Test-Time Calibration: A Framework for Personalized Test-Time Adaptation in Real-World Biosignals

Yong-Yeon Jo [‡] YY.JO@MEDICALAI.COM
Byeong Tak Lee[‡] BYTAKLEE@MEDICALAI.COM
Hak Seung Lee[‡] CARDIOLEE@MEDICALAI.COM
Joon-myoung Kwon[‡] CTO@MEDICALAI.COM
Jeong-Ho Hong[§] NEUROHONG79@GMAIL.COM
Beom Joon Kim[¶] KIM.BJ.STROKE@GMAIL.COM *
[‡] *MedicalAI Co. Ltd*
[§] *Keimyung University Dongsan Medical Center*
[¶] *Seoul National University Bundang Hospital*

Abstract

Test-Time Adaptation (TTA) methods have been widely used to enhance model robustness by continuously updating pre-trained models with unlabeled target data. However, in real-world biosignal applications-where factors such as age, lifestyle, and comorbidities induce significant variability—traditional TTA often falls short in addressing personalization needs. To satisfy such needs, we introduce a novel Test-Time Calibration (TTC) framework that integrates continuous self-supervised adaptation on unlabeled samples with periodic supervised calibration using the sporadically available ground-truth labels. Our approach leverages a model equipped with dual heads for supervised learning (SL) and self-supervised learning (SSL), and further incorporates a dual buffer along with a weighted batch sampling strategy to effectively manage and utilize both data types during the test phase. We evaluate our framework on two distinct datasets: the publicly available PulseDB, a benchmark for cuff-less blood pressure estimation, and a private ICU dataset collected from critically ill patients. Experimental results demonstrate that our approach improves blood pressure prediction accuracy and robustness, highlighting its suitability for dynamic, personalized biosignal applications.

* Correspondence

Data and Code Availability This paper uses the PulseDB dataset Wang et al. (2023), which is available on the PhysioNet repository. While the source code for the experiments is not publicly available, we provide detailed descriptions of all experimental procedures and the hyperparameters chosen. Our implementation is based on the framework presented in Liu et al. (2021).

1. Introduction

A test-time adaptation (TTA) aims to improve model robustness in the presence of distribution shifts on the target domain—a common challenge in real-world applications Ma et al. (2022); He et al. (2021); He (2022). These approaches continuously update pre-trained models using unlabeled samples in the target domain encountered during the test phase Chen et al. (2023); Gan et al. (2023); Sun et al. (2020).

However, such TTA approaches could struggle to cope with scenarios in which personalized samples are received continuously. For example, many real-world biosignal applications require personalization, as individual factors such as age, lifestyle, and comorbidities can affect signal patterns. To address these challenges, many practical systems involve periodic manual calibration to maintain measurement accuracy despite ongoing physiological changes Samsung Electronics (2024); Urden et al. (2013); Galindo and Aleppo (2020).

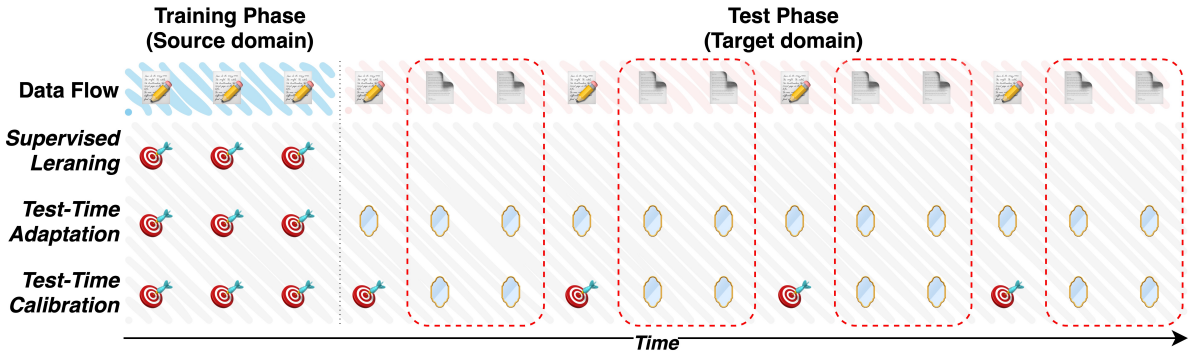


Figure 1: A scenario where personalized data is continuously received and the model is periodically recalibrated after deployment. Before deployment (i.e., during the training phase), the model is trained exclusively on source domain data composed of labeled samples (`Notes with pen`). At this stage, both Test-Time Adaptation (TTA) and Test-Time Calibration (TTC) update the model using a combination of SL and SSL (`Dart`). After deployment (i.e., during the test phase), the model processes incoming data where unlabeled samples (`Plain documents`) are used solely for SSL (`Mirror`). Additionally, TTC performs calibration by incorporating the labeled samples. The round rectangle with a red dot indicates the inference points.

Figure 1 illustrates a practical scenario. In many real-world applications, biosignals are collected continuously, resulting in a constant flow of unlabeled samples. However, the corresponding ground-truth labels, which are often obtained through periodic manual measurements or checkups, appear only intermittently. This scenario reflects the challenge of continuously updating the model with new data while relying on occasional labeled instances for recalibration.

To address this, we propose a *Test-Time Calibration (TTC)* framework that combines adaptation on unlabeled samples with periodic supervised calibration. To accommodate both types of data, our framework employs a model equipped with dual heads for supervised learning (SL) and self-supervised learning (SSL) Sun et al. (2020). Additionally, while unlabeled samples capture the most recent dynamics of an individual, they may be less informative compared to labeled samples. Conversely, because labeled samples are infrequent, they may not fully reflect the individual's current state. To leverage the complementary strengths of these data types, we introduce two key techniques: a dual buffer and a weighted batch sampling strategy. The dual buffer maintains separate buffers for recent unlabeled and labeled samples, ensuring that the less frequent labeled samples are preserved while the model continuously updates with the latest unlabeled data. Meanwhile, the weighted batch sampling strategy creates batches that mix samples from both buffers, allowing the model to benefit from the complementary strengths of each type of data.

