. The F1 score matrix was similar to the accuracy matrix, with slightly lower F1-score values (deviation of up to 0.3 - Train 1 Test 3). Our generalised model exhibited excellent performance with at least 90% accuracy and 0.89 F1-score even in the distribution shift settings, outperforming the previous method across all the patients. In the previous method, seven out of ten patients achieved satisfactory performance (>70%) in the generalisation test. Additionally, the diagonal line (self-test) tends to exhibit higher accuracy, however, this trend was not consistently observed in our data on the accuracy matrix (accuracy of 0.80 for Train 1 Test 10 vs. 0.60 for Train 1 Test 1). One of the patients (Test 5) consistently exhibited lower performance in both metrics, with no value above 0.6. This distribution shift observation highlighted our generalised model's superior performance (Fig.2c) in dealing with this complexity where previous methods often struggled.

C. Generalisation on Cross-cohort

We further evaluated the generalisation performance of our models in data sets with different sampling frequencies of signals and another cohort of patients. Fig. 3b displays our model performance on accuracy and processing time in a range of common sampling frequencies (50, 75, 100, 120, 125, 150, 175, 200 and 240 Hz). The result indicates no notable performance degradation across varying sampling rates. Regarding the time for artefact cleaning, it takes ~19 ms for 120 Hz, which is the frequency that our model trained on, with the longest processing time reaching ~26 ms (150 Hz). This result confirms the robustness of our method in handling diverse frequencies.

To test the model's generalisation further, we conducted a cross-disease cohort validation experiment on randomly selected ABP signals from the MIMIC-III waveform database. Here, the model trained on our internal set (TBI cohort, 120 Hz) was directly applied to MIMIC-III data (general ICU cohort, 125 Hz) to verify the artefact cleaning effects on out-of-distribution patients and inconsistent data collection standards. An accuracy of 95.6% was obtained on five patients from

