



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Test-Time Adaptation for ECG Classification via SQI-Gated Self-Training and Beat-Rhythm Consistency

Wenhan Jiang¹, Zhipeng Deng¹ ✉, Jiale Zhou¹, Haolin Wang², Yafei Ou³, and Yefeng Zheng¹ ✉

¹ Westlake University, Hangzhou, China

² Hokkaido University, Sapporo, Japan

³ RIKEN, Japan

{jiangwenhan, dengzhipeng, zhoujiale, zhengyefeng}@westlake.edu.cn;
yafei.ou@riken.jp

Abstract. Deep learning models for electrocardiogram (ECG) classification often suffer from significant performance degradation when deployed in unseen domains due to shifts in acquisition devices and patient populations. Test-time adaptation (TTA) offers a practical solution by adapting models using only unlabeled data at inference time. However, existing TTA methods often underperform on ECG tasks, since naive online updates ignore the hierarchical beat-rhythm structure of cardiac cycles and are vulnerable to signal artifacts, which leads to unstable adaptation and model drift. We propose **BeatRhythm-TTA**, an ECG-tailored TTA framework that explicitly accounts for ECG’s noisy observations and structured beat-rhythm semantics under domain shift. First, to handle pervasive ECG artifacts, we introduce a Signal Quality Index (SQI)-gated adaptation scheme that selectively filters out low-quality signals to prevent harmful updates. Second, to leverage ECG’s beat-rhythm semantics, we enforce dual-level consistency so the model preserves beat morphology and rhythm dynamics while adapting to shifted acquisition conditions. Extensive experiments on multi-label ECG diagnosis across three adaptation protocols, using PTB-XL as the source domain and CPSC2018/Georgia as two target domains, demonstrate the effectiveness of our method, yielding an average +2.70% relative improvement in Macro-F1 over the best competing method. The code is available at <https://github.com/Vincent-Wenhan/BeatRhythm-TTA>.

Keywords: Test-Time Adaptation · ECG Classification · Domain Shift

1 Introduction

Although deep learning has achieved strong performance on electrocardiogram (ECG) interpretation [17,23], generalizing pretrained models to unseen clinical environments remains challenging due to distribution shifts introduced by heterogeneous hardware and patient cohorts [9,20]. Re-training or fine-tuning for

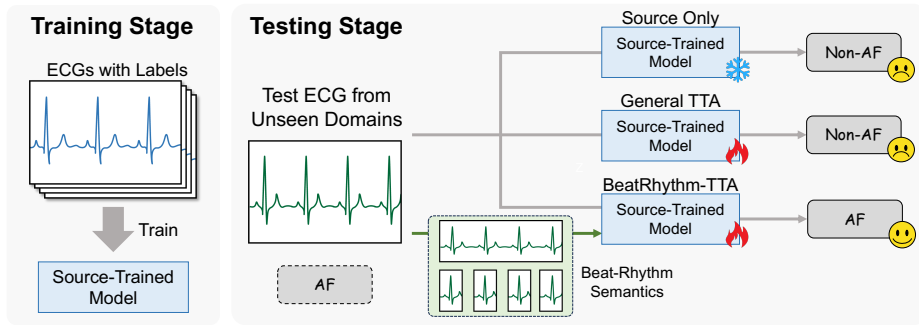


Fig. 1. Motivation of the proposed BeatRhythm-TTA. Comparison of TTA strategies. Generic TTA methods are brittle on ECG, while BeatRhythm-TTA leverages beat-rhythm semantics for robust adaptation. (AF denotes atrial fibrillation, a cardiac rhythm disorder.)

each new target domain is often impractical in clinical practice as it requires substantial expert annotation and repeated engineering effort [20]. This motivates adopting test-time adaptation (TTA) from the broader machine learning literature [27,28], a paradigm that adapts a deployed model using only unlabeled test data, without requiring target labels or retraining from scratch.