To validate our proposed framework, we evaluate our approach on two distinct datasets that are well-suited to our scenario and focused on blood pressure prediction: the publicly available PulseDB dataset Wang et al. (2023) and a private ICU dataset exclusively collected from critically ill ICU patients. Our experimental results demonstrate that TTC achieves significant quantitative performance improvements by reducing prediction errors. In addition, our qualitative analysis shows that TTC effectively captures temporal variations in an individual's blood pressure over time, outperforming other methods. Furthermore, our ablation study confirms the effectiveness of our approach by leveraging the complementary strengths of different data types.

By addressing the previously overlooked scenario of continuously incoming personalized data, our TTC framework overcomes the limitations of conventional TTA methods and holds promising potential for efficient deployment in personalized wearable devices and other real-world applications.

Our contributions are as follows:

- We present the challenge of applying conventional TTA methods in real-world applications where personalized data are continuously received and calibration relies on sporadic labeled samples.

- We propose a novel framework that effectively handles this scenario by leveraging a model capable of processing both unlabeled and labeled samples. Furthermore, we introduce a dual-buffer and weighted batch sampling strategy to exploit the complementary strengths of these two data types, thereby enabling dynamic adaptation during the test phase.
- We extensively evaluate the proposed approach on two real-world datasets, demonstrating significant improvements in blood pressure prediction accuracy through both quantitative and qualitative analyses.

2. Related Work

In our study, we categorize test-time adaptation (TTA) methods into three main paradigms, reflecting the diverse strategies employed in recent literature to address distribution shifts in the target domain:

Test-Time Domain Adaptation (TTDA) : TTDA methods assume access to the entire target domain unlabeled data during the test phase. These methods typically use self-supervised loss functions, such as pseudo-labeling and entropy minimization, to refine predictions on the target unlabeled data collectively. TTDA methods are highly effective in scenarios where comprehensive access to target data is available, allowing models to make informed adjustments over multiple epochs. Key methods in this paradigm include SHOT (Liang et al. (2020)) and NRC (Yang et al. (2021)), which leverage feature alignment and pseudo-label refinement techniques to bridge domain gaps.

Test-Time Batch Adaptation (TTBA) : In TTBA, the model adapts its parameters based on incoming mini-batches of data, enabling instant predictions as new samples are received. TTBA methods such as PredBN (Nado et al. (2020)) and MEMO (Zhang et al. (2022)) have shown effectiveness in handling corruption-based shifts, where data is corrupted in specific ways, such as noise or blurring. These methods are particularly useful when the batch size is limited, as they can utilize techniques like entropy minimization to stabilize adaptation.

Online Test-Time Adaptation (OTTA) : OTTA focuses on sequential adaptation where data arrives continuously in batches, allowing models to adapt iteratively over time. Unlike TTDA and TTBA, OTTA is designed for *real-time updates*, making it suitable for dynamic environments where data distributions may change frequently. OTTA methods like Tent (Wang et al. (2020))), CoTTA (Wang et al. (2022)), and EATA (Niu et al. (2022)) have been shown to perform well on corruption datasets by employing techniques such as entropy minimization and sample selection modules to enhance adaptability.

Among these paradigms, our study focuses on OTTA, as it is particularly suited for scenarios like biosignal applications that demand continuous model adaptation after deployment. Therefore, in this paper, when we refer to TTA, we are specifically discussing methods based on the OTTA.

3. Test-Time Calibration

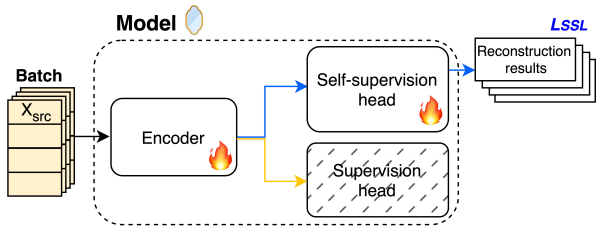
3.1. Motivation

Traditional TTA methods primarily rely on self-supervised learning with continuous streams of unlabeled test samplesWang et al. (2020); Niu et al. (2022); Wang et al. (2022), rendering them susceptible to model drift and diminished predictive performance over time. However, these approaches do not address personalization. In contrast, real-world biosignal applications require personalized solutions, as individual differences in age, lifestyle, and comorbidities can markedly alter signal patterns.

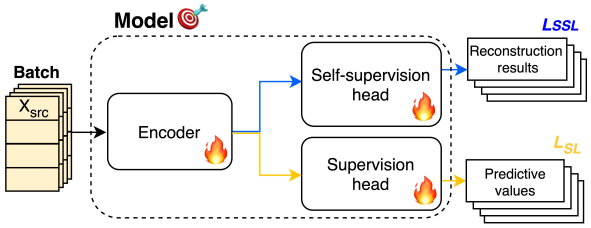
To address these challenges, many practical systems incorporate periodic manual calibration—such as cuff-based validation for blood pressure monitors or finger-prick tests for continuous glucose monitoring Samsung Electronics (2024); Urden et al. (2013); Galindo and Aleppo (2020)—which helps maintain measurement accuracy despite ongoing physiological changes.

Thus, ignoring these real-world constraints limits the clinical utility of traditional TTA methods, underscoring the need for an adaptation mechanism that integrates both unlabeled and labeled signals to achieve robust, personalized updates. To address this gap, we propose a *Test-Time Calibration (TTC)* framework, designed to unify continuous self-supervised learning with intermittent supervised calibration.

