Predicting Hemodynamic Parameters based on Arterial Blood Pressure Waveform Using Self-Supervised Learning and Fine-Tuning

Ke Liao[1], Armagan Elibol[2], Ziyan Gao[1], Lingzhong Meng[3], Nak Young Chong[1*]

[1*]The School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, 923-1292, Ishikawa, Japan.
[2]IAS-8: Data Analytics and Machine Learning, Institute for Advanced Simulation, Forschungszentrum Jülich GmbH, Wilhelm-Johnen-Straße, Jülich, 52428, North Rhine-Westphalia, Germany.
[3]The Department of Anesthesia, Indiana University School of Medicine, 340 West 10th Street, Fairbanks Hall, Suite 6200, Indianapolis, 46202-3082, Indiana (IN), USA.

*Corresponding author(s). E-mail(s): nakyoung@jaist.ac.jp;
Contributing authors: nicolekliao@gmail.com; a.elibol@fz-juelich.de; ziyan-g@jaist.ac.jp; menglz@iu.edu;

Abstract

The arterial blood pressure waveform (ABPW) serves as a less invasive technique for evaluating hemodynamic parameters, offering a lower risk compared to the more invasive pulmonary artery catheter (PAC) thermodilution method. Various studies suggest that deep learning models can potentially predict the hemodynamic parameters of ABPW. However, the scarcity of ground truth data restricts the accuracy of these models, preventing them from gaining clinical acceptance. To mitigate this data and domain challenge, this work proposed a self-supervised generative learning model for hemodynamic parameter prediction, called SSHemo (Self-Supervised Hemodynamic model). Specifically, SSHemo suggests first to leverage large amounts of unlabeled ABPW data to learn the representative embedding and then to fine-tune for the downstream task with a small amount of hemodynamic parameters' ground truth. To verify the effectiveness of SSHemo, we utilize the public available VitalDB data set to train the model, and evaluation was conducted on two public datasets: VitalDB and MIMIC. The experimental

results reveal that SSHemo's regression mean absolute error (MAE) improved significantly from **1.63** L/min to **1.25** L/min when predicting cardiac output (CO). The trending tracking ability for CO changes meets clinical acceptance (radial limit of agreement (LOA) is **±25.56°**, less than **±30°**). In addition, SSHemo demonstrates robust stability in various conditions and cohorts, as evidenced by subgroup analysis, varying range of systemic vascular resistance (SVR) analysis, and rapid CO analysis, compared to the most widely used commercial devices, the EV1000. Computational analysis further underscores the value and potential of practical application of the model in various settings.

Keywords: Hemodynamic prediction; ABPW; self-supervised learning

1 Introduction

Hemodynamic monitoring is the study of blood flow and circulation in the body and plays an important role in perioperative and critical care, acting as a valuable tool for the diagnosis and management of critically ill patients [1, 2]. Among hemodynamic variables, CO is crucial, as the precision of numerous other hemodynamic measures depends on the precision of the estimated CO. Table 1 presents an overview of all hemodynamic variables in conjunction with their respective measurement or calculation methods.

Although CO measurement is crucial, the current gold standard in clinical settings is the thermodilution method through a PAC, termed CO_{TD} [3], this approach is invasive and involves considerable risks associated with placement, which limits its use primarily to cardiac surgeries, liver transplants, and certain critically ill patients. In contrast, recent advances have led to the development of minimally invasive or non-invasive techniques such as Doppler of the aortic, pressure waveform analysis, and bioimpedance [4]. Among these methodologies, CO derived from ABP waveforms (CO_{ABP}) is gaining popularity due to their lower risk of complications. This approach is particularly advantageous, as cannulation of the radial artery is commonly performed in surgical and ICU settings, providing easy access to the necessary data.

CO_{ABP} is typically derived by modeling the dynamics between blood flow, blood pressure, and power, based on the analysis of waveform signals. This process requires extensive experiments to achieve calibration and determine correction factors [5–8]. On the other hand, machine learning-based models [9–12] have recently been introduced for advanced time series and waveform analysis of ABPW, allowing accurate prediction of hemodynamic parameters without a non-trivial procedure. However, data acquisition remains necessary, particularly the CO_{TD} records, which are difficult to obtain due to the invasive nature of the procedures involved, making data acquisition a persistent obstacle that must be addressed. Meanwhile, as a vital signs monitor, the ABP waveform is a common signal with widespread adoption in medical settings.

The ABPW method involves the analysis of the ABP waveform, which is based on the modeling of the waveform and the examination of its characteristics. This

Parameter	Equation	Unit
Cardiac Output (CO)	Measured or $SV \cdot HR/1000$	L/min
Heart Rate (HR)	Measured	bpm
Systolic Blood Pressure (SBP)	Measured	mmHg
Mean Arterial Pressure (MBP)	Measured or $(SBP + 2 \cdot DBP)/3$	mmHg
Diastolic Blood Pressure (DBP)	Measured	mmHg
Central Venous Pressure (CVP)	Measured	mmHg
Right Atrial Pressure (RAP)	Measured	mmHg
Pulmonary Artery Wedge Pressure (PAWP)	Measured	mmHg
Cardiac Index (CI)	CO/BSA[1]	$L/min/m^2$
Stroke Volume (SV)	$(CO/HR) \cdot 1000$	ml/beat
Stroke Volume Index (SI)	$(CI/HR) \cdot 1000$	$ml/beat/m^2$
Systemic Vascular Resistance (SVR)	$80 \cdot (MBP - RAP)/CO$	$dynes \cdot sec/cm^5$
Systemic Vascular Resistance Index (SVRI)	$80 \cdot (MBP - RAP)/CI$	$dynes \cdot sec/cm^5/m^2$
Pulse Pressure (PP)	SBP-DBP	mmHg
Systolic Pressure Variation (SPV)	max(SBP)-min(SBP)	mmHg
Pulse Pressure Variation (PPV)	$max(PP) - min(PP)/[(max(PP) + min(PP))/2] \cdot 100$	%
Stroke Volume Variation (SVV)	$max(SV) - min(SV)/[(max(SV) + min(SV))/2] \cdot 100$	%

[1]BSA (Body Surface Area) is used in medicine to estimate the total surface area of a person's body.

Table 1 The measurement and estimation of hemodynamic parameters. [13, 14]

approach enables a more detailed understanding of the waveform signal, allowing the identification of specific patterns and trends that can affect hemodynamic status.

To improve hemodynamic status prediction, it is crucial to overcome the limitation of requiring a substantial amount of labeled data (gold standard: CO_{TD}). The abundance of ABP waveforms in surgical and ICU settings, as generated by monitoring equipment, presents an opportunity for analysis. Using these unlabeled data, we can pre-train models to recognize basic waveform patterns, similar to how natural language processing (NLP) models predict subsequent tokens based on historical context. This approach enables models to have essential waveform recognition capabilities, thus unlocking the full potential of ABP waveform analysis and significantly improving predictive model performance. This method also facilitates the utilization of previously underutilized data sources, resulting in more accurate and reliable predictions of hemodynamic status. This research work made several contributions.

- This research focuses on the use of the ABP waveform as input to forecast hemodynamic parameters as output. To address the challenge of limited ground truth data, this paper introduces a novel self-supervised learning approach, termed SSHemo (Self-Supervised Hemodynamic predicting model), which offers a promising solution to this issue.
- The pre-trained stage of SSHemo employs a pretext-generative and waveform forecast strategy, leveraging unlabeled historical waveform data to predict subsequent waveform data. Using VitalDB public data sets, a total data set of $3,500$ surgical cases comprising $41,733,506$ samples is used for the waveform forecast task, divided into training (($2,800$ training cases with $32,685,454$ samples), validation (350 validating cases with $4,470,710$ samples), and testing sets (350 testing cases with $4,577,342$ samples) to facilitate model development and evaluation.
- The SSHemo model undergoes fine-tuning for the downstream task - CO prediction, using the limited CO values obtained from the pulmonary artery thermodilution (CO_{TD}), the gold standard for CO monitoring. Using a data set comprising 57 surgery cases with $242,496$ samples, the model is fine-tuned and evaluated, with 47 cases (with $202,091$ samples) used for training, 5 for validation (with $18,807$

samples), and 5 for testing (with 21,598 samples), reflecting the approach used in VitalDB public datasets.

The SSHemo model's performance was assessed in terms of consistency and tracking capabilities. In particular, SSHemo outperformed supervised learning models in both aspects. It achieved MAE of 1.25 L/min, exceeding the MAE of the supervised learning model of 1.63. Furthermore, SSHemo demonstrated superior tracking abilities with a radial LOA of $\pm 25.56^{\circ}$, meeting the clinical standard of $< \pm 30.00^{\circ}$ and outperforming the radial LOA of the supervised learning model of $\pm 37.19^{\circ}$. SSHemo's performance is also evaluated in various clinical scenarios, from minimally invasive surgery to the ICU, with MAE of 1.52 L/min with MIMIC-II and MIMIC-IV datasets, respectively, demonstrating its generalizability in various clinical scenarios.

Furthermore, SSHemo's performance is compared to commercial hemodynamic monitoring devices, including the EV1000 with FloTrac algorithm. Additional analyzes, such as subgroup and varying SVR range evaluations, as well as rapid CO evaluations, demonstrate SSHemo's stability across diverse conditions and cohorts, compared to EV1000 and CO_{TD}. Computational analysis also reveals that this model has practical application and potential value.

2 Related Work

In a previous deep learning-based study, Yang *et al.* [15] used SSL and the transfer learning method to estimate SV or CO values based on the ABP waveform. However, in the pre-training tasks, their work uses the labeled SV values from a commercial APCO device (Vigileo or EV1000 with the FloTrac algorithm). It also relies on labeled data collection, albeit with relatively easier-to-collect labels. Moreover, Yang's pre-trained work relies heavily on the CO or SV from EV1000, which makes its performance highly dependent on EV1000. Meanwhile, the downstream tasks were limited to predicting hemodynamic parameters, which restricted the applicability of the model.

In [16], a CNN-based network has been proposed to predict CO from ABPW data. The training data set consists of labeled Vigilance II data, which provides the gold standard ground truth. The final data set has 293 cases with 232,148 samples for training, 125 cases with 99,492 samples for validation, and 69 cases with 154,555 samples for testing. The performance of the supervised learning method highly depends on the amount of labeled data, which is not only expensive to collect, but also scarce depending on the specific clinical scenarios.

Moon *et al.* [11] built a deep learning-based ABPW algorithm to predict SV values using data from liver transplant patients. However, their work considers the patient information using the individual scale coefficient (ISC) to scale, and the number of training data sets (34 patients, 484,384 samples) is also limited due to the scarce ground truth data.

Kwon *et al.* [12] use the 1D CNN model to estimate the variation of the stroke volume (SVV) from ABPW. In that study, the labeled SVV value used is from a commercial device EV1000. Although the data set in this investigation is improved to 8,512,564 data sets from 557 patients (training (n = 210, data = 3,620,386),

validation (n = 217, data = 3,944,244) and test (n = 130, data = 947,954)), the ground truth may still be associated with the unreliable estimation of SVV.

All the aforementioned works [11, 12, 15, 16] have introduced the deep learning method to the hemodynamic domain, especially the estimation based on ABPW, but some issues still need to be addressed. Research efforts on how to deal with the scarcity of gold standard ground truth is one of the important aspects to be investigated further. Similarly, with a small number of high-quality ground truths, it is worthwhile to investigate improving the prediction accuracy and compare the performance of different deep learning frameworks based on the waveform data. The generalization of deep learning models, when they are extended to different clinical scenarios, also needs to be studied further. This study aims to take a further step in the discussion of these issues.

3 Methods

AI or data-driven models suffer from bias in the data, especially the gold standard or labeled data. In the healthcare or clinical domain, high-quality and valued labeled data require clinical expert annotation, unique experiments, or special measurements, which is costly and time consuming. Using limited labeled data with expert information and achieving high performance is a critical issue in the healthcare and clinical domain [17].

Since valuable data labeled with expert domain knowledge is scarce [18] and the amount of labeled data set affects the final performance [19], it is essential to solve data issues. There are many methods that deal with a small amount of labeled data: semi-supervised learning [20], transfer learning [21], active learning [22], self-supervised learning [23], few-shot learning [24], data augmentation [25], *etc.*

Self-supervised learning (SSL) is a machine learning technique in which a model learns to make predictions about certain aspects of the input data without explicit supervision. The purpose of this project is to learn useful representations from unlabeled data by defining pretext tasks that do not require manual annotations. Furthermore, the learned representations are fine-tuned on a small amount of labeled data for the target task. In recent years, labeled data have gained popularity because they are limited, expensive, or difficult to obtain, but there is an abundance of unlabeled data that can be used to learn useful representations. Therefore, SSL is widely used in NLP, video processing, speech recognition, medical image analysis, *etc* [15, 26].