Most TTA methods have been developed and validated primarily on vision benchmarks [27,28]. These methods are usually based on normalization calibration (e.g., AdaBN [15], MemBN [11]), entropy minimization (e.g., TENT [27], SAR [19], DeYO [14]), or confidence- or energy-based objectives (e.g., COME [32], TEA [30]). Recent work has extended TTA beyond vision, showing promising performance in speech, audio, and AI-for-science under distribution shifts [12,13,31,33]. A closely related paradigm is source-free domain adaptation (SFDA) [10,16,29], which similarly addresses domain shifts without source data but is typically formulated as an offline adaptation stage before deployment. More broadly, annotation-efficient and domain-generalized learning have been investigated in medical AI to reduce supervision cost and improve robustness under domain shifts [4,5]; however, these approaches are typically training-stage solutions, whereas our setting requires adapting a fixed ECG classifier using only unlabeled target samples at inference time. Despite progress in other modalities, ECG-specific TTA remains underexplored, as directly transferring existing methods raises modality-specific challenges. First, adapting models on continuous ECG streams corrupted by transient artifacts (e.g., baseline wander, motion artifacts) amplifies spurious correlations and triggers drift. Second, many existing TTA objectives (e.g., global entropy minimization) ignore the hierarchical beat-rhythm semantics that underlie ECG diagnosis [2,25].

To this end, we propose BeatRhythm-TTA, a semantically grounded TTA framework comprising two complementary components. First, **SQI-gated adaptation** selectively scales or blocks test-time updates based on a signal quality index (SQI) to avoid harmful adaptation on noisy segments. Second, **dual-level**

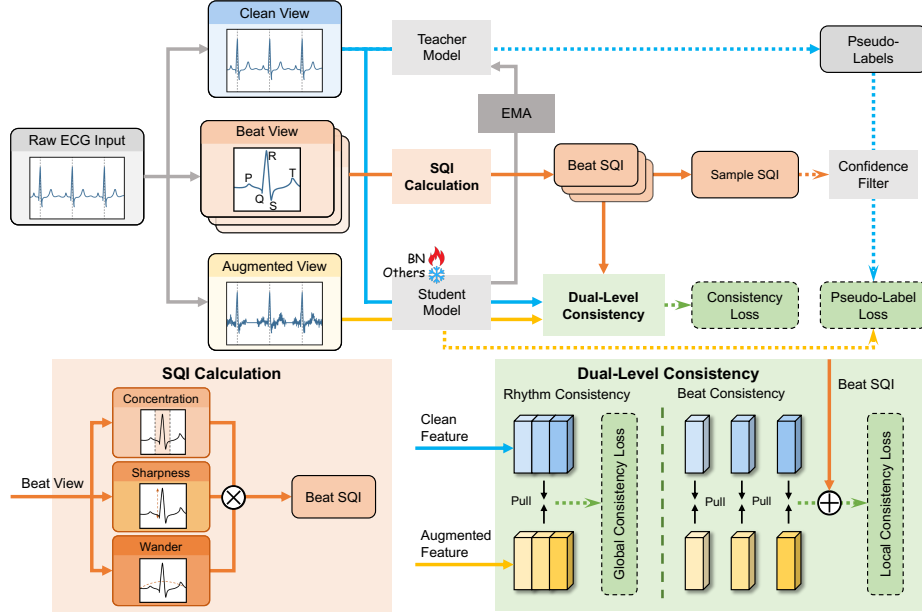


Fig. 2. BeatRhythm-TTA Framework. Given an unlabeled target ECG, adaptation is driven by two complementary modules. (1) **SQI-gated pseudo-labeling:** Blocks self-training updates based on teacher confidence and beat-wise signal quality, where EMA (exponential moving average) maintains a slowly updated teacher for more stable pseudo-labels. (2) **Dual-level consistency:** Enforces rhythm- and beat-level agreement between clean and augmented views to stabilize adaptation.

consistency enforces beat- and rhythm-level agreement to preserve physiologically motivated semantics at two temporal scales [8]: local beat morphology (e.g., ST-segment changes [25]) and global rhythm irregularities (e.g., atrial fibrillation [2]). By coupling reliability-aware gating with physiologically grounded consistency, BeatRhythm-TTA reduces drift under real-world domain shifts. Extensive experiments on multi-label ECG classification across three adaptation protocols, using PTB-XL as the source domain and CPSC2018/Georgia as two target domains, validated the effectiveness of our approach. It outperforms the strongest competing methods with an average +2.70% relative improvement in Macro-F1 across protocols.