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2026.3655095

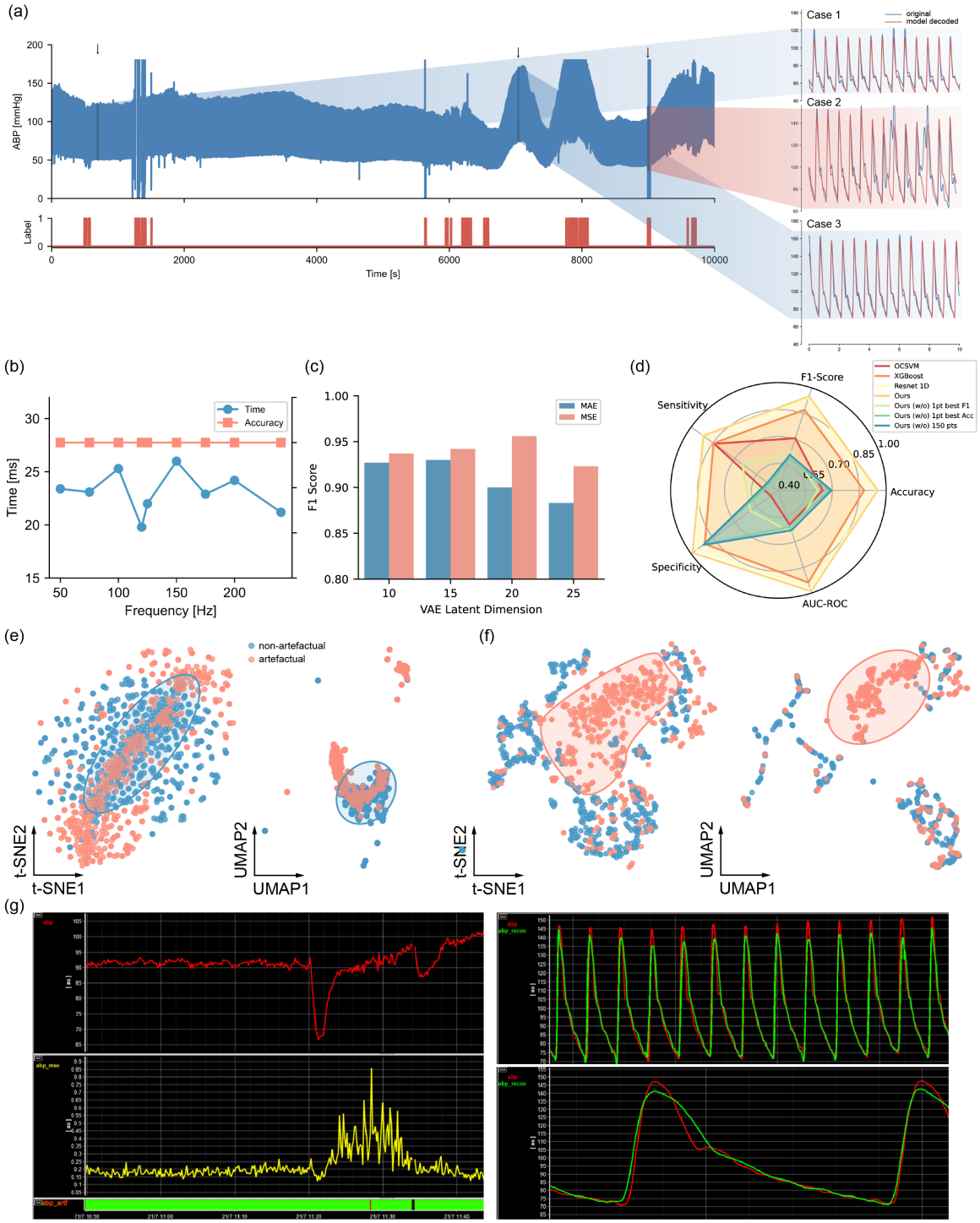

Fig. 3 Model performance visualisations and real-time feasibility. (a) Example of processing of an ABP recording from the MIMIC-III database. The left panel displays a section of the ABP waveform including dynamic changes and artefacts, with the model generated artefacts markup (below). On the right, Case 1 and Case 3 showcase different physiological stages with correctly reconstructed ABP waveforms. Case 2 represents an artefact instance, where the model successfully identifies the section as anomalous, with a clear discrepancy between the original and reconstructed signals. (b) the accuracy and processing time of our method at various input sampling rates (120 Hz is the frequency our model was trained on). (c) Impact of the latent space dimensionality and error evaluation metrics (MSE and MAE) on model performance. (d) Comprehensive comparison of model performance using a radar plot, including baseline models, alongside our model variants with ablation settings, highlights our model's superior generalisation and accuracy across metrics. (e) and (f) Visualisation of data and latent space distributions using t-SNE and UMAP. The left panel illustrates the original data with significant overlap in the blue-highlighted region, indicating limited separability. The right panel displays the transformed latent space, where clear clustering emerges, demonstrating the effectiveness of the model in separating physiological categories post-processing. (g) Left: ICM+ screenshot showing example of monitored signals, with the error (abp_mse) between reconstruction and the origin in the middle, artefact annotation panel (abp_artf) displayed at the bottom. Right: the real-time reconstructed pulse wave, reflecting our explanatory panel.

TABLE II
SPECIFIC PERFORMANCE OF THE PROPOSED ARTEFACT CLEANING METHOD IN CONTINUOUS MONITORING ABP AND PPG DATASET

Metrics	MIMIC-III (ABP)	WESAD (PPG)
Accuracy (%, [95%CI])	95.6 [94.6, 96.4]	85.8 [84.2, 87.3]
Specificity (%, [95%CI])	97.2 [95.9, 98.1]	86.7 [84.0, 89.0]
Sensitivity (%, [95%CI])	94.3 [92.8, 95.5]	85.3 [83.3, 87.1]
HTE before cleaning (N)	2134	N/A
HTE after cleaning (N)	1661	N/A
Reduced HTE proportion	22%	N/A

MIMIC-III, with further quantitative results available in Table 2. Fig.3a displays a data overview of one patient and the detected label from our model, with three expanded illustrative pulse waveform sections, called cases. Cases 1 and 3 show very reasonable model performance on pulses with different morphology and statistical properties (e.g., mean and standard deviation), correctly identified as artefact-free. Case 2 demonstrates a distorted trace, which our model identifies correctly as an artefact, with the decoded waveform being very different from the original. These three cases also illustrate the intra-patient level distribution shift.

To shed light on the potential clinical relevance, we analysed the impact of artefact cleaning on the reduction of detection of hypertension pulse events. Here, we adopted a surrogate pulse-level hypertension detection task, following the standard that systolic blood pressure exceeds 140 mmHg or diastolic blood pressure exceeds 90 mmHg [24]. We obtained positions of systolic peaks P_{sys} and diastolic peaks P_{dia} with the findpeaks function (Scipy package), with a distance of 36 (0.3s × 120 Hz) to avoid over-sensitive detection. Then, Hypertension pulse events (HTEs) were defined as

$$P_{sys} = \{ p_k \mid p_k = findpeaks(x) \} \quad (8)$$

$$P_{dia} = \{ p_k \mid p_k = findpeaks(-x) \} \quad (9)$$

$$HTE = (\max \mathcal{P}_{sys} > 140) \cup (\max \mathcal{P}_{dia} > 90) \quad (10)$$

We applied our model to five patients of the MIMIC-III dataset and counted the reduction in the detected number of hypertensive events after cleaning. Independent reference review verified that 22% of detections suppressed by our method were spurious (false positives).