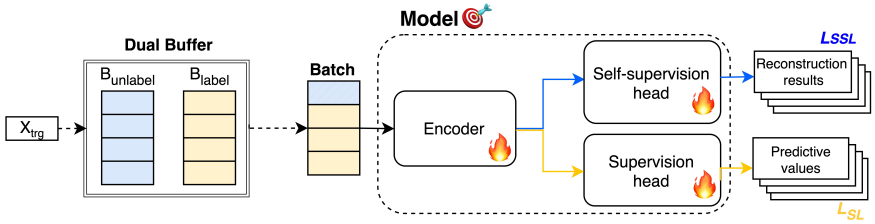
Figure 1 illustrates a real-world biosignal scenario, comparing supervised learning (SL) TTA and TTC across various sample types and learning processes. During the training phase, the source domain comprises data from multiple individuals, whereas in the



(*a*) Model Initialization: In the training phase, the model is initialized using source domain samples through self-supervised learning to warm up.



(*b*) Model Finetuning: In the training phase, the model is further refined on source domain samples by integrating both self-supervised and supervised learning.



(*c*) Model Calibration: In the test phase, the model is adapted to individual target domain samples using a dual buffer and weighted batch sampling strategy. This process leverages both self-supervised and supervised learning to update the model based on the type of incoming sample.

Figure 2: Test-time calibration procedure.

test phase, only samples from a single individual are fed into the model.

In the SL scenario, the model serves as a baseline. Here, the model is trained exclusively on source domain data, relying solely on labeled samples. Once deployed, the model performs inference on incoming data without any further adaptation, meaning that it does not update its parameters even if there are shifts or variations in the data distribution over time.

In contrast, the TTA approach builds on this by training the model with both SL and SSL during the training phase. Once the model is deployed, it continuously adapts by using SSL on each incoming unlabeled sample. This means that while the initial training leverages the explicit information from labels, the test phase updates the model in an unsupervised manner to account for distribution shifts. However, since this adaptation relies solely on SSL during deployment, the model might not fully capture the nuances of the target domain when labels are available, which could limit its ability to maintain calibration over time.

Our proposed TTC framework extends TTA by incorporating periodic supervised calibration. In this scenario, the model continues to adapt through SSL on unlabeled samples as in TTA, but whenever labeled samples become available, the model temporarily suspends inference to perform a calibration step. During these calibration points, the model updates its parameters using both SL and SSL, thereby integrating the precise, ground-truth information from the labeled samples. This dual update strategy ensures that the model remains well-aligned with the evolving biosignal dynamics. By recalibrating periodically, the TTC framework effectively compensates for potential drift that may occur when relying solely on self-supervised adaptation, ultimately preserving and even enhancing the model's prediction accuracy over time.

3.2. Framework

Figure 2 illustrates our proposed framework, which is divided into two distinct phases: the training phase and the test phase.

Training Phase: During this phase, the model is pre-trained on source domain data to achieve both generalizable feature extraction and fine-tuned pre-

Table 1: Summary of datasets

	PulseDB		ICU	
	Training	Test	Training	Test
# of subjects	2,506	279	781	88
# of segments	900,359	111,199	684,105	94,961
SBP (mean, std)	118.60 ± 21.03	118.84 ± 20.59	133.08 ± 22.54	134.82 ± 22.29
DBP (mean, std)	61.86 ± 12.65	62.00 ± 12.27	77.75 ± 14.86	78.35 ± 15.87

dictive capabilities. We adopt a dual-head neural network architecture for our framework, building on prior work in which one head is dedicated to supervised learning (SL) and the other to self-supervised learning (SSL) Gandelsman et al. (2022). Notably, our framework is model-agnostic, meaning that any architecture capable of performing both SL and SSL can be employed.

There are two steps. As the first step, the model is initialized using a reconstruction task—a form of SSL—to learn robust representations of the underlying signal patterns, as shown in Figure 2(a). This step provides a strong initialization by encouraging the model to capture domain-invariant features.

Figure 2(b) illustrates the next step in our training process. Following the conventional protocol Sun et al. (2020), the model is further fine-tuned using source domain data. However, instead of relying solely on SL, we simultaneously perform both SL and SSL during this step. Given that biosignal data can exhibit significant inter-variability even for the same individual under similar conditions, this dual approach is essential. By fusing SL and SSL, we mitigate the risk of the model becoming overly dependent on the supervised objective while still capitalizing on the valuable insights provided by self-supervision. This balanced strategy yields both generalizable feature extraction and finely-tuned predictive capabilities for the source domain samples.

Test Phase: Figure 2(c) illustrates the process during the test phase. In this phase's data flow scenario, the sample distribution is dominated by unlabeled data, while labeled samples remain sparse. Each type of sample exhibits distinct characteristics. Although the abundance of unlabeled samples is beneficial for capturing each individual's most recent physiological state, the lack of ground-truth labels complicates precise model updates. Conversely, labeled samples are more informative for calibrating the pre-trained model's weights, but they might not always reflect the latest conditions.

To fully leverage these data characteristics, we propose two key strategies: a dual buffer and a weighted batch sampling strategy. Using a conventional batch formation approach, the batch is predominantly composed of frequently arriving unlabeled samples, which leads to the frequent loss of the more informative labeled samples. To address this, we design a *dual buffer* that maintains two separate buffers: one for recent unlabeled samples (B_{unlabel}) and one for labeled samples (B_{label}). This design ensures that the model continuously benefits from fresh unlabeled data to capture emerging trends, while preserving labeled data for repeated, impactful parameter updates.

Furthermore, to utilize this buffer effectively, we introduce a *weighted batch sampling strategy*. Since labeled samples do not arrive frequently, they may contain outdated information that could impede the model's ability to update with the latest trends if overrepresented. Therefore, we mix unlabeled and labeled samples in controlled proportions when forming each batch, ensuring that the model remains responsive to new data distributions without overfitting to the limited labeled samples.

The detailed procedure for this test-time calibration is presented in Algorithm 1. In the algorithm, each incoming sample (x, y) is processed by updating the appropriate buffer based on label availability, and calibration training is performed using a weighted batch formed from both B_{unlabel} and B_{label} according to a predefined sampling ratio r.