2 Method

2.1 Preliminaries and Overview

Problem definition. Let $f(\cdot; \theta)$ be an ECG classifier that maps a C -lead segment $x \in \mathbb{R}^{C \times T}$ to logits $z \in \mathbb{R}^K$, with multi-label targets $y \in \{0, 1\}^K$, where

K denotes the number of diseases to classify. We first train θ^S on the labeled source domain $\mathcal{D}^S = \{(x_i^S, y_i^S)\}_{i=1}^{N^S}$:

$$\theta^S = \arg \min_{\theta} \mathbb{E}_{(x^S, y^S) \sim \mathcal{D}^S} \mathcal{L}_{\text{sup}}(f(x^S; \theta), y^S). \quad (1)$$

At deployment, unlabeled target samples $\{x_i^T\}_{i=1}^{N^T}$ arrive from an unseen domain $\mathcal{D}^T$. Test-time adaptation (TTA) updates the model using only x^T by minimizing a self-supervised loss:

$$\theta^T = \arg \min_{\theta} \mathbb{E}_{x^T \sim \mathcal{D}^T} \mathcal{L}_{\text{tta}}(f(x^T; \theta), x^T). \quad (2)$$

Offline and online TTA settings. We consider both offline and online test-time adaptation settings. Offline TTA assumes the entire target test distribution $\{x_i^T\}_{i=1}^{N^T}$ is available for adaptation and we optimize $\mathcal{L}_{\text{tta}}$ on the unlabeled target test set and then evaluate on the same set. Online TTA processes test samples sequentially: as each sample arrives, the model is updated on it and predicts on-the-fly. We consider three protocols: (i) offline, (ii) continual online, and (iii) independent online TTA. Continual online carries updates across samples ($\theta_i \rightarrow \theta_{i+1}$), while independent online resets to θ^S for each sample.

Method Overview. The overview of our method is shown in Fig. 2. In real ECG recordings, signal quality can vary due to artifacts, and diagnostic semantics appear at both the beat and rhythm levels. Motivated by these properties, we propose BeatRhythm-TTA, an ECG-tailored framework with two complementary modules: an SQI-gated pseudo-labeling module that computes a beat-wise SQI to gate teacher-driven self-training, and a dual-level consistency module that enforces rhythm- and beat-level agreement across clean and augmented views.

2.2 SQI-gated Pseudo-labeling

To prevent harmful updates triggered by noisy samples and segments, we perform selective self-training at test time, only updating the model when the current ECG sample is both confident under a teacher model and reliable according to a signal quality index (SQI). Signal quality indices have been widely used to quantify waveform reliability in ECG analysis [3,22]. Unlike prior SQIs that mainly target record-level screening and may require heavier, domain-tuned features, our SQI is a lightweight beat-wise score designed for test-time gating, computed online without labels to control update eligibility and strength.

Given an unlabeled ECG sample $x \in \mathbb{R}^{C \times T}$, we detect R-peaks $\{r_i\}_{i=1}^{M_b}$ using the WFDB toolchain [7], where M_b denotes the number of detected beats in the sample. We select a reference lead based on derivative energy and apply robust normalization with basic physiological sanity checks. We then extract beat-centered windows and scale the window length by the median RR interval to account for heart-rate variation. For each beat window $x^{(i)}$, we compute three lightweight morphology factors:

(i) **QRS concentration** ($\text{conc}^{(i)}$): the QRS complex, the sharp waveform around the R-peak in the P-QRS-T beat structure, carries concentrated energy

in clean beats [26]; we measure it as the energy ratio of a small center region to the whole beat window;

(ii) **sharpness** ($\text{sharp}^{(i)}$): true QRS has steep slopes, measured by a high-quantile magnitude of the first-order difference normalized by signal root mean square;

(iii) **baseline wander ratio** ($\text{bwr}^{(i)}$): strong low-frequency drift, measured as the energy ratio of a moving-average baseline estimate to the whole window, indicates poor motion.