Furthermore, to explore our model's generalisation in other medical pulsatile time series, we used PPG signals as a candidate for testing[10], as artefact cleaning on PPG signals is a challenging task, particularly when the data is collected on wearables and is more sensitive to motion artefacts. We trained our model on the DaLiA dataset and tested it on the WESAD dataset to verify the model's performance on a completely different signal. Our results (Table 2) revealed an 85.8% accuracy and 0.86 F1-score in the PPG dataset. The result is comparable to the results (0.864) from state-of-the-art supervised learning method [10].

D. Model Architecture Analysis

We conducted ablation studies (i.e., modifying the components of our method) and compared the results to baseline methods (Fig. 3d). The performance of five evaluation metrics was exhibited throughout our ablated models and baseline methods. These include one-class support vector machine (OCSVM)[25], a classic label-free method, ResNet1D[26] and XGBoost[27], two famous supervised learning methods proven to achieve decent outcomes for artefacts detection, and our three ablated models (named w/o) without specialised design for generalisation (i.e., distribution shifts). Our model (labelled 'Ours' in the figure) outperformed all the models, even including the supervised learning methods in all five metrics. Training over a large patient dataset (150 patients) without addressing generalisation issues significantly degrades performance. This result demonstrates the validity of our model design and the superior performance as a label-free method.

Moreover, we further compared the variational autoencoder with an autoencoder with the same structure. We found that the variational autoencoder design reduced the reconstruction loss by 7.4% and increased artefact detection accuracy by 2.2%. More implementation details and classic methods, such as template matching and heuristic filters, can refer to Supplementary Materials S2 and S3.

E. Latent Space Representation

We analysed the impact of our latent dimension from the perspective of model design. The latent dimension refers to the size of the compressed representation (i.e., the latent vector) of waveforms, which represents the space in which the input signal is encoded and later reconstructed by the decoder. This latent vector serves as an embedding of the compressed physiological information of the medical pulsatile time series included in our model. Fig. 3c illustrates the impact of varying the latent space dimension in the variational autoencoder on the F1 score, as this metric is stricter than Accuracy as shown in Fig.2c. Our results indicated that a dimension size of 20 provided an optimal balance, with the highest score observed when mean squared error (MSE) was used as the reconstruction error evaluation metric. Deviations from this dimension size resulted in depleted performance.

To evaluate the effectiveness of the representations learned from our model, we conducted a detailed qualitative analysis of the latent vectors. Two widely used dimension reduction methods, t-SNE and Uniform Manifold Approximation and Projection (UMAP), were applied to visualise and disentangle the learned latent representations. Before applying our model, Fig. 3e shows the results from our internal held-out set, where points of both categories are intermixed within the labelled blue region. After applying our model, Fig. 3f displays the learned latent vector maps. The latent space demonstrated improved separation between the two categories. This suggests that the latent space learned by our model effectively captures the underlying data structure, enhancing its ability to distinguish between artefactual and clean signals.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2026.3655095

TABLE III
REAL-TIME FEASIBILITY METRICS OF OUR ARTEFACT CLEANING METHOD MEASURED IN ICM+

Metrics	Value	Description
Processing Time	19 ms	Average time to process a 10-second segment, ensuring real-time feasibility.
Model Computation	3.2 mFLOPS	Computational cost per segment
CPU Utilization Increment	+6%	Average CPU load during processing, compared to baseline system utilization.
Memory Usage Increment	+139 MB	Additional memory used for model inference during the experiment.

TABLE IV
COMPARISON OF KEY CHARACTERISTICS OF EACH STATE-OF-THE-ART METHOD

Literature	Label-free	Multi-Centre Validation	Real-time Stream	Multiple Signals
[8]	✓	×	×	×
[10]	×	✓	✓	✓
[12]	✓	✓	×	×
[13]	×	✓	×	×
[28]	✓	✓	×	✓
[36]	×	✓	✓	×
Ours	✓	✓	✓	✓

F. Real-time Feasibility

Most of the previous methods were demonstrated in offline analysis in retrospective datasets [8], [28], [29], [30]. Real-time artefact cleaning capability is however essential for effectively suppressing false alarms and providing support for clinical decision-making. To evaluate the real-time feasibility of the method, we applied our GenClean method to a simulated normal monitoring data stream using ICM+, a clinical research monitoring software.