Inspired by the SSL approaches used in language models such as BERT [27], waveform data are also treated as a unique "language" of sensor monitors. An automatic self-generated labeling mechanism is utilized to mask the future waveform portions and predict them using previous waveform data (past predictions of the future), eliminating the need for any labeled data. The auto-labeling mechanism enables the effective utilization of large amounts of unlabeled data for pretext tasks, which can subsequently be fine-tuned with a limited set of labeled data to achieve optimal performance in a specific downstream task. Specifically, in this research, we designed SSHemo, a self-supervised learning framework to improve the prediction performance of hemodynamic parameters with limited data labeled with the gold standard and a large amount of

waveform data, as shown in Fig. 1. The model for predicting hemodynamic parameters is based on the ABPW method. SSHemo designs a model for a regression task to predict hemodynamic parameters (CO, HR, SBP, MBP, DBP) using a 10 second waveform with a sliding time window. Our research draws on previous work and clinical experience, using a 10 second ABP waveform as input data. This approach aims to minimize the effects of intrathoracic pressure fluctuations caused by respiratory periods on arterial pressure waveforms, thus improving the precision of our findings [11, 28]. The training process is structured into two stages. Initially, a large dataset of unlabeled ABP waveform data is utilized in a pretext generation and waveform forecasting task to learn meaningful representations of waveforms through a sliding time window. This enables accurate pattern recognition and interpretation of the implied hemodynamic information from the biological waveform signal. Subsequently, the model is fine-tuned with a small amount of labeled data (the golden standard data) to estimate hemodynamic parameters through a regression task, allowing for precise prediction of hemodynamic parameters.

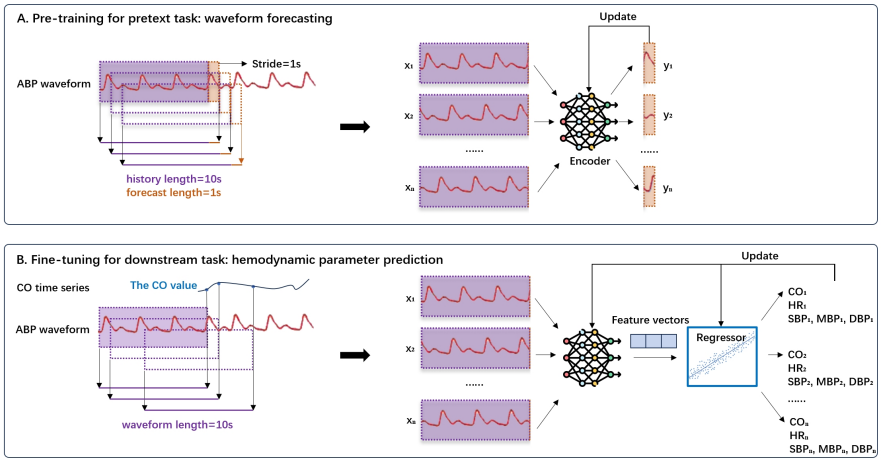

Fig. 1 The schematic diagram of the framework and pipeline of SSHemo.

The framework is described in the following three parts: the definition of the pretext task and the downstream task; input data preparation and preprocessing; and model pretraining and fine-tuning.

3.1 The definition of pretext task and downstream task

3.1.1 Pretext task and generalization

In the field of hemodynamic estimation based on ABPW, algorithms that predict hemodynamic parameters are based on the theoretical assumption that the characteristics of the ABP waveform reflect the underlying hemodynamic state. Accurate feature extraction from biological waveform signals is crucial for model performance, enabling precise recognition of patterns and interpretation of hemodynamic status.

Key features, such as peak and trough points representing SBP and DBP, part area under the pressure curve identifying MBP, and dicrotic notch indicating aortic valve closure. Traditional mathematical and fluid mechanics modeling [29], as well as feature engineering [9], are previous approaches to improving feature extraction and improving hemodynamic parameter prediction.

Our objective in the pre-trained stage is to develop a waveform forecasting model that captures the characteristics and representations of blood pressure waveforms, providing information on the hemodynamic status of a patient and the underlying physiological processes. To achieve this, we design a mechanism that generates auto-label data and creates pretext tasks based on a sliding time window [30], where the model predicts the future waveform based on historical data within a specified time frame, specifically forecasting the next 1 second waveform from a 10 second waveform input.

The reason why a waveform forecasting task is defined as using a second waveform 10 to predict the second waveform 11 is to reduce the impact of respiratory periods on hemodynamic data. Mechanical ventilation and spontaneous breathing cause fluctuations in intrathoracic pressures, affecting heart function. The increase in inspiratory pressure has a dual effect, initially increasing left ventricular stroke volume and arterial pulse pressure, while decreasing right ventricular stroke volume. These fluctuations are used to assess the sufficiency of ventricular filling, particularly in critical care and anesthesiology, where heart-lung interactions are crucial. To estimate hemodynamic parameters, a time window of at least one respiratory cycle (typically 3-5 seconds) is considered, with previous studies using input vectors ranging from 10-20 seconds to account for standard respiratory rates and reduce the impact of respiratory periods on hemodynamic data. Our research adopts a 10-second time window, balancing real-time responsiveness with the need to cover most types of patient, allowing timely and efficient processing. More specifically, a vector length of 5-10 seconds is typically sufficient to cover one complete respiratory cycle, allowing accurate estimation of CO. However, to account for variations in respiratory cycles, particularly in severely ill patients, a 10-second time window is often adopted. This duration balances the need for accurate estimation while minimizing the impact of respiratory cycles, allowing timely and efficient estimation of CO without compromising real-time changes.

Meanwhile, the prediction window is set at 1 second, aligned with the average heartbeat duration of 0.8 to 1 second. This prediction time window effectively captures most of the useful features within a heartbeat cycle, providing a comprehensive representation of cardiovascular characteristics. By accurately forecasting the ABP waveform for one second, the model can effectively capture key characteristics or information involved of the cycle characteristics from heart beats, making it applicable to most downstream tasks, clinical scenarios, and covering almost all kinds of patient. Specific, analyzing heartbeat cycles, and vector length play a crucial role in determining the model's learning capacity. A vector length of less than 1 second often fails to capture a complete cycle, compromising the model's ability to comprehend its features. Conversely, a vector length exceeding 1 second encompasses multiple cycles but may decrease the model's responsiveness and potentially introduce respiratory cycle interference, ultimately impacting predictive accuracy.

Therefore, the waveform forecasting task is designed to utilize a 10-second input waveform to predict the subsequent 1-second waveform, thus mitigating the impact of respiratory periods on hemodynamic data. This configuration enables the model to maintain its responsiveness and real-time capability, yielding accurate and timely outputs. By setting the input vector to 10 seconds and the output vector to 1 second, the model effectively captures both respiratory and cardiac cycles, minimizing the influence of respiratory cycles on CO and allowing extraction of characteristics from the entire cardiac cycle.

The dataset in this stage is automatically generated from the input waveform data itself using a sliding time window, without the need of manually labeled data. Data on ABPW and related vital signs are commonly collected by vital signs monitors, hemodynamic monitors , *etc.*. Therefore, the amount of data on blood pressure waveforms is very large.

The model learns useful representations of the data without explicit supervision based on the data generalization process and the pretext task. Based on a large amount of unlabeled ABPW data from vital signs or hemodynamic monitors, a pre-trained model with good capabilities of representation of waveform signals was trained.

3.1.2 Downstream task and fine-tuning

The ultimate goal of SSL is to take advantage of the representations learned during the pre-training phase to improve performance on downstream tasks that require labeled data. Therefore, the purpose of the downstream task is to transfer the waveform representation capabilities to other tasks. At this stage, we use the gold-standard labeled data to achieve the tuning targets as a regression issue.

Specifically, we used 10 second ABPW data to predict hemodynamic parameters (CO, as well as HR, SBP, DBP, MBP) at the last timestamp in the time window and used expert knowledge [31, 32] to calculate other hemodynamic parameters (CI, SV, SI, SVR, SVRI, PPV, and SVV, *etc.*), as shown in Table 1. The labeled data are CO_{TD} (from the hemodynamic monitor using thermodilution methods [4]) and HR, SBP, DBP, MBP (from the vital signs monitor or the hemodynamic monitor).

The subsequent task is the hemodynamic parameter prediction task. The label data comes from the gold standard hemodynamic parameters. Among hemodynamic parameters, CO is considered the most important and others could be inferred from the combination of CO and some common vital signs (HR, SBP, MBP, DBP and similar others). The gold standard technique in CO estimation is the thermodilution method according to PAC Swan-Ganz to measure the temperature change due to blood flow [33]. This method is limited in surgery or the ICU with a poor and invasive prognosis for patients. Therefore, the number of labeled hemodynamic parameters is small and a scarce resource [34].

3.2 Data preparation

3.2.1 Data collection

In this research work, we utilize the VitalDB public data set for training and testing. VitalDB [35] is a public database collected by Vital Recorder [36], an open source

tool to collect multi-parameter vital signs for surgery and anesthesia. This data set contains both time series data (1 data points per $1-7$ seconds) and high resolution waveform signal data (62.5–500 Hz) from $6,388$ surgery cases at Seoul National University Hospital, Seoul, Korea from August 2016 to June 2017. 12 kinds of monitoring waveform records, 184 intraoperative physiological numeric data tracks, 74 perioperative clinical parameters, and 34 laboratory results. The total number of data tracks is $486,451$ with an average of 2.8 million data points for each case.

According to the requirement of our research, this data set provides ABPW (up to 500 Hz) from the monitor $Solar^{TM}8000M$ (vital sign monitor of GE Healthcare, Chicago, IL, United States) with an ADC equipped with the TramRac-4A module (an analog to digital converter). Vigilance II (hemodynamic monitor of Edwards Lifesciences) provides the labeled gold standard CO parameters. The EV1000 clinical platform (hemodynamic monitor of Edwards Lifesciences, Irvine, CA, United States) provides continuous CO values based on the commercial FloTrac algorithm. Other parameters, such as HR, SBP, MBP and DBP, are provided by the monitor $Solar^{TM}8000M$. An example of segmentation is shown in Fig. 2.

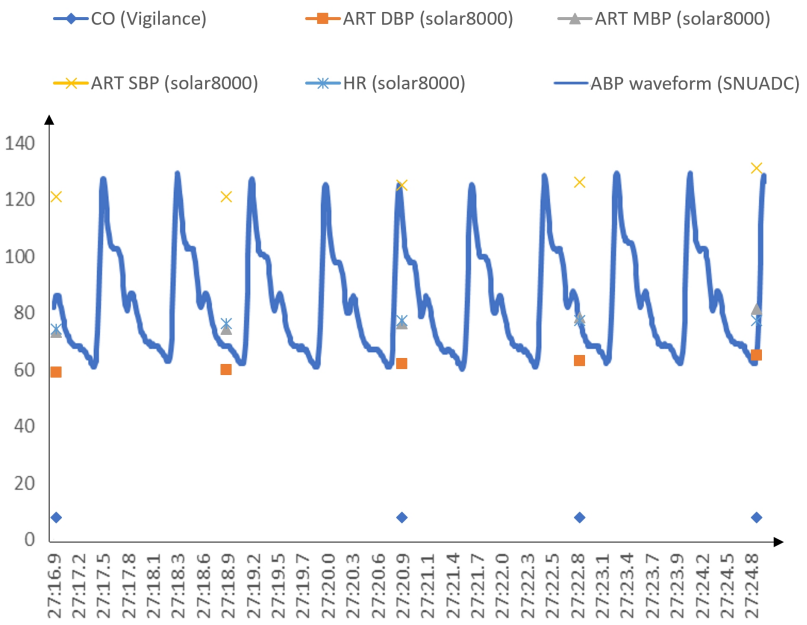

Fig. 2 The example of one data segmentation in VitalDB, showing the sampling situation of real devices.

To extend the diversity of experimental scenarios and to further evaluate the generalization performance of the model, additional datasets,, the MIMIC series datasets,

are used. In this research, we used MIMIC-IV and MIMIC-II public ICU datasets to expand the evaluation.

MIMIC-IV [37] is the fourth version of the MIMIC series dataset, which was released on PhysioNet [38] in 2021 for the ICU (Intensive Care Unit) scenario. It contains patient measurements, orders, diagnoses, procedures, treatments, and free-text deidentified clinical notes, with a decade of admissions between 2008 and 2019, from the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts, USA. Combined with MIMIC-IV, the MIMIC-IV waveform database [39] collects physiological signals and measurements from patients in the ICU, including electrocardiograms, photoplethysmograms, respiration, invasive and non-invasive blood pressure, *etc.* These two databases are linked with *subject_id* and *hadm_id*. MIMIC-IV and MIMIC-IV waveform have almost $180,000$ patients (statistical analysis of the subject ID from admission table) with high-frequency monitoring waveform signals (ECG (Electrocardiogram) waveform, PPG (Pleth, Photoplethysmography), ABP, and Respiration), low-resolution physiological time series records (HR, SBP, MBP, DBP, RR (Respiratory Rate), Temperature, SpO_2 (Oxygen saturation), *etc.*), clinical events, Laboratory events, demographic information, *etc.* The waveform data are recorded in the MIMIC-IV waveform database. The time series data are recorded both in the table *chartevents* of the module *icu* from the MIMIC-IV database and in the numeric files in the MIMIC-IV database.

The MIMIC-II Waveform Database Matched Subset [40] is the second version of the MIMIC series data set. It contains $4,897$ waveform records and $5,266$ numeric records of bedside patient monitors in the ICU of the MIMIC-II Waveform Database, which have been matched and time-aligned with $2,809$ MIMIC-II Clinical Database records [41]. Sun *et al.* [42, 43] extracted data for CO estimation studies based on MIMIC-II. In this research, the specific dataset is utilized for external validation.

Consequently, both data sets provide ABPW (up to 125Hz), along with numeric data from time series: cardiac thermodilution output, HR, SBP, MBP and DBP. When evaluating a model based on an ICU-related dataset, it is crucial to assess its generalization ability from one scenario, such as surgery, to the broader ICU domain. This process involves validating whether the performance of the model remains consistent and reliable when applied to various aspects of intensive care beyond the specific context of surgery.