We combine them to obtain a beat SQI score $\text{SQI}^{(i)}$:

$$\text{SQI}^{(i)} = \sigma(\text{conc}^{(i)}) \cdot \sigma(\text{sharp}^{(i)}) \cdot \sigma(-\text{bwr}^{(i)}), \quad (3)$$

where $\sigma(\cdot)$ is the sigmoid function. Finally, we aggregate beat-wise scores into a sample-level quality weight $w = \frac{1}{M_b} \sum_i \text{SQI}^{(i)}$ (mean over beats).

We adopt a teacher-student scheme to generate pseudo labels from a clean view and supervise the student on an augmented view as:

$$\mathcal{L}_{\text{pl}} = \mathbb{I}[c(x) \geq \tau_c] \mathbb{I}[w(x) \geq \tau_q] \ell_{\text{bce}}(p_\theta(\tilde{x}), p_{\bar{\theta}}(x^c)), \quad (4)$$

where $x^c = x$ is the clean view, $\tilde{x} = \mathcal{A}(x)$ is a time-aligned augmentation, and ℓ_{bce} denotes binary cross-entropy (BCE) loss. We denote $p_\theta(x) = \sigma(f_\theta(x))$ as multi-label probabilities, and use a teacher $f_{\bar{\theta}}$ to produce soft pseudo-labels $p_{\bar{\theta}}(x^c)$. We define the confidence score as $c(x) = \min_k |p_{\bar{\theta}}(x^c)_k - \frac{1}{2}|$, and $\mathbb{I}[\cdot]$ denotes the indicator function, so activate pseudo-label updates only when both the confidence and the SQI thresholds are satisfied. The teacher is updated by exponential moving average (EMA) [24,28]: $\bar{\theta} \leftarrow \mu \bar{\theta} + (1 - \mu)\theta$. This EMA update stabilizes pseudo-labels and mitigates rapid forgetting during online adaptation.

2.3 Dual-level Consistency

ECG semantics appear at two temporal scales: beat-level morphology and rhythm-level dynamics across beats [8]. Diagnoses may rely on subtle beat morphology such as ST-segment elevation/depression and T-wave inversion [25] or rhythm irregularity across multiple beats such as atrial fibrillation (AF) [2]. Accordingly, to stabilize adaptation while preserving clinically meaningful morphology, we regularize test-time updates with a dual-level consistency objective that aligns global sample representations and local beat-centric representations between a clean view x^c and a time-aligned augmented view $\tilde{x}$.

Let $F(\cdot)$ denote intermediate features extracted from a chosen backbone layer. Given the two time-aligned views, we obtain feature of the clean view ($F_c = F(x^c)$), the augmented view ($F_a = F(\tilde{x})$), and minimize a cosine alignment loss:

$$\mathcal{L}_{\text{rhythm}} = 1 - \cos(\text{sg}(F_c), F_a), \quad (5)$$

where $\text{sg}(\cdot)$ denotes stop-gradient operator.

Rhythm alignment captures global dynamics but does not explicitly constrain beat morphology. Using the previously detected R-peaks $\{r_i\}$ and the RR-based

windowing, we extract beat-centric embeddings from F_c and F_a , yielding paired sets $\{b_c^{(i)}\}_{i=1}^{M_b}$ and $\{b_a^{(i)}\}_{i=1}^{M_b}$. We align beat embeddings using cosine distance:

$$\ell_{\text{beat}}^{(i)} = 1 - \cos(\text{sg}(b_c^{(i)}), b_a^{(i)}), \quad (6)$$

and aggregate across beats using the beat-wise SQI weights from Eq. (3), emphasizing reliable beats and suppressing corrupted ones during adaptation:

$$\mathcal{L}_{\text{beat}} = \frac{\sum_i \omega^{(i)} \ell_{\text{beat}}^{(i)}}{\sum_i \omega^{(i)} + \epsilon}, \quad \omega^{(i)} = \max(\text{SQI}^{(i)}, \tau_{\text{floor}}), \quad (7)$$

where $\tau_{\text{floor}} = 0.05$ sets a minimum beat weight.