As shown in Fig. 3g, the left panel displays a long-period monitoring graph, while the right panel provides fine-grained patient data and the decoded waveforms produced by our model. Table 3 summarises the quantitative computation overheads during the running period of our method as a plugin. The increments of running our model on the central processing unit (CPU, +6%) and memory (+139 MB) were within a reasonable range and the processing time negligible, making this approach feasible even when using low specifications hardware for bedside data collection/processing. The computation cost measured in floating-point operations (FLOPs), was significantly lower (3.2 mFLOPS) than deep neural networks like ResNet1D (3 GFLOPS). These findings confirm the practical applicability of our approach in real-time clinical environments.

IV. DISCUSSION

Building upon insights from DeepClean [8], we developed GenClean, a state-of-the-art, generalised, label-free method for artefact cleaning and demonstrated its real-time feasibility on clinical monitoring software running on CPU. Our framework, trained on 180,000 10-second samples of data collected at Addenbrooke's Hospital, integrates a generalisation design and a label-free training strategy. Despite challenges such as site-level harmonisation issues and cross-disease cohort variability, our model achieved over 90% artefact recognition accuracy on the MIMIC-III dataset. When applied to a PPG signal, a widely used non-invasive medical pulsatile signal, our method delivered robust results with an accuracy exceeding 85% on publicly available datasets.

Over the last two decades, a substantial body of work on artefact detection and cleaning has been developed across a range of physiological signals. Early work relied on adaptive filtering and signal quality indices derived from morphology and amplitude heuristics suppress false alarms [6], [31], [32]. Traditional machine learning classifiers were then trained on hand-crafted features for physiological signal quality assessment[33], [34]. These methods are limited by the ability to cope with morphologies and large-scale deployment. More recently, supervised machine learning and deep learning methods have been proposed for artefact detection [35], [36], [37], [38], [39]. However, these models remain largely supervised and typically trained on relatively small, single-centre datasets and rarely address generalisation across patients and sites, as summarised in Table 4. In contrast, GenClean is trained label-free on 180,000 ABP segments from a multi-patient cohort and is explicitly designed to address both patient-level and site-level generalisation. We first investigate patient-level generalisation on ABP and demonstrate improvements under intra- and inter-patient distribution shifts. We then validate the trained model on five patients from the MIMIC-III dataset, with different data sampling frequencies, as a feasibility exploration of cross-site and cross-disease cohort settings. By analysing current bottlenecks of both patient- and site-level generalisation, our approach effectively leverages the large training cohort and achieves competitive performance compared with supervised methods, while mitigating the risks associated with distribution shifts and improving the quality of data available for downstream medical models.

A key innovation of our work lies in its real-time feasibility, a critical requirement for bedside monitoring. While many existing methods have been developed and validated using retrospective datasets, real-time detection is essential for continuously tracking patient status and enabling timely intervention. Our framework addresses this issue by integrating our model as a plugin into a clinical research monitoring software to validate the feasibility of this approach. Due to stringent privacy restrictions in clinical settings and limiting possibilities of streaming data in real-time to high-performance computing infrastructures, real-time feasibility must be achieved on CPU-based bedside hardware, which constrains many existing methods[40], [41]. The reasonable computation overheads of our approach demonstrate its feasibility for application in a modest hardware requirement setting. Furthermore, we also demonstrated that artefact cleaning significantly reduced false hypertension pulse events, which might otherwise trigger frequent monitoring device alarms and even lead to alarm fatigue among clinical staff[42]. Our framework maximises the value of artefact cleaning, potentially improving the accuracy of real-time clinical metrics such as the pressure reactivity index[21]. These improvements are critical

for advancing precision medicine by providing more accurate reflections of patient physiology and supporting individualised care.