3.2.2 Data preprocessing and splitting

In the application scenarios of this research, predictive precision is prioritized at each moment, with performance and loss evaluations based on time windows. To achieve this, both the pre-trained task of waveform forecasting and the fine-tuning task of hemodynamic parameter prediction employ a sliding time window ABP waveform as input for data segmentation. Furthermore, our research assumes that vital patient information can be inferred from intrinsic waveform characteristics. By excluding patient-specific details in both pre-trained and fine-tuning phases, the model remains unbiased toward individual patients or procedures, facilitating its application across

various clinical settings. Therefore, the primary focus of data preprocessing is on waveform data, specifically within defined time windows, to ensure synchronization between various biological signals and hemodynamic parameters.

The process is divided into two parts based on the two training phases: a) the pretraining of the pretext task to learn the capabilities for representation waveform data and b) the fine-tuning of a downstream task to transfer the previous capabilities to the prediction of hemodynamic parameters. In addition, the evaluation process involves preprocessing three databases: vitalDB, MIMIC-IV, and MIMIC-II.

1. Data preprocessing and splitting for pretext task training

   The data pre-processing and splitting procedure for pre-trained and waveform forecasting is shown in Fig. 3. In this task, the VitalDB database is used for the data sources. At this stage, we are only extracting the ABPW, sliced into 11-second segments. Each sample of 11-second segments splits into two parts: the previous 10-second waveform for predicting the future waveform, and the 11th-second ABPW for ground truth. As the ABPW continues to be recorded based on monitoring devices in the clinical scenario, we define a rolling time window to split the waveform to output the 11-second segments (10s + 1s waveform) and the rolling frequency is 1s. The pretext task masks the 11-second waveform patch and uses the 10-second waveform to predict the masked part.

   As blood pressure is a generally low-frequency signal and higher resolution introduces noise and artifacts without adding meaningful information to the waveform, a sampling frequency of 100Hz is used in our research as a balance between capturing essential information and minimizing data processing and storage requirements. The ABPWs are resampled from 500Hz to 100Hz from the VitalDB database in the pre-processing procedure.

   In practice, ABPW data often contains noise and abnormalities (NaN, negative values, values exceed the range of common clinical sense, etc.). To address this, deep neural networks are utilized, taking advantage of their complex structures and non-linear activation functions to learn underlying patterns and exhibit noise tolerance. Meanwhile, the SSL process further enhances the adaptation to noisy data. To prepare the ABP waveform data for modeling, a minimalist approach is employed, involving only resampling and zero-fill methods to ensure correct input for the model. This straightforward technique enables efficient processing of diverse waveforms, streamlining model training and inference. By minimizing preprocessing, the model's generalization capabilities and practical problem-solving abilities are enhanced, allowing it to handle noise or these abnormal signals and perform effectively in practical scenarios with diverse and unpredictable data.

   Finally, out of all $6,388$ cases in vitalDB, we can obtain effective waveform data for $3,500$ cases, while $2,881$ cases have no ABPW data and 7 cases have NaN or negative data.

2. Data preprocessing and splitting for downstream task training

   The data pre-processing and splitting procedure for the downstream task: fine-tuning for waveform regression is shown in Fig. 4. The objective is to predict the last time-stamp hemodynamic values (CO_{TD}, HR, SBP, DBP, MBP) using a continuous

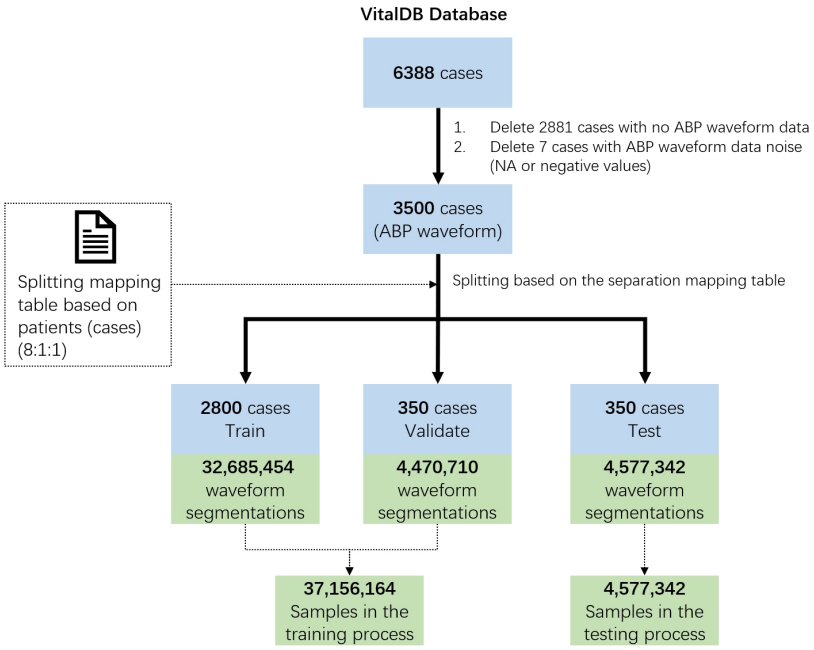

Fig. 3 Data pre-processing and splitting procedure in the pretext task. i) The data used in this task is the ABPW for the forecasting task. ii) The data is split by patients or cases and shared with the next task: the regression task in the fine-tuning stage. This partitioning method can help mitigate overfitting issues caused by data leakage in medical prediction tasks, ensuring better model generalization and evaluation [44]. iii) The $32,685,454$ data samples are for model training; $4,470,710$ data samples are for model validation and early stopping to adjust the hyper-parameters; $4,577,342$ data samples are for model testing.

10-second ABPW. Therefore, the input data is the 10-second ABPW. The labeled data includes hemodynamic parameters such as CO, HR, SBP, MBP, and DBP.

The waveform data is separated into the sample segmentations according to the labeled hemodynamic data's timestamp: get the last timestamp of hemodynamic data at T_0 and ABPW with a time window of 10 seconds (T_{0-10S}, T_0). Combined with the labeled values of hemodynamic parameters, one segmentation or sample contains 10s ABPW data, the hemodynamic parameters at the last timestamp of the 10s time window. According to the pretext task, the ABPW is also re-sampled to 100Hz. Similarly, in the fine-tuning task, the ABPW used is 100Hz. The ABPW process and filter method are the same as the pre-training stage.

3. Data preprocessing and splitting for testing and evaluation

This task utilizes two databases to evaluate the model's generalization ability: the VitalDB public database for the surgery scenario, and the MIMIC public database for the ICU cases, shown in Fig. 5. The first evaluation of the clinical environment is surgery with minimally invasive ABP measurement. The other clinical practice extends from the operating room to the intensive care unit, where more critically ill patients are cared for with more sophisticated blood flow and

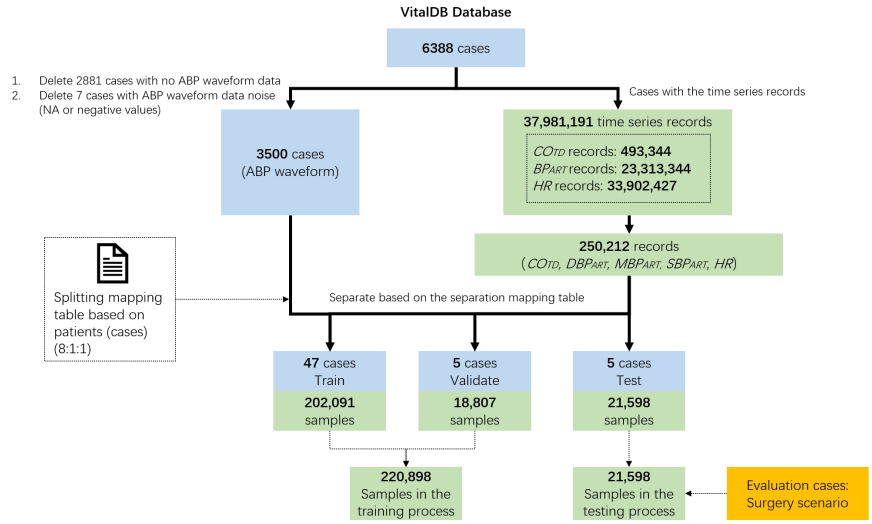

Fig. 4 Data preprocessing and split procedure in the downstream task. i) The dataset is from the VitalDB database and the data used in the multi-regression tasks is ABPW, time series records (CO_{TD} (CO from Vigilance II hemodynamic monitor), and DBP_{ART}, MBP_{ART}, SBP_{ART}, HR (DBP, MBP, SBP, and HR from GE Solar 8000 vital sign monitor)). ii) Data splitting is based on patients or cases, sharing with the previous task: the waveform forecasting in the pretext task to avoid data leakage problems. iii) $202,091$ samples are for training; $18,807$ samples are for validation and early stopping for better hyperparameters; $21,598$ samples belong to the testing case, recognized as one evaluation case: hemodynamic status prediction in the surgery scenario.

circulation in the body. The publicly available data extracted from MIMIC-IV and MIMIC-II is used to evaluate the broader performance of model training in the ICU scenario. In addition to the clinical domain, ethnicity is another variable as the patients recorded in vitalDB are Asian and the patients recorded in MIMIC-IV and MIMIC-II are American.

$21,598$ samples are used to evaluate the basic performance of SSHemo. Besides, $88,679$ samples from 24 cases are extracted from the VitalDB database using the EV1000 (a hemodynamic monitoring system with the FloTrac algorithm from Edwards Lifesciences) to compare the performance between SSHemo and other commercial hemodynamic monitors. In total, $3,103$ samples from 250 patients after the preprocessing procedure were extracted from the MIMIC-IV and the MIMIC-II database. The MIMIC-IV waveform database has an initial ABPW frequency of 64.725Hz, while the MIMIC-II database has a 125Hz ABPW. To keep the same sampling frequency, the MIMIC-IV waveform data was upsampled to 100Hz and the MIMIC-II waveform data was downsampled to 100Hz.

By combining these two databases, the evaluation scenarios extend to clinical practical cases, resulting in a comprehensive evaluation of the model's generalization ability. Moreover, the SSHemo model proposed in this paper is compared with a commercially available hemodynamic monitor device. The device is based on minimally invasive ABP monitoring and provides continuous real-time monitoring of various hemodynamic parameters [45]. As a standard commercial model, the

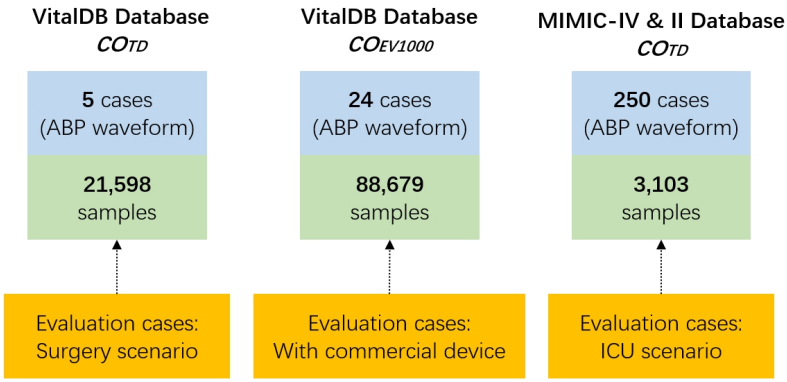

Fig. 5 Evaluation cases for various clinical scenarios or models. i) 21,598 data samples from 5 cases of surgery situations, and 3,103 data samples from 250 cases of ICU scenarios. ii) 21,598 data samples from 5 cases for SSHemo, and 88,679 data samples from 24 cases for the standard of commercial devices: Edwardz EV1000 hemodynamic monitor with the FlocTrac algorithm.

comparison offers a qualitative and intuitive assessment of the existing commercial devices.

3.3 Models

In the field of hemodynamic monitoring, particularly in the area of ABPW analysis (APWA), the model is responsible for extracting features from the waveform signal and predicting hemodynamic parameters. Extensive research has established a strong relationship between ABPW and key hemodynamic parameters such as CO, HR, BP, and SV. The extraction of features from waveform data is an essential technique that significantly improves the performance of the algorithm. There are three main types of modeling methods: circulation mechanics with fluid dynamics modeling [5, 46], feature engineering with machine learning [9, 42] and deep learning [16]. Circulation mechanics with fluid dynamics modeling is based on physical or mathematical modeling, which is sophisticated and requires large-scale experiments to calibrate and obtain correction factors [8]. Feature engineering focuses on signal analysis and extracts statistical, time, and frequency domain features from waveform data. It depends on data analysis and is easily influenced by noise and some artifacts [9, 47]. Recently, more advanced deep learning-based time series or waveform analysis models have been introduced to analyze hemodynamic parameters [11, 12, 16].

Based on the above description, both non-deep learning and deep learning models rely on a large amount of labeled data for calibration or training. The collection of gold standard labeled ground truth (thermodilution-based data) is time-consuming and difficult to collect because it requires invasive procedures. The ABPW is a kind of common data produced by vital signs monitors or hemodynamic monitors because the pressure signal is easy to capture based on the use of a minimally invasive or

non-invasive piezoelectric sensor [31]. In this research, we trained SSHemo, a self-supervised model based on an unlabeled and large amount of ABPW to get feature extraction or waveform representation capabilities. We then transferred these abilities and fine-tuned them to predict the final hemodynamic parameters. Through these two phases of processes, SSHemo does not need too much gold standard labeled data and sophisticated designed models. Meanwhile, performance shows improvement compared to supervised learning, predicting with the full use of unlabeled and labeled data.