We combine the dual-level consistency losses with the SQI-gated pseudo-label loss:

$$\mathcal{L} = \mathcal{L}_{\text{pl}} + \lambda_{\text{beat}} \mathcal{L}_{\text{beat}} + \lambda_{\text{rhythm}} \mathcal{L}_{\text{rhythm}}. \quad (8)$$

λ_{beat} and λ_{rhythm} are scalar hyperparameters that balance the contributions of beat-level and rhythm-level consistency against the pseudo-labeling loss. By optimizing this objective, the model is updated via SQI-gated self-training, while dual-level consistency regularizes test-time updates to preserve both beat- and rhythm-level semantics.

3 Experiments

3.1 Datasets and Experiment Setup

We evaluated our method on three public 12-lead ECG datasets from the PhysioNet/CinC Challenge 2020 [21]. Specifically, we used PTB-XL (Germany, 21,837 records) as the source domain, and evaluated cross-domain adaptation on CPSC2018 (China, 6,877 records) and Georgia (USA, 10,344 records) as two distinct target domains. All datasets were projected into a shared 6-class label space (atrial fibrillation, first-degree atrioventricular block, left bundle branch block, right bundle branch block, premature atrial contraction, and normal sinus rhythm), discarding unshared labels and empty samples.

Evaluation metrics. We report Macro-F1 and Macro-AUC on target test sets, averaged over 3 runs. Macro averaging gives equal weight to each class and is suitable for imbalanced multi-label ECG diagnosis.

Implementation details. Experiments were implemented in PyTorch on an RTX 4090 GPU. All ECGs were resampled to 500 Hz and padded to a fixed length of 10 s. Following [1,6,18], a 1D ResNet-18 was pretrained on PTB-XL (8:2 train/val split) for 100 epochs using the BCE loss, standard augmentations, and the Adam optimizer ($lr = 10^{-4}$, batch size 512). The best validation checkpoint was used as the common source model for all methods. For fair comparison, all methods shared the same source checkpoint, preprocessing, target stream, and protocol-specific adaptation budget, followed official/default protocols, and disabled dropout during adaptation and evaluation. Offline TTA adapted on

the target set for 1 epoch (batch size 256). Continual and independent online TTA processed single samples (batch size 1) for 5 and 10 internal steps using the Adam optimizer with $lr=10^{-6}$ and 10^{-4} , respectively. BeatRhythm-TTA updated only BatchNorm affine parameters. Thresholds in Eq. (4) were fixed to $\tau_c=0.2$ and $\tau_q=0.05$ across all target datasets and protocols without target-label tuning. The loss weights in Eq. (8) were set to $\lambda_{\text{beat}}=0.5$ and $\lambda_{\text{rhythm}}=1.0$ for all experiments. Strong-view augmentation used Gaussian noise, gain jitter, and one temporal mean-mask, without time shift, cropping, or warping to preserve beat-window alignment.

3.2 Results and Analysis

We compared our method against the source-only baseline and representative TTA methods (AdaBN [15], MemBN [11], TENT [27], SAR [19], DeYO [14], COME [32] and TEA [30]). For the offline setting, we additionally evaluated source-free domain adaptation baselines (SHOT [16], SF(DA)² [10] and UCon [29]). Tables 1–2 show that our method consistently achieves the best average Macro-AUC and Macro-F1 across offline and online protocols. Paired t-tests show statistically significant Macro-F1 gains over the strongest baseline across all three protocols, while class-wise inspection suggests that the gains are not driven by a single diagnostic class.

Offline TTA. Under batch adaptation (Table 1), BeatRhythm-TTA improves average Macro-F1 from 60.64% to 62.42% and Macro-AUC from 88.13% to 88.84%, showing effective adaptation when the target set can be revisited.