In addition, our framework benefits significantly from the variational autoencoder (VAE) backbone. Previously widely used in image generation[43], the VAE's effectiveness in this domain has been validated in DeepClean[8]. Within our framework, the VAE model provides three distinct advantages. First, our design allows artefact detection to be more effectively achieved through the decoder's outputs. By generating non-artefactual segments based on input signal features, deviations (e.g., Fig. 3a, cases 1–3) can be directly observed to assess signal integrity and potentially recover contaminated sections. This approach is more interpretable than black-box feature engineering algorithms. Second, compared to similar methods[28], our robust VAE backbone can obtain more standardised and transparent metrics for artefact identification among patients, which avoids the use of subsequent complex machine learning methods on latent vectors to identify artefacts. Also, unlike autoencoder methods, the probabilistic modelling of VAEs provides continuous probability modelling in the latent space, ensuring a smooth and continuous representation of the data. This continuity allows for a more structured latent space, enabling meaningful dimensionality reduction, as demonstrated by t-SNE or UMAP latent space visualisations. Finally, the latent vector may add value to the understanding of patients' states, as the essence of understanding artefacts lies in identifying and distinguishing features of normal signals. Thus, continuous monitoring facilitates the analysis of temporal trajectories of patient status which could potentially highlight patient-specific conditions and distributions. To achieve this, potential spatial transitions corresponding to different clinical states or labels may need to be identified.

In selecting medical pulsatile signals, we focused on two representative cases, invasive ABP, the clinical gold standard for continuous blood pressure monitoring, and PPG, one of the most used non-invasive pulsatile signals in consumer healthcare electronics. Testing our model on these two signals demonstrated strong performance, indicating its potential applicability to a broader spectrum of medical pulsatile signals. Medical pulsatile time series are generally driven by cardiac activity, characterised by consistent waveforms, and governed by well-defined physiological principles. This inherent stability ensures the presence of reliable non-artefactual sections, which our method leverages to learn physiological patterns while addressing generalisation challenges across samples. By doing so, our framework enables large-scale label-free training and holds promise for application to other medical signals. Furthermore, our work does not diminish the value of alternative approaches, such as supervised or statistical methods. On the contrary, we advocate for a multi-stage methodology: our label-free framework can be employed for large-scale artefact cleaning during the initial stages of data processing, while identified labelled cases from this process can further inform supervised training. We believe that semi-supervised and supervised approaches remain highly complementary and encourage their continued exploration and development alongside label-free methods to advance this domain, for example, potentially enable foundation models[44] and agent-based systems[45], [46] that rely on high-quality streaming data.

Despite its strengths, our study has certain limitations. First, although our framework demonstrates promising generalisation to unseen patients, the diversity and size of the current dataset are still limited. The training cohort primarily consisted of TBI patients, and the external testing set (MIMIC) involved only a small number of subjects. As such, our results should be interpreted within this context rather than as a full solution to the patient-level distribution shift problem. Expanding the dataset to a broader population and modalities can facilitate its applicability. We acknowledge that collecting these larger datasets and addressing the site-level generalisation problems within the data (e.g., differences in collection protocols and device configurations) still requires a concerted effort from the research community. Second, as a probabilistic model, the VAE may struggle with out-of-distribution cases not covered in the training set, such as cardiac arrhythmias, as the model learns to reconstruct common patterns from the training data. This limitation is particularly relevant for signals that exhibit large inter-individual or site-specific differences, for example, variations in waveform morphology. While our framework captures a wide range of physiological diversity within the available data, it may not fully represent such extreme or rare conditions. When faced with unseen physiological variations, the model may incorrectly assign them to the closest learned distribution in common cases. Addressing those special cases may require targeted solutions. Third, implementation for real-time processing of signals from patient monitors requires more attention. In practical implementation, artefactual segments should be masked rather than deleted to ensure data continuity and preserve potentially useful information for other downstream analyses. Future efforts will focus on deploying and testing the framework in bedside environments to assess its practical impact on patient care.

V. Conclusion

In this work, we presented GenClean, a generalised, label-free and real-time artefact cleaning framework for high-resolution medical pulsatile time series. Our primary contribution is to provide a practical real-time implementation that can be integrated into clinical research monitoring software to support continuous bedside use. Through extensive experiments on ABP and PPG signals from diverse cohorts, we demonstrated that our method achieves strong artefact detection performance under both intra- and inter-patient distribution shifts; these studies serve as initial explorations of its generalisation across sites, diseases and signal types. We also verified the real-time feasibility of the whole framework within ICM+, confirming that it can be deployed on standard clinical hardware. We believe our method can help ensure high-quality data collection and real-time data curation, which are essential steps towards robust data-driven precision medicine and individualised patient care.

REFERENCES

[1] S. Brouwers, I. Sudano, Y. Kokubo, and E. M. Sulaica, 'Arterial hypertension', *The Lancet*, vol. 398, no. 10296, pp. 249–261, July 2021, doi: 10.1016/S0140-6736(21)00221-X.