3.3.1 Pre-trained model

In this task, the pre-trained model is a forecasting model and is responsible for extracting the features or representation from the waveform signal instead of traditional mathematical modeling and feature engineering. These models input historical data to predict future data within the sliding time window and do not require human-labeled data. A high-performance forecasting model is indicative of its strong capability to extract features or patterns from waveform data.

In the field of deep learning, the models to process the waveform data is a general aspect of the time series or sequence process research, and each network architecture (RNN (Recurrent Neural Network [48])-based, CNN (Convolutional Neural Network [49])-based, and Transformer [50]-based) has its associated representative models.

Specifically, our research addresses the input data consisting of 10 seconds of waveform samples at 100 Hz, resulting in 1000 data points. This long sequence is a significant challenge in time-series modeling. The CNN-based model achieves strong capabilities in processing long sequences by employing multiple layers to expand the receptive field, effectively capturing contextual information. Meanwhile, the transformer uses the attention mechanism to capture long-distance dependencies, allowing it to proficiently handle past memories and lengthy data sequences. Therefore, we employ two prominent backbone architectures: CNN-based and Transformer-based models, represented by InceptionTime [51] and patchTST (Patch Time Series Transformer) [52], respectively. These models are chosen for their ability to effectively process long sequences, outperforming traditional RNN-based approaches, which are often hampered by the limitations of the sequence length [53].

1. InceptionTime and CNN-based models. CNN models are known for their ability to effectively extract features from input data. Convolutional layers are used to detect various features, such as contours, inflection points, and patterns. When combined with subsequent layers, higher-level representations of the input data are produced. In the field of time-series and waveform data processing, time-series data is usually defined as one-dimensional (1D) data, and 1D-CNN can be used to process them. Recently, transformer-based architectures have shown a good ability to handle longer sequences by using receptive fields instead of memories.

   InceptionTime is a deep CNN model inspired by the Inception v4 architecture. The model uses the Inception module for different lengths of input time series. In addition, residual connection, cascading of multiple Inception modules, and global

average pooling improve the performance of the network. Compared with the HIVE-COTE algorithm, InceptionTime is more accurate and scalable.

2. PatchTST and Transformer-based models. Transformer models and their self-attention mechanisms have demonstrated strong performance and effective processes in capturing short-term or long-term dependencies and learning complex patterns within sequential data. In time series forecasting and other sequential data processing tasks, transformer models have shown promise in capturing temporal patterns, handling irregular time intervals, and making accurate predictions. Moreover, transformer-based architectures show a good ability to deal with longer sequences using attention rather than memory.

   PatchTST is a transformer-based model for long-term time series forecasting. PatchTST leverages patching and transformer structures to handle multivariate time series, extract local semantic information, reduce model complexity, and enable longer input sequences. In research experiments, PatchTST outperformed MLP-based models such as N-BEATS and N-HiTS, demonstrating superior performance with the lowest MAE and mean squared error (MSE). The model's evaluation highlights its potential in long-term forecasting tasks.

3.3.2 Fine-tuning model

During the downstream task and the fine-tuning stage, we preserve the feature processing layers to maintain the ability to process waveforms and feature extraction. We then modify the pre-trained backbone model structure, particularly the last few modules or layers, to meet the final requirements for multi-parameter regression. Modified network structure trained using a smaller amount of labeled data with pre-trained parameters, we fine-tuned the final results of the entire model. The main principle for change and fine-tuning is clearly outlined in the following content, as demonstrated in Fig. 6.

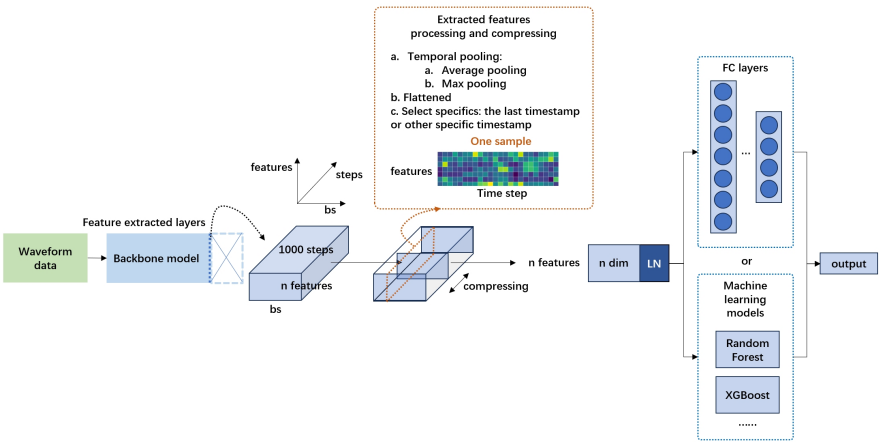

Fig. 6 The schematic diagram of the network structure and layer definition of fine-tuning.

1. The design of features extracted parts. The final model structure comprises a feature extraction module and an output module. The feature extraction module inherits from the backbone model and extracts features from the input waveform data. The output module assembles and processes the extracted features to predict multiple outputs. The features of the backbone models are usually two-dimensional with time-step information and feature information. The main part of feature extraction is to compress or extract the features and adjust the dimensionality of the feature for downstream tasks. Specifically, Global Average Pooling (GAP) layers are used for the InceptionTime backbone model and history-extracting layers for the PatchTST backbone model.
2. The design of output parts. The research objective for the fine-tuning task was shifted from waveform forecasting to multiparameter regression using labeled data (*, e.g.* CO_{TD}, HR, SBP, DBP, and MBP). One is to leverage fully connected (FC) layers to transfer features or information for the outputs. The loss function and validation metrics were adjusted to align with the final tasks. The other is to utilize features as input and other machine learning models for the following prediction tasks, such as Random Forest [54], Decision Tree [55], XGBoost [56], and similar others.

   Meanwhile, if shifting from a regression task with a single parameter, such as CO, to multiple parameter regression (e.g. CO, HR, MAP, SBP, DBP), two key adjustments are necessary to modify the output part. The model's final layer must be redesigned to accommodate multiple outputs, allowing it to efficiently process and generate results with several parameters. Furthermore, a well-structured loss function that considers multiple parameters and incorporates normalization techniques is crucial to prevent biases and ensure accurate loss calculation, with the option of using a combination of linear functions as a loss function based on the range of hemodynamic parameters, as Equation 2 shows. The weights are determined and adjusted on the basis of the range of different parameters.

$$L_{\text{total}} = \sum_i w_i L_i \tag{1}$$

$$= w_{CO}L_{CO} + w_{HR}L_{HR} + w_{SBP}L_{SBP} + w_{MBP}L_{MBP} + w_{DBP}L_{DBP} \tag{2}$$

3. Parameters freezing and fine-tuning. During the fine-tuning procedure, the feature-extracting module's parameters are initialized based on the trained parameters of pre-trained models. The remaining parameters in the output module are randomly initialized. During the fine-tuning process, there are two options for updating parameters: freezing the parameters in the feature-extracting module and training the output module's parameters, or unfreezing all parameters and updating them during training with labeled data to predict the output. The optimal choice should be based on the results of the evaluation.

4 Results

4.1 Experiments

In the pre-trained process, the input data is the first 10 seconds (100Hz × 10s=1,000 data points) and the output is the last second (100Hz × 1s=100 data points) of the 11 second waveform segmentation. In total, 3,500 surgical cases with 41,733,506 samples (2,800 training cases with 32,685,454 samples, 350 validating cases with 4,470,710 samples, and 350 testing cases with 4,577,342 samples) with a split ratio: of 8 : 1 : 1 by the patient, without overlapping patients in both cohorts, as shown in Fig. 3. During the fine-tuning process, the input data is the 10-second waveform (100Hz × 10s=1,000 data points). The output is the hemodynamic parameters (CO, HR, SBP, DBP, MBP) at the last timestamp (1 data point for uni-parameter regression or 5 data points for multi-parameter regression). In the VitalDB database, we finally extracted 57 surgery cases with 242,496 samples (47 training cases with 202,091 samples, 5 validation cases with 18,807 samples, and 5 test cases with 21,598 samples). Similarly, the separation principles are the same for pretraining and patient-based, as Fig. 4 shows.

We use two different time-series forecasting models as our backbone model: InceptionTime and PatchTST. The model structure and initial parameters are successfully taken from the backbone model in pre-trained tasks, and modified according to the fine-tuning tasks. We keep the layers and parameters for feature extraction and change the various output modules or layers of the model for the predicted purpose [57].

For both a forecasting task and a multiregression task, the loss function is MSE [58]. Adam (a gradient descent optimizer) and early stopping were used to update the parameters [59–61]. The batch size for pre-trained training is 1,024 and the batch size for fine-tuning training is 256. Model training was stopped when validation errors (MSE) stopped decreasing after 10 iterations with a dynamically changing learning rate of 0.5 times.

The GPU we used is Nvidia RTX 4090 with 24G GPU (Graphics Processing Units) memories. For the pre-trained model training, 4 GPU units are used for parallel training, and for the fine-tuning model training, 1 GPU unit is used. The experiment environment is based on Pytorch Version 2.0.1 [62], Fastai Version 2.7.13 (a deep learning library that provides the high-level components to provide state-of-the-art results in standard deep learning domains) [63], and Tsai version 0.3.7 [64] library (state-of-the-art Deep Learning Library for Time Series and Sequences).

4.2 Result analysis

4.2.1 Cohort analysis

The characteristics of the patient cohort and the dataset are shown in Table 2. In the pre-trained task, the patient demographics were similar across the training, validation, and test datasets due to the large amount of data. In the fine-tuned task, the patient demographics are slightly different among the training, validation, and testing datasets. Meanwhile, the evaluated datasets are analyzed. Based on demographic statistics of patients, the more data, the more similar the distribution of demographic information between different data sets.

Dataset	Pre-trained dataset			Fine-tuned dataset			External dataset	
	Train n[1]=32,685,454 p[2]=2,800	Validate n=4,470,710 p=350	Test n=4,577,342 p=350	Train n=202,091 p=47	Validate n=18,807 p=5	Test n=21,598 p=5	$CO_{EV\,1000}$ n=88,679 p=24	CO_{ICU} n=3,103 p=250
Data source	VitalDB	VitalDB	VitalDB	VitalDB	VitalDB	VitalDB	VitalDB	MIMIC-IV & II
Demographics								
Age[3] (years)	0-90 60 [51-70] 58 ± 15	0-90 60 [49-68] 57 ± 16	0-86 60 [50-68] 58 ± 15	18-82 55 [48-61] 54 ± 12	37-63 59 [49-61] 54 ± 11	24-73 57 [52-73] 56 ± 20	18-73 56 [49-62] 54 ± 15	37-90 70 [61-77] 65 ± 11
Sex[4] (male)	1240 (44.3%)	192 (54.9%)	206 (58.9%)	34 (72.3%)	4 (80.0%)	2 (40.0%)	18 (75.0%)	156 (62.4%)
Height[3] (cm)	42.0-188.6 162.8 [156.4-169.1] 162.4 ± 10.3	64.1-187.5 162.3 [156.2-169.0] 161.9 ± 12.9	64.1-185.2 163.7 [157.8-169.6] 163.2 ± 10.3	139.0-181.0 166.4 [161.4-170.5] 165.9 ± 8.2	153.9-176.0 167.0 [166.0-171.0] 166.8 ± 8.2	154.5-177.4 160.4 [155.4-169.5] 163.4 ± 9.8	152.5-179.0 165.7 [160.5-169.6] 165.6 ± 7.3	NA[6]
Weight[3] (kg)	4.8-111.5 60.2 [52.9-68.1] 60.9 ± 11.7	8.9-96 59.8 [52.2-68.2] 60.1 ± 12.5	8.9-139.7 61.1 [53.6-70.5] 62.4 ± 13.8	43.8-81.4 61.6 [54.1-66.2] 61.3 ± 8.9	50.0-65.9 59.2 [50.4-61.4] 57.4 ± 7.0	42.2-77.9 59.9 [58.9-63.0] 60.4 ± 12.7	42.2-78.9 59.1 [53.2-65.5] 59.6 ± 9.9	NA[6]
Hemodynamic characteristics								
HR[5] (bpm)	0-300 72 [62-83] 74 ± 17	0-306 71 [62-84] 74 ± 17	0-300 73 [63-84] 75 ± 17	0-300 82 [70-95] 83 ± 17	34-185 88 [74-97] 86 ± 15	0-174 90 [80-100] 90 ± 16	0-185 87 [74-96] 85 ± 16	0-154 86 [78-93] 85 ± 14
SBP_{ART}[5] (mmHg)	0-350 115 [102-129] 115 ± 27	0-350 115 [102-129] 115 ± 27	0-350 115 [102-129] 116 ± 26	0-350 104 [89-119] 105 ± 26	0-350 96 [85-108] 99 ± 32	0-331 95 [87-107] 97 ± 21	0-349 102 [89-116] 104 ± 25	0-197 114 [102-127] 111 ± 32
DBP_{ART}[5] (mmHg)	0-350 62 [54-70] 62 ± 19	0-350 61 [54-70] 62 ± 18	0-350 62 [54-71] 63 ± 18	0-350 51 [43-60] 53 ± 19	0-350 47 [42-54] 53 ± 34	0-350 45 [39-56] 48 ± 15	0-350 49 [42-57] 51 ± 19	0-103 55 [49-62] 54 ± 16
MBP_{ART}[5] (mmHg)	0-350 81 [72-92] 83 ± 26	0-350 81 [71-92] 83 ± 25	0-350 82 [72-92] 83 ± 25	0-350 69 [59-80] 74 ± 31	0-338 64 [57-73] 73 ± 40	0-347 62 [55-72] 66 ± 22	0-350 66 [58-77] 72 ± 30	0-291 76 [68-84] 76 ± 21
CO_{TD}[5] (L/min)	0-17.4 7.0 [5.5-8.6] 7.1 ± 2.2	3.9-15.1 6.5 [5.9-8.6] 7.4 ± 2.4	2.7-10.5 7.3 [6.1-8.4] 7.1 ± 1.8	0-17.4 7.0 [5.5-8.6] 7.1 ± 2.2	3.9-15.1 6.5 [5.9-8.6] 7.4 ± 2.4	2.7-10.5 7.3 [6.1-8.4] 7.1 ± 1.8	2.1-17.4 7.1 [5.7-8.5] 7.2 ± 2.2	1.55-19.1 4.02 [4.9-6.0] 5.1 ± 1.8

[1] n: values are expressed as the number of samples.