4. Experiment

4.1. Dataset

For the evaluation of the TTC framework, we employ two distinct biosignal datasets: the publicly available PulseDB and a private dataset collected from an in-

Algorithm 1 Test-Time Calibration Procedure

Require: Dataset D, Model M, Buffer sizes B_{label}, $B_{unlabel}$, Sampling ratio r

Initialize labeled buffer B_{label} and unlabeled buffer $B_{unlabel}$
for each incoming sample (x, y) from data loader **do**
  if y is available **then**
    if B_{label} is full **then**
      Remove oldest sample from B_{label}
    end if
    Append (x, y) to B_{label}
  else
    if $B_{unlabel}$ is full **then**
      Remove oldest sample from $B_{unlabel}$
    end if
    Append x to $B_{unlabel}$
  end if
  Construct training batch by sampling from B_{label} and $B_{unlabel}$ according to ratio r
  Update calibration parameters of M using the constructed batch
end for

tensive care unit (ICU dataset). Table 1 summarizes the dataset.

PulseDB The PulseDB dataset Wang et al. (2023) serves as a benchmark for cuff-less blood pressure estimation. It integrates data from the MIMIC-III waveform Johnson et al. (2016) and VitalDB Lee et al. (2022) databases, comprising over 5.2 million 10-second segments of electrocardiogram (ECG), photoplethysmogram (PPG), and arterial blood pressure (ABP) waveforms. The dataset applies rigorous quality filtering: segments that do not contain a complete cycle during the 10-second interval are excluded, segments with negative skewness in PPG cycles are discarded, and segments with a correlation of less than 0.9 between PPG and ABP are removed. These criteria ensure that only high-quality signals are retained, though they may result in an irregular temporal distribution of samples. Consequently, on average, each subject contributes about 360 continuous segments to the training set and 400 continuous segments to the test set. ABP is used to derive ground-truth labels for systolic and diastolic blood pressure (SBP and DBP, respectively).

ICU Dataset We use a private ICU dataset recorded with Vitalrecord[1]. This dataset is exclusively collected from critically ill ICU patients. For ease of training and processing, the continuous waveforms are segmented into 1-minute intervals. Each segment undergoes the same quality checks for waveforms as those applied in PulseDB. As a result, due to variations in the duration of ICU stays, the number of segments per patient ranges widely from 60 to 6000. In contrast to PulseDB, periodic cuff-based blood pressure measurements in the ICU serve as the ground-truth labels, with SBP/DBP values being recorded.

Comparing the two datasets, not only does the distribution of segments per patient differ, but the patients in the private ICU dataset also tend to have underlying conditions such as hypertension or atrial fibrillation. These conditions result in highly fluctuating and non-stationary blood pressure patterns, which is reflected in the higher standard deviations of the actual BP values reported in the table.

4.2. Implementation Details

Training Phase: We design a model that employs a ResNet-inspired module as the embedding layer He et al. (2016) and utilizes the Vision Transformer (ViT) Dosovitskiy et al. (2020) as the backbone encoder. For both self-supervised and supervised learning, the decoder and regressor are implemented using a single Transformer block, respectively. To optimize hyperparameters, we employed Population Based Training Jaderberg et al. (2017). For the loss functions, Shrinkage Loss Lu et al. (2018) is applied for the supervised learning component, while mean squared error (MSE) is used for self-supervised learning. For optimization, we used the LAMB optimizer You et al. (2019). Detailed hyperparameter settings are provided in the Appendix A.

Although the backbone network remains consistent, the processing of the input signals varies according to the dataset. Since the PulseDB dataset consists of 10-second segments, we subdivide them into finer overlapping patches to extract more detailed information. Specifically, both ECG and PPG signals are resampled to 125 Hz, and the input signals are divided into overlapping patches of size 30 with a shift size of 15. In contrast, the ICU dataset is provided in relatively longer 1-minute segments. For this dataset, both ECG and PPG signals are also resam-

1. https://vitaldb.net/vital-recorder/

Table 2: Blood pressure prediction results on different datasets. Metrics reported: MAE with 95% confidence interval and standard deviation; bold indicates the best performance

Dataset	Model	SBP	DBP
PulseDB	SL	13.65 ± 0.10 (17.36)	7.97 ± 0.06 (10.08)
	TTA	12.96 ± 0.10 (16.66)	7.48 ± 0.06 (9.64)
	TTC	**8.84 ± 0.07 (11.84)**	**4.94 ± 0.04 (6.78)**
ICU	SL	14.62 ± 0.32 (19.0)	10.90 ± 0.24 (15.0)
	TTA	13.30 ± 0.85 (20.0)	9.51 ± 0.61 (14.0)
	TTC	**11.75 ± 0.26 (16.0)**	**8.57 ± 0.20 (12.0)**

pled to 125 Hz, but the input signals are segmented into non-overlapping patches of size 125.

Test Phase: In the test phase, a dual buffer and a sampling strategy are employed. The buffer is configured with an 8:1 ratio, such that B_{unlabel} holds 64 unlabeled samples and B_{label} holds 8 labeled samples. For the sampling strategy, we fix the total batch size at 32 and select unlabeled and labeled samples in a 3:1 ratio. Since the model continues to learn during the test phase, an optimizer is required. We employ Stochastic Gradient Descent (SGD) Robbins and Monro (1951) with a fixed learning rate of 0.001, momentum of 0.9, and weight decay of 0.001. This adaptation process is conducted individually for each subject, with model updates repeated 5 times per batch.

4.3. Experimental Result

To evaluate both the general calibration scenario and the effectiveness of our proposed method, we introduce certain assumptions into our experimental setup. For the PulseDB dataset, where ABP is used, ground-truth labels for SBP and DBP are available for every segment. Since calibrating at every segment would be equivalent to a SL setting, we simulate periodic checkups by assuming that one labeled sample appears after every ten unlabeled samples, thereby approximating a realistic calibration scenario for TTC. Consequently, performance is evaluated on the samples that are not used as calibration points.