[2] Y. Nguyen and V. Bora, 'Arterial Pressure Monitoring', in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025. Accessed: Feb. 16, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK556127/

[3] H. Kim, S.-B. Lee, Y. Son, M. Czosnyka, and D.-J. Kim, 'Hemodynamic Instability and Cardiovascular Events After Traumatic Brain Injury Predict Outcome After Artifact Removal With Deep Belief Network Analysis', *J. Neurosurg. Anesthesiol.*, vol. 30, no. 4, p. 347, Oct. 2018, doi: 10.1097/ANA.0000000000000462.

[4] W.-T. M. Au-Yeung, A. K. Sahani, E. M. Isselbacher, and A. A. Armoundas, 'Reduction of false alarms in the intensive care unit using an optimized machine learning based approach', *Npj Digit. Med.*, vol. 2, no. 1, pp. 1–5, Sept. 2019, doi: 10.1038/s41746-019-0160-7.

[5] Md. K. Islam, A. Rastegarnia, and S. Sanei, 'Signal Artifacts and Techniques for Artifacts and Noise Removal', in *Signal Processing Techniques for Computational Health Informatics*, vol. 192, M. A. R. Ahad and M. U. Ahmed, Eds, in Intelligent Systems Reference Library, vol. 192. , Cham: Springer International Publishing, 2021, pp. 23–79. doi: 10.1007/978-3-030-54932-9_2.

[6] Q. Li, R. G. Mark, and G. D. Clifford, 'Artificial arterial blood pressure artifact models and an evaluation of a robust blood pressure and heart rate estimator', *Biomed. Eng. OnLine*, vol. 8, no. 1, p. 13, Dec. 2009, doi: 10.1186/1475-925X-8-13.

[7] S.-C. Wang *et al.*, 'Arterial blood pressure waveform in liver transplant surgery possesses variability of morphology reflecting recipients' acuity and predicting short term outcomes', *J. Clin. Monit. Comput.*, vol. 37, no. 6, pp. 1521–1531, Dec. 2023, doi: 10.1007/s10877-023-01047-9.

[8] T. Edinburgh, P. Smielewski, M. Czosnyka, M. Cabeleira, S. J. Eglen, and A. Ercole, 'DeepClean: Self-Supervised Artefact Rejection for Intensive Care Waveform Data Using Deep Generative Learning', in *Intracranial Pressure and Neuromonitoring XVII*, B. Depreitere, G. Meyfroidt, and F. Güiza, Eds, in Acta Neurochirurgica Supplement. , Cham: Springer International Publishing, 2021, pp. 235–241. doi: 10.1007/978-3-030-59436-7_45.

[9] Y. Zheng, C. Wu, P. Cai, Z. Zhong, H. Huang, and Y. Jiang, 'Tiny-PPG: A lightweight deep neural network for real-time detection of motion artifacts in photoplethysmogram signals on edge devices', *Internet Things*, vol. 25, p. 101007, Apr. 2024, doi: 10.1016/j.iot.2023.101007.

[10] S. F. Chen, Z. Guo, C. Ding, X. Hu, and C. Rudin, 'Sparse learned kernels for interpretable and efficient medical time series processing', *Nat. Mach. Intell.*, vol. 6, no. 10, pp. 1132–1144, Sept. 2024, doi: 10.1038/s42256-024-00898-4.

[11] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, 'TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis', Apr. 11, 2023, *arXiv*: arXiv:2210.02186. Accessed: Aug. 20, 2024. [Online]. Available: http://arxiv.org/abs/2210.02186

[12] L. Xing and A. J. Casson, 'Deep Autoencoder for Real-Time Single-Channel EEG Cleaning and Its Smartphone Implementation Using TensorFlow Lite With Hardware/Software Acceleration', *IEEE Trans. Biomed. Eng.*, vol. 71, no. 11, pp. 3111–3122, Nov. 2024, doi: 10.1109/TBME.2024.3408331.

[13] M.-B. Hossain, H. F. Posada-Quintero, and K. H. Chon, 'A Deep Convolutional Autoencoder for Automatic Motion Artifact Removal in Electrodermal Activity', *IEEE Trans. Biomed. Eng.*, vol. 69, no. 12, pp. 3601–3611, Dec. 2022, doi: 10.1109/TBME.2022.3174509.

[14] D. Marzorati, A. Dorizza, D. Bovio, C. Salito, L. Mainardi, and P. Cerveri, 'Hybrid Convolutional Networks for End-to-End Event Detection in Concurrent PPG and PCG Signals Affected by Motion Artifacts', *IEEE Trans. Biomed. Eng.*, vol. 69, no. 8, pp. 2512–2523, Aug. 2022, doi: 10.1109/TBME.2022.3148171.