[2] p: values are expressed as the number of patients.

[3] Age, Height, Weight: values are expressed as min-max, median [quartile 25% - quartile 75%], mean±std.

[4] Sex: values are expressed as percent.

[5] HR, SBP_{ART}, DBP_{ART}, MBP_{ART}, CO_{TD} as measured by monitoring.

[6] MIMIC-II matched dataset misses the height and weight information, therefore there is no statistic data.

Table 2 Characteristics of datasets

4.2.2 The performance comparison between supervised learning and self-supervised learning

In this paper, we integrate three types of evaluation methods to evaluate the performance of the proposed model. The box plot shows the distribution or spread between the model prediction results and the ground truth. The final task is to predict the hemodynamic parameters for the last time in a time window. Therefore, regression task evaluation metrics are introduced to evaluate the accuracy of the model. MSE, MAE, Root Mean Squared Error (RMSE), R-squared (R2), mean absolute percentage error (MAPE), *etc.* [65] to summarize predictive capabilities.

In addition, in the hemodynamic domain, especially in the CO prediction task, the meta-analysis describes the relationship between CO_{TD} and predicted CO [66, 67]. The meta-analysis includes four plots and their associated parameters: the scatter-plot with slope and correlation coefficient (r-value); the Bland-Altman plot with bias and limits of agreement (lower LOA and upper LOA); the four-quadrant plot with concordance rate; and the polar plot with angular bias and radial LOA. The scatter plot and fitted curves illustrate the linear trend or consistency between the ground truth and the predicted value. The data points are close to the line of identity x = y to obtain good agreement, where x is the labeled data and y is the predicted results [68]. The plot Bland-Altman is a way to assess the bias between the mean difference and an agreement interval, illustrating the mean of the differences (the middle horizontal line) and the 95% LOA (thick horizontal lines top and bottom) between the reference method and the test method [69]. In addition to absolute accuracy and precision, it is also important to assess the ability to track changes in predicted values. The ability of the multi-beat analysis method to track relative changes is shown in a four-quadrant plot with a central exclusion zone of 15%, an area of nonsignificant change [70, 71]. The polar plot is used for the analysis of trend capability based on the polar coordinate transformation [72].

Each model or method is assessed using the aforementioned three types of evaluation to measure both accuracy and tracking performance.

Fig. 7 shows the evolution of performance from supervised learning to self-supervised learning.

From the results in Fig. 7, the SSL training framework shows better concordance and tracking capabilities compared to the direct supervised learning method. Consider the backbone network as a feature extractor that follows a machine learning model, such as a random forest that gets more accurate results. Meanwhile, SSL shows better generalization abilities from surgery cases to ICU cases when evaluated based on the MIMIC series dataset. The radial LOA of SSHemo is $\pm 29.38°$ and $\pm 25.56°$, within the clinically acceptable trend tracking ability for CO changes (radial LOA$\leq \pm 30°$) [31, 66].

4.2.3 The comparison between SSHemo and other commercial devices

To verify performance in the medical practical situation, a comparative performance analysis with commercial hemodynamic monitors has been devised. This experimental

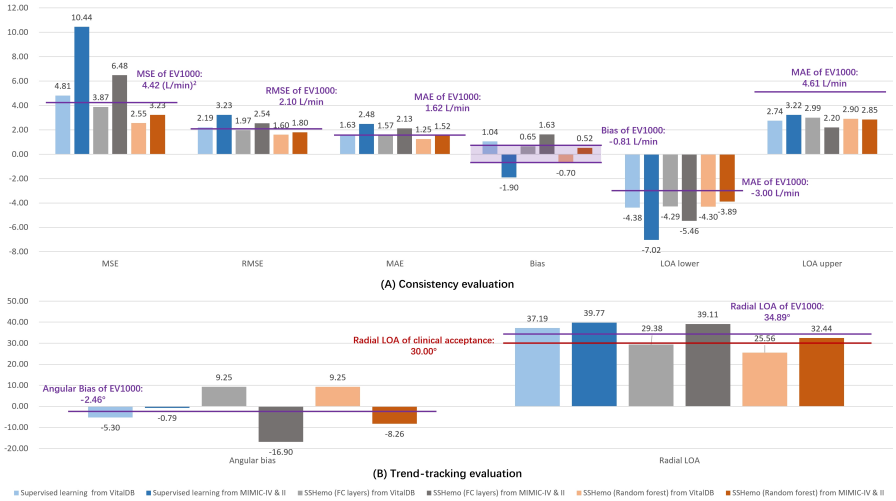

Fig. 7 The performance comparison between supervised learning and SSL, showing SSL's effect on prediction ability improvement. Supervised learning means directly training with the backbone model. i) SSHemo (FC layers) means pre-trained with pretext task and fine-tuned based on labeled data, keeping the feature-extracted layers from the backbone network and changing the output module to the FC layers for downstream prediction tasks. ii) SSHemo (Random forest) means considering the backbone network as a feature extractor and following a random forest as a regressor in the final prediction task. iii) The VitalDB dataset is for basic test and evaluation and the MIMIC series dataset is for generalization evaluation. iv) (A) is the comparison of consistency or concordance evaluation. The grey line in this sub-figure means the MAE result of the standard commercial device (EV1000) and below it, the model's performance is better than the current standard commercial device. v) (B) is the comparison of trend-tracking abilities. The orange line in this sub-figure means the radial LOA result of clinical acceptance and below it, the tracking ability meets the requirement of clinical acceptance and has real practical meanings.

framework not only assesses the capabilities of SSHemo but also replicates and builds upon existing research models in the field. Specifically, EV1000 with the FlocTrac algorithm represents the commercial standard for ABPW modeling, whereas DLAPCO and related research embody the current state of AI algorithms in this field [15, 16]. To accommodate data set limitations (VitalDB, MIMIC-II and MIMIC-IV), expand various clinical scenarios, and ensure consistent comparative conditions, the DLAPCO model has been tailored by integrating relevant research and omitting disparate elements, thus synthesizing a unified approach. A network architecture is achieved, and we call it the Yang-modified model. The Yang-modified model builds on the same network architectures with inception blocks. Key modifications include adjusting the duration of the input vector from 20 seconds to 10 seconds, eliminating the ICD-9-CM-3 codes, and shifting the output from SV to CO. These adjustments aim to tailor the model to the specific requirements of our research for comparison. Both algorithms use minimally invasive measurement techniques during surgery, and the evaluation parameters are calculated between CO_{model} (the prediction of a model) and CO_{TD}. The results are shown in Fig. 8 with the boxplot and the four meta-analysis plots.

The results indicate that the SSHemo predicted values are more closely aligned with CO_{TD}. In particular, SSHemo demonstrates better concordance performance,

with a slope of 0.30 in the scatter plot, outperforming EV1000's slope of 0.19 and Yang's modified model of 0.16. Furthermore, SSHemo exhibits better tracking ability, with an angular LOA of ±27.35 °, surpassing the LOA of EV1000 of ±35.78 °. A comparative analysis reveals that Yang's modified model exhibits lower predictive accuracy, yet demonstrates a similar trend performance to SSHemo. In contrast, the EV1000 algorithm displays diminished capabilities in tracking trends, as is evident from the results.

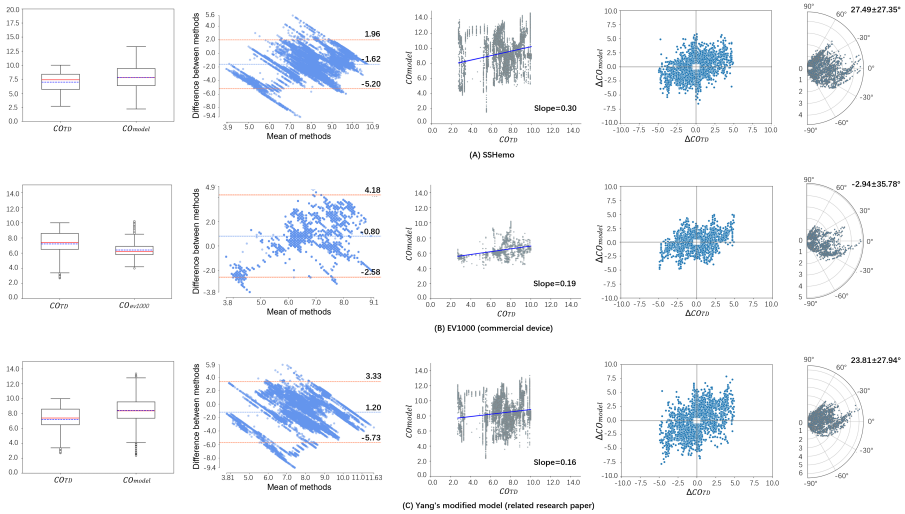

Fig. 8 The evaluated performance between SSHemo, commercial device, and related research paper [73], showing the comparison of different models or algorithms integrated with commercial devices. (A) are the five results plots of SSHemo; (B) are the five results plots of a standard commercial device: EV1000 hemodynamic monitor with the FlocTrac algorithm from Edwards Lifesciences; (C) are the five results plots from the model of reproducing and modifying current research papers [15, 16].

Due to data issues, in this investigation, we only compare SSHemo with a commercial device, EV1000 from Edwards Lifesciences, as the standard hemodynamic monitor. Because the objective quantitative comparison is based on clinically collected data and evaluation metrics, more performance results about other commercial hemodynamic monitors are investigated [73]. Table 3 concludes the results of the quantitative evaluation and the conclusions of the investigation. All results are compared to CO_{TD}, the CO of the intermittent transpulmonary thermodilution. The Bland-Altman analysis shows that the commercial devices exhibit a smaller bias. However, a larger radial LOA from the trend analysis for commercial devices did not meet the criteria for acceptable performance. Notably, SSHemo meets the clinical acceptance based on the radial LOA evaluation results.

In particular, reducing the number of samples in the newly generated data set produces different evaluation results compared to the test data sets. This discrepancy

arises primarily from the removal of samples lacking EV1000 recordings. As such, the model evaluation is contingent on the test data set, and the current results are specific to the existing data set. Therefore, compared with commercial devices mentioned in the other literature, the data is necessary for experimentation, and the current results have a limited reference value.

Model or device	Bias (L/min)	LOA lower (L/min)	LOA upper (L/min)	Radial LOA (°)
SSHemo[1]	1.62	-5.20	1.96	±27.35
EV1000[1]	-0.80	-2.58	4.18	±35.78
FloTrac/Vigileo[2]	-0.28	-2.39	2.39	−−
LiDCORapid[2]	-0.26	-2.81	2.81	−−
PiCCO2[2]	-0.86	-1.41	1.41	−−
Nexfin[2]	-0.93	-2.25	2.25	−−

[1]The evaluated results are based on experiments using publicly available datasets.

[2]The evaluated results are derived from investigations of published research papers.

Table 3 Bland-Altman analysis of SSHemo and other commercial devices

4.2.4 The analysis of predicting performance in cases of varying SVR

As Fig. 9 shows, the experimental results indicate that the predictive performance of the model remains relatively stable across various SVR ranges, with a notable exception where it performs better at lower SVR levels.

Furthermore, compared to the results obtained from the EV1000 device, it is observed that the EV1000 predictive results tend to be lower than the gold standard CO_{TD}, whereas the SSHemo model's predictions are more closely aligned with the gold standard. This suggests that the SSHemo model demonstrates superior predictive accuracy, particularly compared to existing devices like the EV1000, and its performance is less affected by variations in SVR.

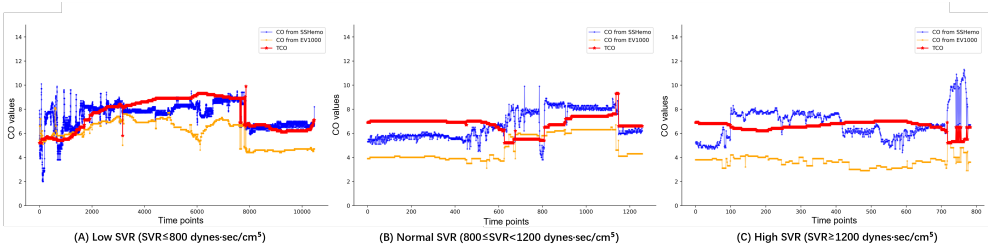

Fig. 9 A comparative analysis of CO values among CO_{TD} (red), CO_{SSHemo} (blue) and CO_{EV1000} (orange) with differences across various SVR ranges.