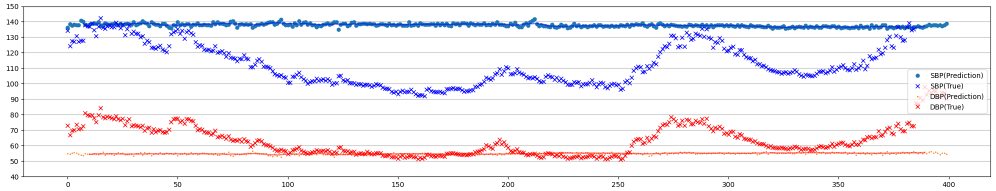
On the ICU dataset, labeled data are only available at periodic checkup points, which typically occur roughly once per hour. In an ICU setting, although biosignals are continuously recorded for several to tens of hours, the actual BP measurements are taken infrequently. Predicting BP one hour ahead based on biosignals recorded an hour earlier does not adequately reflect the effectiveness of our proposed method. Therefore, we limit our evaluation to segments near the time of BP measurement—within approximately ±5 segments—based on the assumption that blood pressure remains relatively stable within this narrow time window[2]. This approach ensures that our predictions are more accurate and clinically relevant.

For evaluation metrics, we report the mean absolute error (MAE) along with its standard deviation (std) and the 95% confidence interval (ci, expressed as ± values) for both SBP and DBP. In Table 2, the best performance is highlighted in bold.

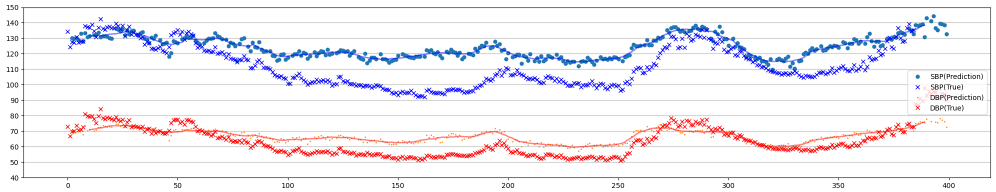
4.3.1. Performance Evaluation

Quantitative: In Table 2, we summarize the blood pressure prediction performance across the two datasets. Note that all values represent MAE, where lower values indicate better performance. The results demonstrate that the TTC framework achieves notably lower prediction errors for SBP compared to SL and TTA. Specifically, the MAE for SBP under TTC was considerably reduced, with clear improvements reflected not only in mean error values but also in reduced variability and narrower confidence intervals. These results emphasize that TTC provides meaningful performance gains over SL and TTA, particularly in personalized biosignal scenarios. Notably, the overall MAE on the ICU dataset is higher than on PulseDB, reflecting the greater standard deviation in SBP and DBP values within ICU data and the increased difficulty of the prediction task.

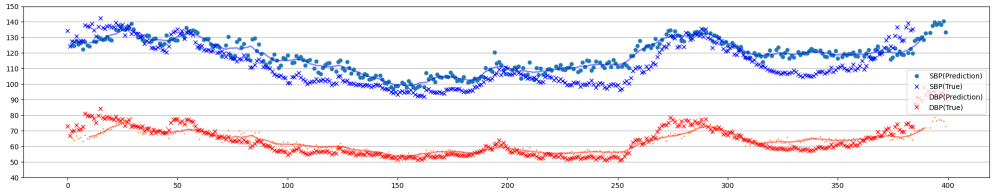
2. In Table 1, the number of valid segments specified for the test set is 94,961; however, since labels were obtained at approximately one label per hour, only five segments before and after each labeled segment were evaluated. Thus, the actual number of segments used for evaluation is approximately 22,240.



(*a*) Supervised learning



(*b*) Test-time adaptation



(*c*) Test-time calibration

Figure 3: Qualitative results of the prediction performance.

Qualitative Performance Assessment: We conducted a qualitative performance assessment. Figure 5 presents a visualization of the prediction results for a subject from the PulseDB dataset. For the PulseDB dataset, as described in Section 4.1, in the TTC scenario, a labeled sample is inserted for calibration once every ten samples, while all labeled samples are annotated. In the figure, markers labeled 'X' denote the actual measured SBP/DBP values, while markers labeled 'o' indicate the predicted values. The line means the overall trend of the model's predicted values over time. In this visualization, all labeled sample positions are annotated to show how well the model follows the temporal trend.

As shown in Figure 5(a), the SL baseline produces nearly constant, linear outputs, making its predictions less reliable. In contrast, Figure 5(b) and 5(c) reveal that both TTA and TTC display more variable prediction trajectories as the model continuously adapts to incoming samples. Notably, TTC demonstrates a more dynamic and responsive adaptation to shifting biosignal patterns than TTA. This superior performance stems primarily from the calibration effect in TTC, which significantly enhances alignment between predicted and ground-truth blood pressure measurements.

Additional examples from the PulseDB dataset are provided in Appendix B.[3] Although cases with minimal signal variation might sometimes favor SL, such instances are rare, and overall, both TTA and TTC consistently outperform the SL baseline. These findings, as confirmed by our supplementary materials, highlight the robustness of TTC in adapting to the dynamic nature of biosignals.

4.3.2. Ablation Study

We conducted an ablation study to evaluate the impact of different techniques in our proposed framework. Specifically, we analyzed how model performance changes with varying sizes of the dual buffer and different sampling weights.

3. The ICU dataset was not visualized because ground-truth labels are only available approximately once per hour, making it difficult to assert that dynamic changes are captured. However, supplementary results confirm that TTC exhibits similarly dynamic adaptation on the ICU dataset as on PulseDB.