[15] E. Beqiri *et al.*, 'Common Data Elements for Disorders of Consciousness: Recommendations from the Working Group on Physiology and Big Data', *Neurocrit. Care*, vol. 39, no. 3, pp. 593–599, Dec. 2023, doi: 10.1007/s12028-023-01846-7.

[16] R. Hogan *et al.*, 'Scaling convolutional neural networks achieves expert level seizure detection in neonatal EEG', *Npj Digit. Med.*, vol. 8, no. 1, p. 17, Jan. 2025, doi: 10.1038/s41746-024-01416-x.

[17] D. Conen *et al.*, 'Age-Specific Differences Between Conventional and Ambulatory Daytime Blood Pressure Values', *Hypertens. Dallas Tex 1979*, vol. 64, no. 5, pp. 1073–1079, Nov. 2014, doi: 10.1161/HYPERTENSIONAHA.114.03957.

[18] A. E. W. Johnson *et al.*, 'MIMIC-III, a freely accessible critical care database', *Sci. Data*, vol. 3, no. 1, p. 160035, May 2016, doi: 10.1038/sdata.2016.35.

[19] S. Singh, M. Kozłowski, I. García-López, Z. Jiang, and E. Rodriguez-Villegas, 'Proof of Concept of a Novel Neck-Situated Wearable PPG System for Continuous Physiological Monitoring', *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021, doi: 10.1109/TIM.2021.3083415.

[20] Y. Liu, H. Wu, J. Wang, and M. Long, 'Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting', presented at the Advances in Neural Information Processing Systems, Oct. 2022. Accessed: May 30, 2025. [Online]. Available: https://openreview.net/forum?id=ucNDIDRNjjv

[21] M. Czosnyka, P. Smielewski, P. Kirkpatrick, R. J. Laing, D. Menon, and J. D. Pickard, 'Continuous Assessment of the Cerebral Vasomotor Reactivity in Head Injury', *Neurosurgery*, vol. 41, no. 1, p. 11, July 1997.

[22] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, 'Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection', in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, in ICMI '18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 400–408. doi: 10.1145/3242969.3242985.

[23] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, 'Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks', *Sensors*, vol. 19, no. 14, Art. no. 14, Jan. 2019, doi: 10.3390/s19143079.

[24] B. Williams *et al.*, '2018 ESC/ESH Guidelines for the management of arterial hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH)', *Eur. Heart J.*, vol. 39, no. 33, pp. 3021–3104, Sept. 2018, doi: 10.1093/eurheartj/ehy339.

[25] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, 'Support Vector Method for Novelty Detection', in *Advances in Neural Information Processing Systems*, MIT Press, 1999. Accessed: Dec. 26, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/1999/hash/8725fb777f25776ffa9076e44fcfd776-Abstract.html

[26] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, 'Very deep convolutional neural networks for raw waveforms', in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 421–425. doi: 10.1109/ICASSP.2017.7952190.

[27] T. Chen and C. Guestrin, 'XGBoost: A Scalable Tree Boosting System', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[28] H. Haule, I. Piper, P. Jones, C. Qin, T.-Y. M. Lo, and J. Escudero, 'VAE-IF: Deep feature extraction with averaging for fully unsupervised artifact detection in routinely acquired ICU time-series', *Comput. Biol. Med.*, vol. 186, p. 109610, Mar. 2025, doi: 10.1016/j.compbiomed.2024.109610.

[29] Z. Nowroozilarki, B. J. Mortazavi, and R. Jafari, 'Variational Autoencoders for Biomedical Signal Morphology Clustering and Noise Detection', *IEEE J. Biomed. Health Inform.*, vol. 28, no. 1, pp. 169–180, Jan. 2024, doi: 10.1109/JBHI.2023.3320585.

[30] K. JM, M. DM, and B. JG, 'Optimized Arterial Line Artifact Identification Algorithm Cleans High-Frequency Arterial Line Data With High Accuracy in Critically Ill Patients.', *Crit. Care Explor.*, vol. 4, no. 12, p. e0814, Dec. 2022, doi: 10.1097/CCE.0000000000000814.

[31] W. Zong, T. Heldt, G. B. Moody, and R. G. Mark, 'An open-source algorithm to detect onset of arterial blood pressure pulses', in *Computers in Cardiology, 2003*, Thessaloniki Chalkidiki, Greece: IEEE, 2003, pp. 259–262. doi: 10.1109/CIC.2003.1291140.