Sub-group	Evaluated metrics	Male (N: 9459)	Female (N: 12139)	All (N: 21598)
Sex	MAE	1.22	1.28	1.25
	Radial LOA	25.63	25.49	25.56

Table 4 Sub-group analysis for Sex.

Sub-group	Evaluated metrics	Lower range	Normal range	Higher range	All
		<18.5 (N: 5648)	18.5∼24.0 (N: 4779)	>24.0 (N: 11171)	(N: 21598)
BMI	MAE	1.66	1.22	1.05	1.25
	Radial LOA	26.79	24.32	27.78	25.56
		<18 (N: 0)	18∼60 (N: 13661)	>60 (N: 7937)	(N: 21598)
Age	MAE	N/A	1.36	1.05	1.25
	Radial LOA	N/A	26.68	24.77	25.56
		<4.0 (N: 2141)	4.0∼8.0 (N: 11066)	>8.0 (N: 8391)	(N: 21598)
CO	MAE	2.7	1.13	1.05	1.25
	Radial LOA	35.43	25.66	24.43	25.56
		<60 (N: 0)	60-100 (N: 15804)	>100 (N: 5794)	All (N: 21598)
HR	MAE	N/A	1.29	1.14	1.25
	Radial LOA	N/A	26.7	25.21	25.56
		<70 (N: 15776)	70-105 (N: 5750)	>105 (N: 72)	All (N: 21598)
MBP	MAE	1.21	1.34	1.43	1.25
	Radial LOA	24.75	28.46	26.67	25.56
		<100 (N: 13657)	100-140 (N: 7773)	>140 (N: 168)	All (N: 21598)
SBP	MAE	1.28	1.2	1.43	1.25
	Radial LOA	25.59	27.95	22.87	25.56
		<60 (N: 17681)	60-90 (N: 3881)	>90 (N: 36)	All (N: 21598)
DBP	MAE	1.24	1.31	1.39	1.25
	Radial LOA	25.11	28.76	33.62	25.56

Table 5 Sub-group analysis for different cohorts.

4.2.5 Sub-group and patient-Specific analysis

To further enhance the study, additional experiments and discussions are analyzed, including subgroup analysis based on key demographic and physiological factors. Specifically, dividing the data by age, sex, BMI, HR, SBP, DBP, MBP, and CO_{TD} will provide valuable insight into how these variables influence the results. The results are shown in Table 4 and Table 5.

The SSHemo model exhibits better performance in terms of concordance and tracking ability when applied to populations with specific characteristics, including a higher body mass index (BMI $>$ 18.5), advanced age (age $>$ 60), lower CO (CO $<$ 4 L/min), increased heart rate (HR $>$ 100 bpm), and relative lower blood pressure (MBP $<$ 70 mmHg, SBP $<$ 140 mmHg or DBP $<$ 60 mmHg).

In particular, the concordance evaluation of the SSHemo model showed more differences in these subgroups, indicating its improved ability to accurately predict hemodynamic parameters in these specific patient populations. This suggests that the SSHemo model may be particularly useful in clinical settings where patients exhibit these characteristics, providing valuable information for personalized patient care and management.

Furthermore, a comparative analysis of two distinct patients highlights the efficacy of SSHemo. Notably, the results obtained from SSHemo demonstrated a closer alignment with the established gold standard, underscoring its reliability and accuracy in producing high-quality outcomes, as Fig. 10 shows.

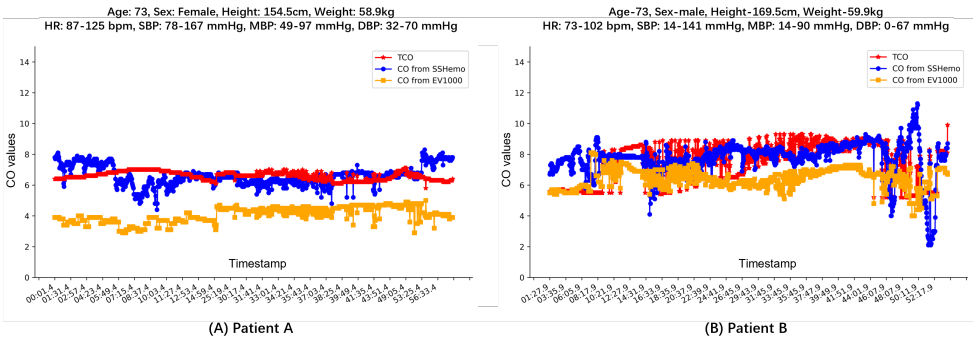

Fig. 10 The CO value comparison among the CO_{TD} (red), CO_{SSHemo} (blue) and CO_{EV1000} (orange) for two patients in a single surgery.

4.2.6 Rapid change for hemodynamic status

Our research work also focuses on the detailed analysis of rapid changes in CO during a single surgery or over a specific period. This in-depth examination aims to assess the predictive performance of our model, particularly its tracking ability, in response to rapid fluctuations in CO. By dissecting the CO changes within a specific interval, we seek to evaluate the model's capacity to accurately capture and adapt to these dynamic changes. Some example results are shown in Fig. 11 and Fig. 12.

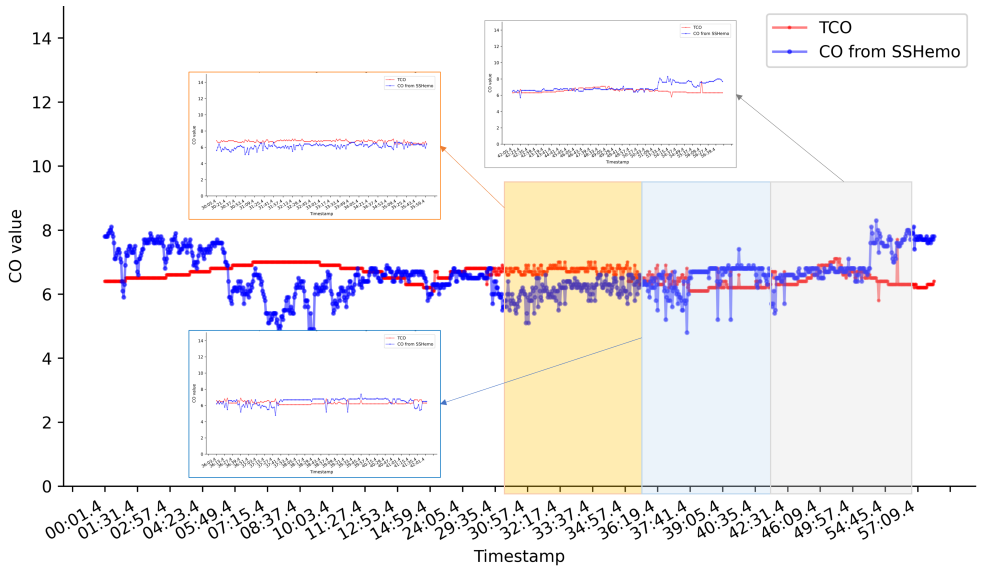

Fig. 11 The first example of rapid CO.

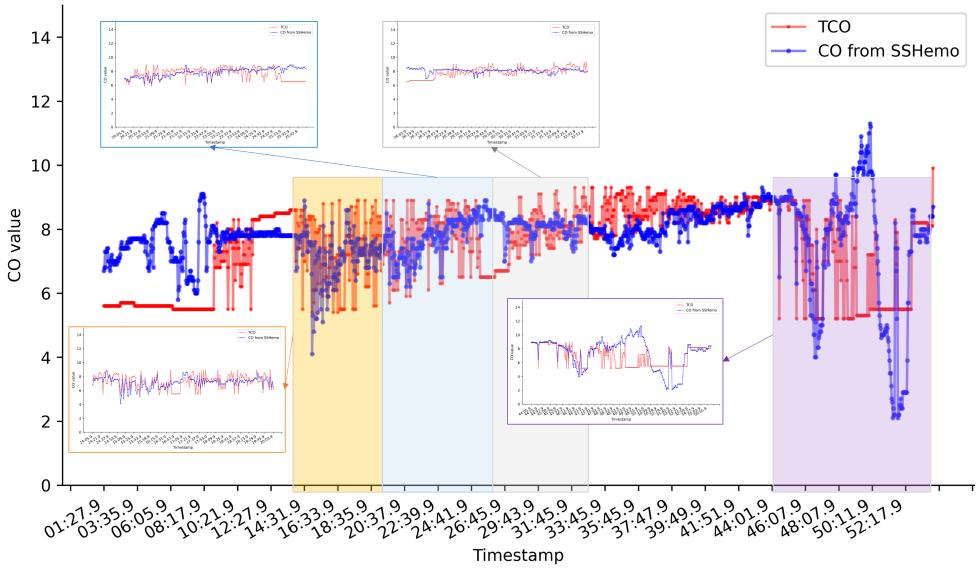

Fig. 12 The second example of rapid CO.

The experimental results show that when rapid changes in CO occur, the model often exhibits opposite or more subdued predictions compared to the gold standard, which requires a certain delay to catch up with changes. However, the overall trend remains consistent with the gold standard.

4.2.7 Other parameters prediction results

The prediction of hemodynamic parameters is not only the estimation of CO values, but also other related parameters. Table 1 shows the related parameters derived from CO and other primary hemodynamic variables (HR, SBP, MBP, DBP, CVP and others). Therefore, in this research, the model also estimates HR, SBP, MBP and DBP, and other hemodynamic parameters are calculated from mathematical equations in Table 1. It can be seen that HR, SBP, MBP, and DBP are parameters directly estimated from the ABPW signals. The supervised learning with the InceptionTime network obtained acceptable results, shown in Fig. 13. From the results, the consistency error of the parameters related to BP is approximately 2 mmHg and the HR prediction error is less than 3 bpm, which are acceptable results.

4.2.8 Computing stabilities in training process

The direct supervised learning process is often hampered by limited labeled data, resulting in instability and fluctuations. Utilizing a pre-trained backbone model can significantly enhance stability and robustness during fine-tuning, yielding more stable validation results. The pre-trained model provides a strong foundation, allowing the model to leverage knowledge and features learned from a large-scale dataset,

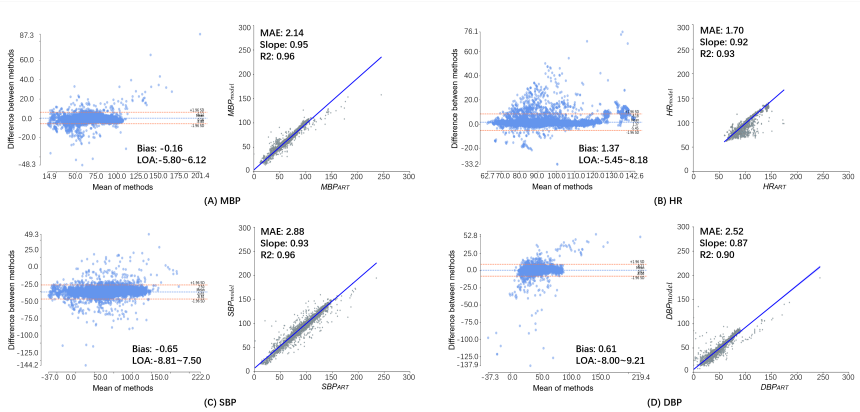

Fig. 13 The performance of predicting other hemodynamic parameters: HR, MBP, SBP, DBP. i) Two plots with parameters represent the consistency of the estimated results: Bland-Altman plot with bias (mmHg) and LOA (mmHg) and scatter plot with slope and r value (−1 1). Besides, MAE as the regression evaluated parameters are also noticed. ii) Subfigure (A)(B)(C)(D) stands for the MBP, HR, SBP, and DBP evaluation results.

thus improving generalization and reducing overfitting risks. By adopting a pre-trained backbone model, the fine-tuning process becomes more stable and effective, particularly when dealing with limited labeled data, as shown in Fig. 14.

The efficiency of the training procedure also shows significant variability. Notably, SSL leverages the pre-trained model's extracted waveform features, enabling a rapid fine-tuning process that adapts and specializes to achieve faster and more stable convergence in subsequent downstream tasks, such as hemodynamic regression, with enhanced efficiency and an accelerated convergence rate. Direct supervised learning often faces challenges during the initial training phase, exhibiting unstable convergence and requiring extensive training to understand the intricate relationship between waveform data and hemodynamic status. This leads to a larger amount of time and resources being devoted to achieving convergence, since multiple iterations are necessary to refine the model's understanding of these complex relationships.

4.2.9 Execution time and real-time efficiency

To expand the real scenario practice, the models are evaluated on four computing platforms, including CPU, NVIDIA Jetson Nano 8G edge computing platforms, NVIDIA Jetson AGX Xavier edge computing platforms, and all-in-one PCs with NVIDIA RTX 3000 graphics card. The running time for an estimation of one sample (10 seconds ABP waveform to predict the CO value) is shown in Table 6.

Based on the execution time of different models with various computing platforms, the efficient performance of real-time prediction capabilities is demonstrated, as shown in Table 6. The model facilitates real-time prediction of hemodynamic parameters, operating in a subsecond time frame. This enables beat-by-beat or second-by-second analysis, making it highly suitable for practical application scenarios where timely and accurate data is crucial.