Table 3: Mean absolute errors with 95% confidence interval and standard deviation for varying sizes of $B_{unlabel}$ and B_{label}; bold indicates the best performance

(a) PulseDB

$B_{unlabel}$	**8** B_{label}	**16** B_{label}
	SBP/DBP	**SBP/DBP**
16	10.41 ± 0.08 (13.66) / 5.86 ± 0.05 (7.87)	11.35 ± 0.09 (14.75) / 6.36 ± 0.05 (8.47)
32	9.53 ± 0.07 (12.61) / 5.33 ± 0.04 (7.20)	10.65 ± 0.08 (13.88) / 5.96 ± 0.05 (7.95)
64	**8.84 ± 0.07 (11.84) / 4.94 ± 0.04 (6.78)**	**9.63 ± 0.07 (12.66) / 5.32 ± 0.04 (7.16)**

(b) ICU

$B_{unlabel}$	**8** B_{label}	**16** B_{label}
	SBP/DBP	**SBP/DBP**
16	11.75 ± 0.26 (16.0) / **8.55 ± 0.20 (12.0)**	12.03 ± 0.26 (16.0) / **8.84 ± 0.21 (13.0)**
32	**11.74 ± 0.26 (16.0)** / 8.56 ± 0.20 (12.0)	12.03 ± 0.26 (16.0) / 8.88 ± 0.21 (13.0)
64	11.75 ± 0.26 (16.0) / 8.57 ± 0.20 (12.0)	**12.01 ± 0.26 (16.0)** / 8.89 ± 0.21 (13.0)

Varying the size of the dual buffer: We set different sizes for the dual buffer to assess their impact on model performance. The batch size was maintained at 32, with a 1:3 ratio of unlabeled to labeled samples. A larger buffer size can help the model retain and utilize more historical data, aiding in learning from past trends, yet it may reduce the model's ability to adapt to recent changes. Because obtaining enough labeled samples to fill B_label can be challenging, we fixed B_label = 8 or 16 and varied B_unlabel = 16, 32, and 64.

Table 3 shows the results depending on the dual-buffer. For the PulseDB dataset, increasing B_unlabel consistently reduces the MAE for both SBP and DBP. For example, raising B_unlabel lowers the SBP MAE regardless of the B_label size. This implies that maintaining a larger amount of unlabeled historical data can improve performance over time. In the ICU dataset, unlike in PulseDB, we observed relatively stable performance across different buffer sizes without a pronounced trend. This difference appears to stem from how performance is evaluated: the ICU dataset contains limited labels, and assessment focuses on each label and its surrounding samples (as shown red rectangle in Figure 1). Consequently, it is difficult to confirm whether the entire distribution is fully captured, a limitation also noted in qualitative experimental results. We additionally observe that the MAE is consistently lower for B_label = 8 than for B_label = 16 in both datasets. A larger labeled buffer retains more older samples, potentially reducing adaptability to newer data. Conversely, a smaller labeled buffer updates more frequently, mitigating the influence of outdated samples and facilitating adaptation.

As a result, there is a trade-off: a larger unlabeled buffer helps capture more long-term context, whereas a smaller labeled buffer enables the model to incorporate fresh labeled data more rapidly.

Varying the Proportion of Labeled Samples in a Batch: We evaluated the impact of varying the ratio of labeled to unlabeled samples in a batch on model performance. Labeled samples provide more accurate information about the subject's status, and increasing their proportion can enhance the model's calibration. In this experiment, we fixed the batch size at 32 and varied the proportion of labeled samples to 25% (8 labeled samples), 50% (16 labeled samples), and 75% (24 labeled samples) using a dual buffer with $B_{\text{label}} = 8$ and $B_{\text{unlabel}} = 64$.

As shown in Table 4, increasing the proportion of labeled samples leads to a substantial reduction in MAEs for both SBP and DBP on the PulseDB dataset. On the ICU dataset, although the improvement in performance was not as pronounced, a sim-

Table 4: Mean absolute errors with 95% confidence interval and standard deviation for varying the proportion of labeled samples; bold indicates the best performance

Dataset	Type	25%	50%	75%
PulseDB	**SBP**	15.47 ± 0.11 (19.33)	12.49 ± 0.09 (16.01)	**8.84 ± 0.07 (11.84)**
	DBP	8.94 ± 0.07 (11.34)	7.16 ± 0.06 (9.32)	**4.94 ± 0.04 (6.78)**
ICU	**SBP**	11.91 ± 0.26 (16.0)	**11.74 ± 0.26 (16.0)**	11.75 ± 0.26 (16.0)
	DBP	8.66 ± 0.20 (12.0)	8.60 ± 0.20 (12.0)	**8.57 ± 0.20 (12.0)**

ilar trend was observed. These results suggest that for datasets with higher variability in signal patterns, a higher proportion of labeled samples significantly improves model calibration and overall performance, effectively acting as a more thorough fine-tuning for the target domain.

The ablation study shows that optimizing the buffer size and increasing the proportion of labeled samples improve model performance by balancing historical data retention and adaptation to new data. Higher proportions of labeled samples, in particular, enhance calibration and fine-tuning for the target domain.

5. Conclusion

In this work, we presented a novel Test-Time Calibration (TTC) framework designed specifically for real-world biosignal applications, where models must continuously adapt to a constant flow of unlabeled data while relying on sporadic labeled samples for calibration. By integrating self-supervised adaptation with periodic supervised calibration and leveraging a dual buffer along with a weighted batch sampling strategy, our framework dynamically adjusts to changing data distributions and effectively captures personalized signal dynamics.

Our extensive evaluation on both the publicly available PulseDB dataset and a private ICU dataset demonstrates that TTC significantly reduces prediction errors for blood pressure measurements compared to conventional SL and TTA approaches. Both quantitative results and qualitative analyses confirm that TTC not only improves accuracy but also robustly tracks temporal variations in individual biosignals. These findings highlight the promise of our approach for enhancing real-time monitoring and prediction in diverse healthcare settings.

Our current investigation is limited to blood pressure prediction and does not include clinical analyses such as subgroup analyses based on demographic or clinical factors. Moreover, our evaluation is restricted to a limited set of biosignal datasets, and the real-time adaptation and calibration processes in our approach may introduce latency in practical applications (e.g., wearable devices). Additionally, the substantial differences in preprocessing requirements across datasets pose challenges for scalability and practical applicability. In future work, we plan to extend our TTC framework to other biosignal modalities—such as ECG-based arrhythmia detection or glucose monitoring—to demonstrate its versatility in personalized healthcare solutions. We also aim to improve the model architecture, develop more efficient training procedures, and evaluate the framework across a broader range of biosignal datasets and clinical scenarios. Furthermore, we will explore more streamlined, dataset-specific preprocessing strategies and incorporate detailed clinical analyses, including subgroup analyses based on demographic and clinical factors, to further validate and refine our approach.