[32] D. Pollreisz and N. TaheriNejad, 'Detection and Removal of Motion Artifacts in PPG Signals', *Mob. Netw. Appl.*, vol. 27, no. 2, pp. 728–738, Apr. 2022, doi: 10.1007/s11036-019-01323-6.

[33] V. G. Almeida *et al.*, 'Machine Learning Techniques for Arterial Pressure Waveform Analysis', *J. Pers. Med.*, vol. 3, no. 2, pp. 82–101, May 2013, doi: 10.3390/jpm3020082.

[34] C. Tronstad, I. N. Omenås, and L. A. Rosseland, 'An improved artifact removal algorithm for continuous cardiac output and blood pressure recordings', *2015 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC*, pp. 5968–5972, Jan. 2015, doi: 10.1109/EMBC.2015.7319751.

[35] L. D, G. L, P. M, L. JC, L. Y, and H. J, 'A Deep Learning-Based Automated Framework for Subpeak Designation on Intracranial Pressure Signals.', *Sensors*, vol. 23, no. 18, p. nan, Sept. 2023, doi: 10.3390/s23187834.

[36] C. Mataczynski, A. Kazimierska, A. Uryga, M. Burzynska, A. Rusiecki, and M. Kasprowicz, 'End-to-End Automatic Morphological Classification of Intracranial Pressure Pulse Waveforms Using Deep Learning', *IEEE J. Biomed. Health Inform.*, vol. 26, no. 2, pp. 494–504, Feb. 2022, doi: 10.1109/JBHI.2021.3088629.

[37] D. CH, G. D, and T. A, 'Error-checking intraoperative arterial line blood pressures.', *J. Clin. Monit. Comput.*, vol. 33, no. 3, pp. 407–412, June 2019, doi: 10.1007/s10877-018-0167-7.

[38] M.-T. I, W. JE, J. M, and A. M, 'Empirical Mode Decomposition-Based Method for Artefact Removal in Raw Intracranial Pressure Signals.', *Acta Neurochir. Suppl.*, vol. 131, pp. 201–205, 2021, doi: 10.1007/978-3-030-59436-7_39.

[39] M. Wu, P. Branco, J. X. C. Ke, and D. B. MacDonald, 'Artifact Detection in Invasive Blood Pressure Data using Forecasting Methods and Machine Learning', *2020 IEEE Int. Conf. Bioinforma. Biomed. BIBM*, pp. 843–850, Jan. 2020, doi: 10.1109/BIBM49941.2020.9313540.

[40] J. Xu, H. Wu, J. Wang, and M. Long, 'Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy', June 29, 2022, *arXiv*: arXiv:2110.02642. Accessed: Dec. 06, 2023. [Online]. Available: http://arxiv.org/abs/2110.02642

[41] A. Deng and B. Hooi, 'Graph Neural Network-Based Anomaly Detection in Multivariate Time Series', *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 5, Art. no. 5, May 2021, doi: 10.1609/aaai.v35i5.16523.

[42] E. Yanar and Y. S. Dogrusoz, 'False Ventricular-Fibrillation/Flutter Alarm Reduction of Patient Monitoring Systems in Intensive Care Units', *2018 IEEE Int. Symp. Med. Meas. Appl. MeMeA*, pp. 01-May, Jan. 2018, doi: 10.1109/MeMeA.2018.8438601.

[43] D. P. Kingma and M. Welling, 'Auto-Encoding Variational Bayes', Dec. 10, 2022, *arXiv*: arXiv:1312.6114. doi: 10.48550/arXiv.1312.6114.

[44] A. Das, W. Kong, R. Sen, and Y. Zhou, 'A decoder-only foundation model for time-series forecasting', presented at the Forty-first International Conference on Machine Learning, June 2024. Accessed: Dec. 03, 2025. [Online]. Available: https://openreview.net/forum?id=jn2iTJas6h

[45] X. Chen, Z. Song, D. Ji, S. Gao, and L. Zhu, 'SID: Multi-LLM Debate Driven by Self Signals', Oct. 08, 2025, *arXiv*: arXiv:2510.06843. doi: 10.48550/arXiv.2510.06843.

[46] G. Lee, W. Yu, K. Shin, W. Cheng, and H. Chen, 'TimeCAP: Learning to Contextualize, Augment, and Predict Time Series Events with Large Language Model Agents', *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 17, pp. 18082–18090, Apr. 2025, doi: 10.1609/aaai.v39i17.33989.