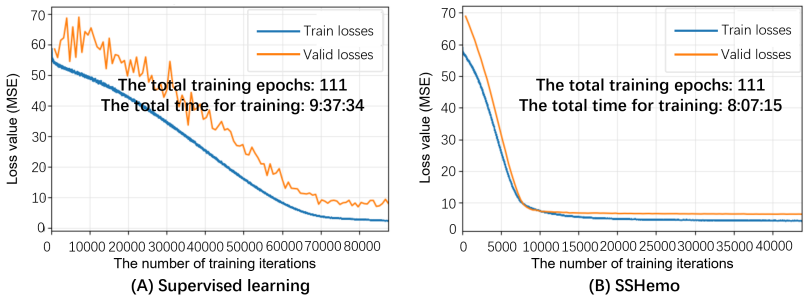

Fig. 14 One case of training process comparison between SSHemo and supervised learning, showing the promotion of computing efficiency and stability from supervised learning to SSL. i) the backbone network is both based on InceptionTime and the fine-tuning methods are FC layers. ii) the training processes are based on 111 epochs to ensure complete convergence in the training process, 0.0001 learning rate, and 1024 batch size. iii) (A) The training and evaluation loss curves for supervised learning show fluctuations during the initial stages. As the training progresses, the loss starts to decline in a fluctuating trend, indicating that the model is gradually converging towards a more stable state. iv) (B) shows the training and evaluation loss curves of SSHemo, an SSL framework. The training process was remarkably stable, with a steady decrease in loss in the evaluated data set, resulting in rapid convergence. In general, the model training was very efficient. Compared to direct supervised learning, the model converged significantly faster, with an efficiency improvement of approximately threefold.

Computing platforms	Baseline model	
	InceptionTime	patchTST
CPU (12th Gen Intel(R) Core(TM) i7-1260P 2.10 GHz)	87.01 ms	62.65 ms
Edge-computing platform (NVIDIA Jetson AGX Xavier)	16.40 ms	12.50 ms
Edge-computing platform (NVIDIA Jetson Orin Nano 8G)	10.46 ms	10.30 ms
All-in-One PCs (with graphics card of NVIDIA RTX 3000)	10.39 ms	7.85 ms

Table 6 The running time comparison among various baseline models and different computing platforms.

5 Discussion

5.1 Other methods for limited labeled data

In order to optimize the learning process with data or computational constraints and using the information present in unlabeled data, several methods have been proposed, such as semi-supervised learning [20] and the teacher-student method [74]. Semi-supervised learning is a machine learning procedure in which a model is trained on a combination of labeled and unlabeled data, helping to improve the model's generalization capabilities and performance. The teacher-student method is also a technique in machine learning in which a smaller and more computationally efficient model (the student model) learns from a larger and more complex model (the teacher model). By transferring knowledge from a larger teacher model to a smaller student model, the teacher-student method improves efficiency and performance on a variety of specific tasks.

Semi-supervised learning uses the structure or distribution of the unlabeled data to improve learning. Self-supervised learning [23] depends on learning rich data representations from large amounts of unlabeled data by creating a pretext task with automatically generated labeled training data, such as predicting part of the data given another part. The teacher-student approach aims to transfer the knowledge and replicate the teacher's performance more efficiently. It is used for model compression and to improve computational efficiency, allowing the deployment of high-performing models on devices with limited computational resources.

In the hemodynamic status prediction research domain, the key issue is fully understanding the representation of the ABPW and the relationship between waveform data and hemodynamic parameters. Therefore, SSL is much more suitable for the research field.

5.2 Contributions for limited labeled data in medical domain

This research first utilizes the SSL framework for the ABPW-hemodynamic prediction domain and improves the model performance in terms of both accuracy and tracking capabilities. The SSL method solves the problem of limited labeled data in the hemodynamic domain. Meanwhile, scarce and high-cost labeled data is a general issue in the medical domain, and there is a large amount of unlabeled data due to the wide use of EHR and monitors. SSL provides the ability for better representation based on signal or EHR data and is finally used in many other clinical domains. The basic method from SSHemo could be used in many other medical scenarios: clinical prediction based on monitoring signals, medical image processing, and patient status estimation based on vital signs time series, *etc.*

5.3 Limitations

In this study, the training data is from surgical cases within the VitalDB database. Incorporating data from broader sources provides data diversity in this research field. This expansion not only includes a wider range of cohorts (race, age, etc.), but also clinical scenarios (ICU or emergency cases), different types of surgery, various patient acuity statuses, multiple centers, *etc.* Through the improvement of data diversity, the generalizability and applicability of the model would improve.

Secondly, in the research work, information on patient demographics is not considered, due to the fact that the number of patients for training is too small (only 52 patients). Adding demographic information would increase the instability or overfitting issues for the model. Another kind of clinical issue is also not considered: medications or fluid treatments. The influence of clinical events has a significant change in hemodynamic status, especially the use of vasodilators [75]. From the above description, some clinical information, such as demographics of the patients and medical events, must be considered in the research to meet practical situations.

In addition, the CO_{TD} in the VitalDB dataset is from the Vigilance II hemodynamic monitor and the CO_{TD} in the MIMIC series dataset is from intermittent transpulmonary thermodilution. Therefore, CO_{TD} from the VitalDB data set continues, and CO_{TD} from the MIMIC series data set is discrete. In this research, we

consider both as the golden standard data. Meanwhile, from the sensor and data process aspect, a deeper understanding of sensing principles, time shifts for the results, and calibrations for continuous measurement are needed for more accurate and acceptable results.

5.4 Future work

From the evaluation results and the limitation analysis, some future works are considered.

1. The current model is divided into two parts: feature extraction and hemodynamics regressor. Currently, the feature extraction module is trained based on waveform forecast. In the future, more changes could be added to the pre-text task to get better abilities for waveform representations. Masking the waveform randomly to fill the blank masked patch is one kind of method to improve the pre-text task. Moreover, the Large Time Series Model [76], a large-scale collection of open time series datasets to empower pre-training for universal forecasting and feature extracting, is capable of addressing diverse downstream forecasting tasks, which is also another choice for the feature extractor.
2. Secondly, increasing the amount of data, not only the labeled data but also the unlabeled data, has a great impact on the model performance. In addition, adding more diversity of datasets, such as multiple research centers, various clinical scenarios, and different cohorts, would increase the generalization and practical performance of the model. Meanwhile, comparisons with commercial devices or other methods require evaluation under unified datasets or experimental conditions to ensure fairness in testing.
3. The hemodynamic status does not only depend on the ABPW, but also on other factors, such as demographic information of the patient, medical events, *etc.* In future research work, more clinical factors are considered in model research and data preprocessing procedure to improve the accuracy of the prediction model.

6 Conclusion

In this study, SSHemo, a self-supervised learning framework for hemodynamic parameter estimation, is proposed to address the limited problem of labeled data. SSHemo utilizes a large amount of unlabeled data for waveform forecasting as a pretext task and a small amount of golden standard labeled data for hemodynamic parameters regression as a downstream task. The training stability and model performance of SSHemo outperform the supervised learning model and some commercial devices, *i.e.*, EV1000. The final results show that the tracking abilities of SSHemo meet the clinical acceptance with $\pm 25.56^{\circ}$ radial LOA. Based on evaluation datasets from various clinical scenarios such as surgery and the ICU, SSHemo demonstrates strong generalizability, particularly in terms of performance tracking. Furthermore, subgroup analysis, varying SVR range analysis, and rapid CO analysis demonstrate that SSHemo exhibits relative stability across different conditions and cohorts. Computational analysis also reveals that this model has practical application value and potential. Future work

would include more diversity of data, model improvement, and other considerations of clinical parameters.

7 Declarations

- Funding
  Not applicable.
- Conflict of interest/Competing interests
  The authors declare that there are no conflicts of interest or competing interests related to this research.
- Ethics approval and consent to participate
  This research utilized publicly available datasets, and as such, ethics approval was not required in accordance with the guidelines established by the Japan Advanced Institute of Science and Technology for studies involving secondary data. The datasets used in this study include the VitalDB dataset, the MIMIC-IV dataset, the MIMIC-IV waveform dataset, the MIMIC-II dataset, and the MIMIC-II waveform dataset.
  All public datasets were anonymized, ensuring that individual participant identities cannot be ascertained. Although informed consent for participation was not required for the use of these publicly available datasets, ethical considerations were adhered to throughout the research process, including the responsible reporting and analysis of the data.
  This study followed the principles outlined in the Declaration of Helsinki and the Health Insurance Portability and Accountability Act (HIPAA) to ensure ethical research practices. The MIMIC-related databases used in this study has received ethical approval for its use, requiring researchers to complete human subjects training and adhere to data use agreements that protect patient privacy. Similarly, the VitalDB database follows ethical guidelines and collects de-identified data in compliance with local regulations on data privacy. Researchers are encouraged to reference the datasets accordingly in their work.
- Data availability
  This research utilizes multiple publicly available datasets. Details regarding each dataset are provided below:
  - VitalDB
    - Source URL: https://vitaldb.net/dataset/
    - Description: This is a comprehensive dataset of 6,388 surgical patients composed of intraoperative biosignals and clinical information. The biosignal data included in the dataset is high quality data such as 500 Hz waveform signals and numeric values at intervals of 1-7 seconds. More than 60 surgery related clinical information is also provided to help interpret the signals.
      The dataset is provided free of charge to help researchers who want to study and develop new medical AI algorithms using monitoring signals from surgical patients. We expect that the distribution of this world's

largest biosignal dataset will greatly contribute to the advancement of medical AI research.

  * Access Conditions: Users are permitted to use dataset from the provider website subject to our terms of use.
    The dataset will be released to users under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. This license type has been applied to the VitalDB open dataset in order to maximize the dissemination and use of the data.
  * Citation: [35]

- MIMIC-IV

  * Source URL: https://physionet.org/content/mimiciv/2.0/
  * Description: Retrospectively collected medical data has the opportunity to improve patient care through knowledge discovery and algorithm development. Broad reuse of medical data is desirable for the greatest public good, but data sharing must be done in a manner which protects patient privacy. The Medical Information Mart for Intensive Care (MIMIC)-III database provided critical care data for over 40,000 patients admitted to intensive care units at the Beth Israel Deaconess Medical Center (BIDMC). Importantly, MIMIC-III was deidentified, and patient identifiers were removed according to the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision. MIMIC-III has been integral in driving large amounts of research in clinical informatics, epidemiology, and machine learning. Here we present MIMIC-IV, an update to MIMIC-III, which incorporates contemporary data and improves on numerous aspects of MIMIC-III. MIMIC-IV adopts a modular approach to data organization, highlighting data provenance and facilitating both individual and combined use of disparate data sources. MIMIC-IV is intended to carry on the success of MIMIC-III and support a broad set of applications within healthcare.
  * Access Conditions: Only credentialed users who sign the DUA can access the files.
  * Citation: [37] and [38]

- MIMIC-IV waveform

  * Source URL: https://physionet.org/content/mimic4wdb/0.1.0/
  * Description:The MIMIC-IV Waveform Database is a large collection of physiological signals and measurements from patients in intensive care units, including electrocardiograms, photoplethysmograms, respiration, invasive and non-invasive blood pressure, and more. These measurements and signals are obtained directly from the bedside monitor, and provide a detailed view into the physiology of critically ill patients.
    Combining this database with the clinical information found in MIMIC-IV provides a broad, cross-sectional example of the data available to caregivers in a modern ICU. We hope that this database will provide a foundation

for future improvements to monitoring technology as well as data-driven diagnosis and treatment.
    - Access Conditions: Anyone can access the files, as long as they conform to the terms of the specified license (Open Data Commons Open Database License v1.0).
    - Citation: [39] and [38]
  - Data for CO Estimation Studies base on MIMIC-II
    - Source URL: https://archive.physionet.org/physiotools/cardiac-output/data/
    - Description: This dataset included 120 pairs of records that included both ABP waveforms and CO_{TD} measurements, and all of these were used in the study. The ABP signals, sampled at 125 Hz, can be found in the waveform records, and the reference CO_{TD} measurements, occurring at irregular intervals, can be found in the companion "numerics" records. Demographic and clinical information, including age, gender, ICD9 codes, and medications, was available in the MIMIC II Clinical Database for 117 of the study records.
    - Access Conditions: The databases are distributed freely via PhysioNet. There are no restrictions on access to the MIMIC II Waveform Database. Access to the MIMIC II Clinical Database is available to qualified researchers who obtain human subjects training and sign a simple data use agreement.
    - Citation: [43] and [42]

- Materials availability
  Not applicable.
- Code availability
  Not applicable.
- Author contribution
  All authors contributed significantly to the completion of this research, detailed as follows:
  - Author 1 (Ke Liao): Contributed to conceptualization, methodology design, literature review, data collection, statistical analysis, interpretation of findings, and is the lead author for manuscript preparation and editing.
  - Author 2 (Armagan Elibol): Contributed to statistical analysis and interpreted the findings. And participated in manuscript revision for important intellectual content related to results and discussion.
  - Author 3 (Ziyan Gao): Contributed to literature review, modification of experiment process, result analysis and interpretation of the data.
  - Author 4 (Lingzhong Meng): Contributed to the medical explainations for the comceptualization and methodology design. And participated in revising the manuscript, especially the clinical descriptions.

– Author 5 (Nak Young Chong): Supervision of the research paper. Contributed to writing the result analysis, interpreted parts, conclusion section and revising the whole manuscript.

All authors read and approved the final version of the manuscript, and all agreed to the submission of the work for publication.