Continual online TTA. In the streaming setting without resets (Table 2), BeatRhythm-TTA improves Macro-F1 by +1.12% and Macro-AUC by +0.65%, indicating that SQI gating helps suppress unstable accumulated updates.

Independent online TTA. In the independent setting, BeatRhythm-TTA achieves the largest Macro-F1 gain (+1.97%), suggesting stable sample-wise adaptation under individual variation in morphology, rhythm, and noise.

Ablation Study. Table 3 shows that each component contributes to performance. Removing either SQI gating or consistency regularization degrades performance, while beat-level consistency brings further gains, confirming their complementarity.

4 Conclusion

We presented BeatRhythm-TTA, an ECG-tailored framework combining SQI-gated pseudo-labeling with dual-level consistency regularization. Across two target domains, our method consistently outperformed representative TTA and SFDA baselines under three protocols. The results highlight that reliability-aware updates and semantically grounded beat-rhythm constraints are useful for stable ECG adaptation under real-world shifts. Our current evaluation focuses on closed-set adaptation with a shared 6-class label space and relies on R-peak-based SQI estimation; future work will explore unseen diagnoses, detector errors, noise robustness, and broader clinical deployment.

Table 1. Offline TTA results. **Bold:** best, underlined: second best. All values are reported in percent (%).

Method	CPSC2018		Georgia		Avg	
	AUC	F1	AUC	F1	AUC	F1
Source Only	87.62	60.84	88.59	60.35	88.11	60.59
AdaBN [15]	83.25	46.71	87.33	52.29	85.29	49.50
TENT [27]	87.57	60.63	88.57	60.29	88.07	60.46
COME [32]	87.28	60.14	88.43	59.83	87.86	59.99
SAR [19]	83.43	47.63	87.48	53.40	85.45	50.52
MemBN [11]	83.51	47.58	87.58	54.32	85.55	50.95
TEA [30]	87.47	60.36	<u>88.62</u>	60.28	88.04	60.32
DeYO [14]	87.55	60.62	88.57	60.30	88.06	60.46
SHOT [16]	87.35	60.20	88.47	59.76	87.91	59.98
SF(DA) ² [10]	87.46	60.38	88.52	60.07	87.99	60.22
UCon [29]	<u>87.65</u>	<u>60.92</u>	88.61	<u>60.36</u>	<u>88.13</u>	<u>60.64</u>
Ours	88.70	62.57	88.97	62.26	88.84	62.42

Table 2. Online TTA results under continual and independent settings. **Bold:** best, underlined: second best within each setting. All values are reported in percent (%).

Method	CPSC2018		Georgia		Avg	
	AUC	F1	AUC	F1	AUC	F1
Source Only	87.62	60.84	88.59	60.35	88.11	60.59
<i>Continual Online TTA</i>						
AdaBN [15]	86.11	58.81	87.94	61.93	87.03	60.37
TENT [27]	84.92	53.68	87.40	56.15	86.16	54.92
COME [32]	87.29	60.63	88.54	60.33	87.92	60.48
SAR [19]	<u>87.63</u>	<u>60.86</u>	88.60	60.37	<u>88.12</u>	<u>60.62</u>
MemBN [11]	87.43	60.44	88.38	60.65	87.91	60.55
TEA [30]	87.52	60.77	88.59	60.25	88.06	60.51
DeYO [14]	87.38	60.55	88.54	60.24	87.96	60.39
Ours	88.98	62.53	<u>88.55</u>	<u>60.94</u>	88.77	61.74
<i>Independent Online TTA</i>						
AdaBN [15]	87.25	59.37	88.44	58.75	87.85	59.06
TENT [27]	87.20	60.37	88.43	59.87	87.82	60.12
COME [32]	87.35	60.47	88.50	60.19	87.93	60.33
SAR [19]	<u>87.64</u>	<u>60.90</u>	<u>88.61</u>	<u>60.41</u>	<u>88.13</u>	<u>60.66</u>
MemBN [11]	85.17	45.42	85.55	46.47	85.36	45.95
TEA [30]	87.55	60.68	88.59	60.23	88.07	60.45
DeYO [14]	87.35	60.39	88.47	59.96	87.91	60.18
Ours	88.27	62.97	88.63	62.30	88.45	62.63

Table 3. Ablation study under offline setting. All values are reported in percent (%).