Institutional Review Board (IRB) All data collection for the retrospective analysis was approved by the Institutional Review Board of Seoul National University Bundang Hospital (IRB No: B-2502-957-103).

Acknowledgments

This work was supported by (1) Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by Ministry of Science and ICT (MSIT) in the Korea government (RS-2024-00444014, Global AI-ECG-based Software Medical Device Approval and Commercialization International Collaboration) and (2) the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. 2022R1A2C1009047).

References

Liang Chen, Yong Zhang, Yibing Song, Ying Shan, and Lingqiao Liu. Improved test-time adaptation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24172–24182, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Rodolfo J Galindo and Grazia Aleppo. Continuous glucose monitoring: the achievement of 100 years of innovation in diabetes technology. *Diabetes research and clinical practice*, 170:108502, 2020.

Yulu Gan, Yan Bai, Yihang Lou, Xianzheng Ma, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7595–7603, 2023.

Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35:29374–29385, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Yufan He, Aaron Carass, Lianrui Zuo, Blake E Dewey, and Jerry L Prince. Autoencoder based self-supervised test-time adaptation for medical image analysis. *Medical image analysis*, 72:102136, 2021.

Zhiqiang He. Ecg heartbeat classification under dataset shift. *J. Intell. Med. Healthc*, 1:79–89, 2022.

Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3 (1):1–9, 2016.

Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. Vitaldb, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1):279, 2022.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, pages 6028–6039. PMLR, 2020.

Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34:21808–21820, 2021.

Xiankai Lu, Chao Ma, Bingbing Ni, Xiaokang Yang, Ian Reid, and Ming-Hsuan Yang. Deep regression tracking with shrinkage loss. In *Proceedings of the European conference on computer vision (ECCV)*, pages 353–369, 2018.

Wenao Ma, Cheng Chen, Shuang Zheng, Jing Qin, Huimao Zhang, and Qi Dou. Test-time adaptation with calibration of medical image classification nets for label distribution shift. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 313–323. Springer, 2022.

Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning (ICML)*, pages 8187–8202. PMLR, 2022.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Samsung Electronics. Samsung health monitor, 2024. URL https://www.samsung.com/ca/apps/samsung-health-monitor/. Accessed: 2024-08-31.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9229–9248. PMLR, 13–18 Jul 2020.

Linda Diann Urden, Kathleen M Stacy, and Mary E Lough. *Critical care nursing, diagnosis and management, 7: critical care nursing*. Elsevier Health Sciences, 2013.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7201–7211, 2022.

Weinan Wang, Pedram Mohseni, Kevin L Kilgore, and Laleh Najafizadeh. Pulsedb: A large, cleaned dataset based on mimic-iii and vitaldb for benchmarking cuff-less blood pressure estimation methods. *Frontiers in Digital Health*, 4:1090854, 2023.

Shiqi Yang, Joost Van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in neural information processing systems*, 34:29393–29405, 2021.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

Appendix A. Model details

A.1. PulseDB

Training phase We adopt a hybrid architecture by integrating ResNet He et al. (2016) and Vision Transformer Dosovitskiy et al. (2020) to process ECG and PPG signals. Two input channels are sampled at 125 Hz over a 10-second window, which is then divided into patches of size 30 with a 15-step overlap. These segmented signals are first passed through a single ResNet block (one layer, inplanes 256, expand 2, kernel size 5, and ReLU activation) to generate patchified features. The resulting features then feed into a Transformer encoder configured with a 512-dimensional embedding, eight attention heads, four consecutive blocks, and GELU activation, relying on fixed positional encoding. For outputs, a single Transformer block (decoder depth of 1, 512-dimensional embedding, and eight decoder heads) reconstructs the original signal. Additionally, a simple linear layer regresses systolic and diastolic blood pressure, integrating auxiliary inputs (age and gender) that are embedded into a 32-dimensional vector.

In the initialization setting, the LAMB optimizer You et al. (2019) is employed with a learning rate of 0.002516152722130999 and a weight decay of 0.00012775200014634934, determined via a population-based training (PBT) scheduler Jaderberg et al. (2017). The loss function combines a shrinkage loss, specified by `p` = 2.0, `speed` = 10, and `loc` = 0.0, with mean squared error (MSE). In the shrinkage loss, the parameter `p` controls the power used to measure deviations from the target value. The parameter `speed` dictates how aggressively these deviations are pulled toward the reference point, while `loc` defines the specific target (or location) toward which errors are shrunk. In the fine-tuning setting, the LAMB optimizer is used with a reduced learning rate of 0.000570373305632515 and a weight decay of 0.0015399401149371588, again discovered via population-based training. The shrinkage loss is now configured with `p` = 1.0, `speed` = 10, and `loc` = 0.1, with MSE.

Test phase the pretrained model is utilized without additional modification. The same shrinkage loss configuration (`loc` = 0.1, `p` = 1.0, `speed` = 10) is combined with MSE. For optimization, the SGD optimizer is fixed with a learning rate of 0.001, momentum of 0.9, and a weight decay of 0.001.

A.2. ICU dataset

Training phase We likewise construct a hybrid architecture by combining ResNet He et al. (2016) and Vision Transformer Dosovitskiy et al. (2020). This model is configured to handle a maximum of 7500 time steps at 125 Hz for two input leads, which are segmented into patches of size 125 with no overlap (patch shift of 125). The segmented signals are first passed through two ResNet blocks (one layer each, inplanes 128, expand 2, kernel size 5, and ReLU activation) to derive patchified features. These features are then processed by a Transformer encoder comprising four consecutive blocks, each with a 512-dimensional embedding, eight attention heads, a 0.15 drop path probability, fixed positional encoding, and GELU activation. The reconstruction head uses a single-layer Transformer block with a 512-dimensional embedding, eight attention heads, and four blocks under fixed positional encoding to reconstruct the original signal. For blood pressure regression, the model includes a dedicated Transformer block with the same specifications as the reconstruction head, aiming to predict systolic and diastolic values without auxiliary inputs.