References

[1] Meng, L.: Heterogeneous impact of hypotension on organ perfusion and outcomes: a narrative review. British journal of anaesthesia **127**(6), 845–861 (2021)

[2] Cecconi, M., De Backer, D., Antonelli, M., Beale, R., Bakker, J., Hofer, C., Jaeschke, R., Mebazaa, A., Pinsky, M., Teboul, J., Vincent, J., Rhodes, A.: Consensus on circulatory shock and hemodynamic monitoring. task force of the european society of intensive care medicine. Intensive Care Medicine **40**, 1795–1815 (2014) https://doi.org/10.1007/s00134-014-3525-z

[3] Esper, S.A., Pinsky, M.R.: Arterial waveform analysis. Best Practice & Research Clinical Anaesthesiology **28**(4), 363–380 (2014)

[4] Kolb, B., Kapoor, V.: Cardiac output measurement. Anaesthesia & Intensive Care Medicine **20**(3), 193–201 (2019)

[5] Frank, O.: The basic shape of the arterial pulse. first treatise: mathematical analysis. Journal of molecular and cellular cardiology **22**(3), 255–277 (1990)

[6] Rhodes, A., Sunderland, R.: Arterial pulse power analysis: The LiDCOplus system, London, pp. 183–192 (2005). https://doi.org/10.1007/3-540-26900-2_14

[7] Mayer, J., Boldt, J., Schöllhorn, T., Röhm, K., Mengistu, A., Suttner, S.: Semi-invasive monitoring of cardiac output by a new device using arterial pressure waveform analysis: a comparison with intermittent pulmonary artery thermodilution in patients undergoing cardiac surgery. British journal of anaesthesia **98**(2), 176–182 (2007)

[8] Jin, W.: Cardiovascular function assessment using computational blood flow modelling and machine learning. PhD thesis, King's College London (2021)

[9] Ke, L., Elibol, A., Wei, X., Cenyu, L., Wei, W., Chong, N.Y.: Machine learning algorithm to predict cardiac output using arterial pressure waveform analysis. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1586–1591 (2022). IEEE

[10] Yang, H.-L., Lee, H.-C., Jung, C.-W., Kim, M.-S.: A deep learning method for intraoperative age-agnostic and disease-specific cardiac output monitoring from arterial blood pressure. In: 2020 IEEE 20th International Conference on

Bioinformatics and Bioengineering (BIBE), pp. 662–666 (2020). IEEE

[11] Moon, Y.-J., Moon, H.S., Kim, D.-S., Kim, J.-M., Lee, J.-K., Shim, W.-H., Kim, S.-H., Hwang, G.-S., Choi, J.-S.: Deep learning-based stroke volume estimation outperforms conventional arterial contour method in patients with hemodynamic instability. Journal of clinical medicine **8**(9), 1419 (2019)

[12] Kwon, H.-M., Seo, W.-Y., Kim, J.-M., Shim, W.-H., Kim, S.-H., Hwang, G.-S.: Estimation of stroke volume variance from arterial blood pressure: Using a 1-d convolutional neural network. Sensors **21**(15), 5130 (2021)

[13] Levick, J.R.: An Introduction to Cardiovascular Physiology. Butterworth-Heinemann, London (2013)

[14] Lopes, A.A., O'Leary, P.W.: Measurement, interpretation and use of hemodynamic parameters. Cardiology in the Young **19**(S1), 8–12 (2009)

[15] Yang, H.-L., Jung, C.-W., Yang, S.M., Kim, M.-S., Shim, S., Lee, K.H., Lee, H.-C.: Development and validation of an arterial pressure-based cardiac output algorithm using a convolutional neural network: retrospective study based on prospective registry data. JMIR Medical Informatics **9**(8), 24762 (2021)

[16] Yang, H.-L., Lee, H.-C., Jung, C.-W., Kim, M.-S.: A deep learning method for intraoperative age-agnostic and disease-specific cardiac output monitoring from arterial blood pressure. In: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 662–666 (2020). IEEE

[17] Ho, N., Kim, Y.-C.: Evaluation of transfer learning in deep convolutional neural network models for cardiac short axis slice classification. Scientific reports **11**(1), 1839 (2021)

[18] Yu, Y., Li, M., Liu, L., Li, Y., Wang, J.: Clinical big data and deep learning: Applications, challenges, and future outlooks. Big Data Mining and Analytics **2**(4), 288–305 (2019)

[19] Liao, K., Wang, W., Elibol, A., Meng, L., Zhao, X., Chong, N.Y.: Does deep learning really outperform non-deep machine learning for clinical prediction on physiological time series? arXiv preprint arXiv:2211.06034 (2022)

[20] Pise, N.N., Kulkarni, P.: A survey of semi-supervised learning methods. In: 2008 International Conference on Computational Intelligence and Security, vol. 2, pp. 30–34 (2008). https://doi.org/10.1109/CIS.2008.204

[21] Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering **22**(10), 1345–1359 (2010) https://doi.org/10.1109/TKDE.2009.191

[22] Siméoni, O., Budnik, M., Avrithis, Y., Gravier, G.: Rethinking deep active learning: Using unlabeled data at model training. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 1220–1227 (2021). https://doi.org/10.1109/ICPR48806.2021.9412716

[23] Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., Tao, D.: A Survey on Self-supervised Learning: Algorithms, Applications, and Future Trends (2023)

[24] Song, Y., Wang, T., Cai, P., Mondal, S.K., Sahoo, J.P.: A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. ACM Comput. Surv. **55**(13s) (2023) https://doi.org/10.1145/3582688

[25] Mumuni, A.G., Mumuni, F.: Data augmentation: A comprehensive survey of modern approaches. Array **16**, 100258 (2022)

[26] Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2009)

[27] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019). https://arxiv.org/abs/1810.04805

[28] Nirmalan, M., Dark, P.M.: Broader applications of arterial pressure wave form analysis. Continuing Education in Anaesthesia, Critical Care & Pain **14**(6), 285–290 (2014)

[29] Kouz, K., Scheeren, T.W., Backer, D., Saugel, B.: Pulse wave analysis to estimate cardiac output. Anesthesiology **134**(1), 119–126 (2021)

[30] Ribeiro, M., Grolinger, K., ElYamany, H.F., Higashino, W.A., Capretz, M.A.: Transfer learning with seasonal and trend adjustment for cross-building energy forecasting. Energy and Buildings **165**, 352–363 (2018)

[31] Critchley, L.A.H.: Minimally invasive cardiac output monitoring in the year 2012. Artery Bypass (2013)

[32] Campos, M.M., Sellés, A.F., Vera, G.G., Febrer, J.M., Cloarec, C.S., Hernández, Y.P., Nogales, X.G.: Techniques available for hemodynamic monitoring. advantages and limitations. Medicina Intensiva (English Edition) **36**(6), 434–444 (2012)

[33] Mathews, L., Singh, R.K.: Swan-ganz catheter in hemodynamic monitoring. Journal of Anaesthesiology Clinical Pharmacology **22**(4), 335–345 (2006)

[34] Pinsky, M.R., Teboul, J.-L., Vincent, J.-L., Intensive Care Medicine, E.S., *et al.*: Hemodynamic Monitoring. Springer, USA (2019)

[35] Lee, H.-C., Park, Y., Yoon, S.B., Yang, S.M., Park, D., Jung, C.-W.: Vitaldb, a

high-fidelity multi-parameter vital signs database in surgical patients. Scientific Data **9**(1), 279 (2022)

[36] Lee, H.-C., Jung, C.-W.: Vital recorder—a free research tool for automatic recording of high-resolution time-synchronised physiological data from multiple anaesthesia devices. Scientific reports **8**(1), 1527 (2018)

[37] Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B.: Mimic-iv, a freely accessible electronic health record dataset. Scientific data **10**(1), 1 (2023)

[38] Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. circulation **101**(23), 215–220 (2000)

[39] Moody, B., Hao, S., Gow, B., Pollard, T., Zong, W., Mark, R.: MIMIC-IV Waveform Database (version 0.1.0). PhysioNet (2022). https://doi.org/10.13026/a2mw-f949 . https://doi.org/10.13026/a2mw-f949

[40] Saeed, M., Villarroel, M., Reisner, A.T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T.H., Moody, B., Mark, R.G.: Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. Critical care medicine **39**(5), 952 (2011)

[41] Saeed, M., Lieu, C., Raber, G., Mark, R.G.: Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring. In: Computers in Cardiology, pp. 641–644 (2002). IEEE

[42] Sun, J., Reisner, A., Saeed, M., Mark, R.: Estimating cardiac output from arterial blood pressurewaveforms: a critical evaluation using the mimic ii database. In: Computers in Cardiology, 2005, pp. 295–298 (2005). IEEE

[43] Sun, J.X.: Cardiac output estimation using arterial blood pressure waveforms. PhD thesis, Massachusetts Institute of Technology (2006)

[44] Poirot, M.G., Vepakomma, P., Chang, K., Kalpathy-Cramer, J., Gupta, R., Raskar, R.: Split learning for collaborative deep learning in healthcare. arXiv preprint arXiv:1912.12115 (2019)

[45] Wu, N.-H., Hsieh, T.-H., Chang, C.-Y., Shih, P.-C., Kao, M.-C., Lin, H.-Y.: Validation of cardiac output estimation using the fourth-generation flotrac/ev1000™ system in patients undergoing robotic-assisted off-pump coronary artery bypass surgery. Heart and Vessels **38**(3), 341–347 (2023)

[46] Westerhof, N., Lankhaar, J.-W., Westerhof, B.E.: The arterial windkessel. Medical & biological engineering & computing **47**(2), 131–141 (2009)

[47] Chen, J., Sun, K., Sun, Y., Li, X.: Signal quality assessment of ppg signals using stft time-frequency spectra and deep learning approaches. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1153–1156 (2021). IEEE

[48] Sherstinsky, A.: Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D: Nonlinear Phenomena **404**, 132306 (2020)

[49] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)

[50] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[51] Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D.F., Weber, J., Webb, G.I., Idoumghar, L., Muller, P.-A., Petitjean, F.: Inceptiontime: Finding alexnet for time series classification. Data Mining and Knowledge Discovery **34**(6), 1936–1962 (2020)

[52] Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. ArXiv **abs/2211.14730** (2022)

[53] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W.: Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 11106–11115 (2021)

[54] Breiman, L.: Random forests, machine learning 45. Journal of Clinical Microbiology **2**, 199–228 (2001)

[55] Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A.F.M., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M.S., Hand, D.J., Steinberg, D.: Top 10 algorithms in data mining. Knowledge and Information Systems **14**, 1–37 (2007)

[56] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. ACM (2016)

[57] Dai, W., Yang, Q., Xue, G.-R., Yu, Y.: Boosting for transfer learning. In: Proceedings of the 24th International Conference on Machine Learning, pp. 193–200 (2007)

[58] Brockwell, P.J., Davis, R.A.: Introduction to Time Series and Forecasting. Springer, USA (2002)

[59] Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2016)

[60] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[61] Caruana, R., Lawrence, S., Giles, C.: Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. Advances in neural information processing systems **13** (2000)

[62] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

[63] Howard, J., Gugger, S.: Fastai: A layered api for deep learning. Information **11**(2), 108 (2020)

[64] Oguiza, I.: tsai - A state-of-the-art deep learning library for time series and sequential data. Github (2023). https://github.com/timeseriesAI/tsai

[65] Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction vol. 2. Springer, USA (2009)

[66] Critchley, L.A., Yang, X.X., Lee, A.: Assessment of trending ability of cardiac output monitors by polar plot methodology. Journal of cardiothoracic and vascular anesthesia **25**(3), 536–546 (2011)

[67] Critchley, L.: Meta-analyses of Bland–Altman-style cardiac output validation studies: good, but do they provide answers to all our questions? Oxford University Press (2017)

[68] Tufte, E.R.: The Visual Display of Quantitative Information vol. 2. Graphics press Cheshire, CT, USA (2001)

[69] Giavarina, D.: Understanding bland altman analysis. Biochemia medica **25**(2), 141–151 (2015)

[70] Saugel, B., Grothe, O., Wagner, J.Y.: Tracking changes in cardiac output: statistical considerations on the 4-quadrant plot and the polar plot methodology. Anesthesia & Analgesia **121**(2), 514–524 (2015)

[71] Montenij, L., Buhre, W., Jansen, J., Kruitwagen, C., De Waal, E.: Methodology of method comparison studies evaluating the validity of cardiac output monitors: a stepwise approach and checklist. British Journal of Anaesthesia **116**(6), 750–758 (2016)

[72] Desebbe, O., Henaine, R., Keller, G., Koffel, C., Garcia, H., Rosamel, P., Obadia, J.-F., Bastien, O., Lehot, J.-J., Haftek, M.: Ability of the third-generation flotrac/vigileo software to track changes in cardiac output in cardiac surgery patients: a polar plot approach. Journal of cardiothoracic and vascular anesthesia **27**(6), 1122–1127 (2013)

[73] Geisen, M., Ganter, M.T., Hartnack, S., Dzemali, O., Hofer, C.K., Zollinger, A.: Accuracy, precision, and trending of 4 pulse wave analysis techniques in the postoperative period. Journal of Cardiothoracic and Vascular Anesthesia **32**(2), 715–722 (2018) https://doi.org/10.1053/j.jvca.2017.09.006

[74] Hu, C., Li, X., Liu, D., Wu, H., Chen, X., Wang, J., Liu, X.: Teacher-student architecture for knowledge distillation: A survey. ArXiv **abs/2308.04268** (2023)

[75] Ream, A.K.: Cardiovascular hemodynamics. JAMA The Journal of the American Medical Association **204**(8), 742 (1968)

[76] Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., Sahoo, D.: Unified training of universal time series forecasting transformers. ArXiv **abs/2402.02592** (2024)