SQI	Consistency		CPSC2018		Georgia		Avg	
	Rhythm	Beat	AUC	F1	AUC	F1	AUC	F1
✓	✓	✓	87.62	60.84	88.59	60.35	88.11	60.59
			88.45	61.73	88.89	61.58	88.67	61.66
			88.55	61.95	88.91	61.82	88.73	61.89
✓	✓		88.65	62.38	88.95	62.13	88.80	62.26
✓	✓	✓	88.70	62.57	88.97	62.26	88.84	62.42

Acknowledgments. This work was supported by Zhejiang Leading Innovative and Entrepreneur Team Introduction Program (2024R01007) and the “Pioneer” and “Leading Goose” Research and Development Program of Zhejiang (2025C02077).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abbaspourazad, S., Elachqar, O., Miller, A., Emrani, S., Nallasamy, U., Shapiro, I.: Large-scale training of foundation models for wearable biosignals. In: International Conference on Learning Representations (ICLR) (2024)
2. Carrington, M., Providência, R., Chahal, C.A.A., Ricci, F., Epstein, A.E., Gallina, S., Fedorowski, A., Sutton, R., Khanji, M.Y.: Monitoring and diagnosis of intermittent arrhythmias: evidence-based guidance and role of novel monitoring strategies. *European Heart Journal Open* **2**(6), oeac072 (2022)
3. Clifford, G.D., Behar, J., Li, Q., Rezek, I.: Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms. *Physiological Measurement* **33**(9), 1419–1433 (2012)
4. Deng, Z., Xu, Z., Isshiki, T., Zheng, Y.: Fedsemidg: Domain generalized federated semi-supervised medical image segmentation. *Medical Image Analysis* **112**, 104096 (2026). <https://doi.org/10.1016/j.media.2026.104096>
5. Deng, Z., Yang, Y., Suzuki, K.: Federated active learning framework for efficient annotation strategy in skin-lesion classification. *Journal of Investigative Dermatology* **145**(2), 303–311 (2025). <https://doi.org/10.1016/j.jid.2024.05.023>
6. Gao, Y., Kim, S., McIntosh, C.: EchoingECG: An electrocardiogram cross-modal model for echocardiogram tasks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 175–185. Springer (2025)
7. 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), e215–e220 (2000)
8. Hong, S., Xiao, C., Ma, T., Li, H., Sun, J.: MINA: Multilevel knowledge-guided attention for modeling electrocardiography signals. In: 28th International Joint Conference on Artificial Intelligence, IJCAI 2019. pp. 5888–5894 (2019)
9. Huang, Z., MacLachlan, S., Yu, L., Herbozo Contreras, L.F., Truong, N.D., Ribeiro, A.H., Kavehei, O.: Generalization challenges in electrocardiogram deep learning:

- insights from dataset characteristics and attention mechanism. *Future Cardiology* **20**(4), 209–220 (2024)
10. Hwang, U., Lee, J., Shin, J., Yoon, S.: SF(DA)2: Source-free domain adaptation through the lens of data augmentation. In: 12th International Conference on Learning Representations, ICLR 2024 (2024)
 11. Kang, J., Kim, N., Ok, J., Kwak, S.: MemBN: Robust test-time adaptation via batch norm with statistics memory. In: European Conference on Computer Vision (ECCV). pp. 467–483. Springer (2024)
 12. Kim, C., Park, J., Shim, H., Yang, E.: SGEM: Test-time adaptation for automatic speech recognition via sequential-level generalized entropy minimization. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. vol. 2023, pp. 3367–3371 (2023)
 13. Kim, H., Kim, S., Mok, J., Yoon, S.: Battling the non-stationarity in time series forecasting via test-time adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 17868–17876 (2025)
 14. Lee, J., Jung, D., Lee, S., Park, J., Shin, J., Hwang, U., Yoon, S.: Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In: 12th International Conference on Learning Representations, ICLR 2024 (2024)
 15. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Revisiting batch normalization for practical domain adaptation. arXiv preprint arXiv:1603.04779 (2016)
 16. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International Conference on Machine Learning. pp. 6028–6039. PMLR (2020)
 17. Liu, C., Cheng, S., Ding, W., Arcucci, R.: Spectral cross-domain neural network with soft-adaptive threshold spectral enhancement. *IEEE Transactions on Neural Networks and Learning Systems* **36**(1), 692–703 (2025)
 18. Liu, C., Wan, Z., Ouyang, C., Shah, A., Bai, W., Arcucci, R.: Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In: International Conference on Machine Learning. pp. 31949–31963. PMLR (2024)
 19. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. In: International Conference on Learning Representations (ICLR) (2023)
 20. Ong Ly, C., Unnikrishnan, B., Tadic, T., Patel, T., Duhamel, J., Kandel, S., Moayed, Y., Brudno, M., Hope, A., Ross, H., et al.: Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. *npj Digital Medicine* **7**(1), 124 (2024)
 21. Perez Alday, E.A., Gu, A., Shah, A.J., Robichaux, C., Wong, A.K.I., Liu, C., Liu, F., Bahrami Rad, A., Elola, A., Seyedi, S., Li, Q., Sharma, A., Clifford, G.D., Reyna, M.A.: Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement* **41**(12), 124003 (2020). <https://doi.org/10.1088/1361-6579/abc960>
 22. Rahman, S., Karmakar, C., Natgunanathan, I., Yearwood, J., Palaniswami, M.: Robustness of electrocardiogram signal quality indices. *Journal of the Royal Society Interface* **19**(189) (2022)
 23. Ribeiro, A.H., Ribeiro, M.H., Paixão, G.M., Oliveira, D.M., Gomes, P.R., Canazart, J.A., Ferreira, M.P., Andersson, C.R., Macfarlane, P.W., Meira Jr, W., et al.: Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications* **11**(1), 1760 (2020)
 24. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems* **30** (2017)

25. Vogel, B., Claessen, B.E., Arnold, S.V., Chan, D., Cohen, D.J., Giannitsis, E., Gibson, C.M., Goto, S., Katus, H.A., Kerneis, M., et al.: ST-segment elevation myocardial infarction. *Nature Reviews Disease Primers* **5**(1), 39 (2019)
26. Wagner, G.S., Strauss, D.G.: *Marriott's Practical Electrocardiography*. Lippincott Williams & Wilkins (2013)
27. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: TENT: Fully test-time adaptation by entropy minimization. In: *International Conference on Learning Representations* (2021), <https://openreview.net/forum?id=uXl3bZLkr3c>
28. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 7201–7211 (2022)
29. Xu, G., Guo, H., Yi, L., Ling, C., Wang, B., Yi, G.: Revisiting source-free domain adaptation: a new perspective via uncertainty control. In: *The Thirteenth International Conference on Learning Representations* (2025), <https://openreview.net/forum?id=nx9Z5Kva96>
30. Yuan, Y., Xu, B., Hou, L., Sun, F., Shen, H., Cheng, X.: TEA: Test-time energy adaptation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 23901–23911 (2024)
31. Zhang, C., Zheng, H., You, X., Zheng, Y., Gu, Y.: PASS: Test-time prompting to adapt styles and semantic shapes in medical image segmentation. *IEEE Transactions on Medical Imaging* **44**(4), 1853–1865 (2025)
32. Zhang, Q., Bian, Y., Kong, X., Zhao, P., Zhang, C.: COME: Test-time adaption by conservatively minimizing entropy. In: *International Conference on Learning Representations (ICLR)* (2025)
33. Zhou, J., Wang, W., Li, S., Qu, X., Guo, X., Liu, Y., Tang, W., Lin, X., Zheng, Y.: TopoTTA: Topology-enhanced test-time adaptation for tubular structure segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 24123–24134 (2025)