In the initialization setting, the LAMB optimizer You et al. (2019) is used with a learning rate of 0.0009895934981800387, a weight decay of 6.1418×10^{-5} as determined by the PBT scheduler Jaderberg et al. (2017). The loss function combines the shrinkage loss (with `p`=1.0, `speed`=10, and `loc`=0.1 and MSE. In the fine-tuning phase, the LAMB optimizer is reconfigured with a learning rate of 0.0007908773293493147 and a weight decay of 0.0005890236102532118. The shrinkage loss parameters are adjusted to `p`=2.5, `speed`=10, and `loc`=0.0, again combined with MSE.

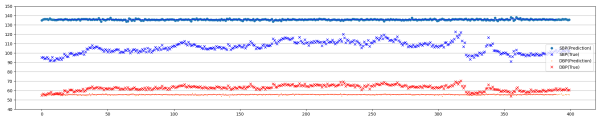
Test phase The pretrained model is used without additional modification. The shrinkage loss is set to `p`=2.5, `speed`=10, `loc`=0.0 with MSE. For optimization, the SGD optimizer is deployed with a learning rate of 0.001, momentum of 0.9l and `weight_decay`=1×10^{-6}.

Appendix B. Qualitative Results

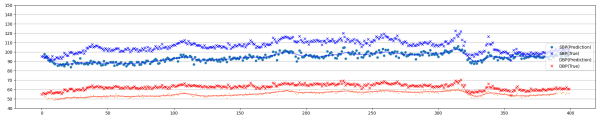
In the figure, markers labeled 'X' denote the actual measured SBP/DBP values, while markers labeled 'o' indicate the predicted values. The line means the over all trend of the model's predicted values over time. In this visualization, all labeled sample positions are

annotated to show how well the model follows the temporal trend.

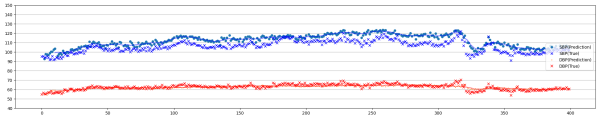
PulseDB A labeled sample is inserted for calibration once every ten samples, causing the TTC framework to produce predictions more closely aligned with the ground-truth labels. As shown in Section 4.3.1, we observed a similar trend in those results.



(*a*) Supervised learning

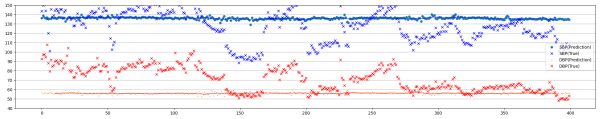


(*b*) Test-time adaptation

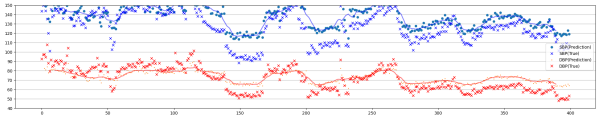


(*c*) Test-time calibration

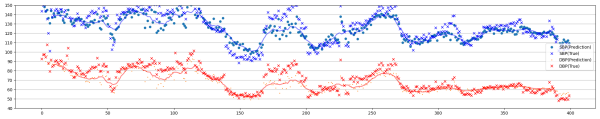
Figure 4: Qualitative results of `p028331_0` on PulseDB.



(*a*) Supervised learning



(*b*) Test-time adaptation

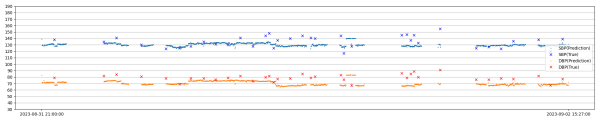


(*c*) Test-time calibration

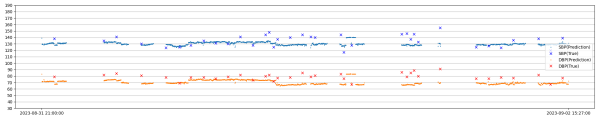
Figure 5: Qualitative results of `p064771_0` on PulseDB.

ICU dataset Labels are only available around once per hour, meaning that label inputs occur far less frequently than the predicted values (denoted by 'o').

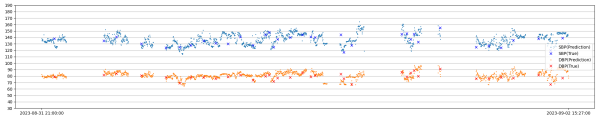
We observe that the TTC approach exhibits more dynamic changes compared to TTA or SL.



(*a*) Supervised learning

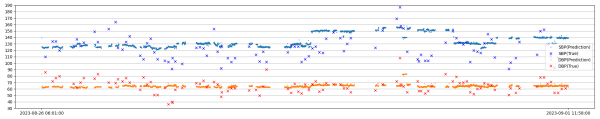


(*b*) Test-time adaptation

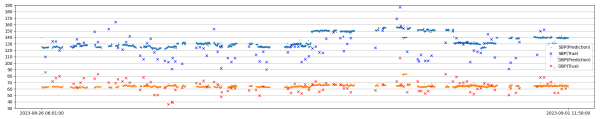


(*c*) Test-time calibration

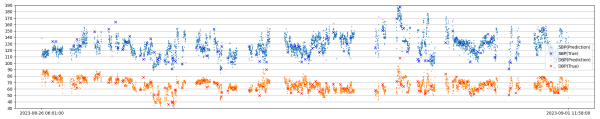
Figure 6: Qualitative results of `index 8` on ICU dataset.



(*a*) Supervised learning



(*b*) Test-time adaptation



(*c*) Test-time calibration

Figure 7: Qualitative results of `index 23` on ICU dataset.