AGILE: A Multi-task Contrastive Learning Framework with Adversarial Gradient Iterative Learning for Bio-signal Anonymization

Tamonash Bhattacharyya*, Farshad Firouzi*, Amir M. Rahmani†,
Sanaz Rahimi Mousavi‡, Krishnendu Chakrabarty*
* Arizona State University, AZ, USA
† University of California, Irvine, CA, USA
‡ California State University, Dominguez Hills, CA, USA

***Abstract*—Bio-signals have become a de-facto method for identity verification and subject re-identification, while providing crucial clinically relevant information for telediagnosis, thereby raising significant privacy concerns. Existing anonymization methods often operate on limited modalities leaving parts of the spatio-temporal signal space exposed to re-identification attacks while degrading signal fidelity through indiscriminate noise. Thus, we proposes a multi-task contrastive learning framework that jointly suppresses biometric features while preserving clinically relevant characteristics. The framework iteratively perturbs the signal using a novel adversarial fast gradient sign method (A-FGSM) for targeted noise injection that maximizes identity loss while minimizing diagnostic loss. Evaluated on PTB, CODE-15%, and MIMIC-IV-ECG datasets, our method reduces biometric identification to 15.74% while maintaining a clinical classification accuracy of 94.7%, establishing a new benchmark for bio-signal anonymization.**

***Index Terms*—Bio-signal Anonymization, ECG, Healthcare Systems, Telediagnosis**

I. Introduction

Healthcare has been significantly transformed by deep learning, enabling automated disease diagnosis, enhanced risk stratification, and the rapid expansion of telemedicine. These advances have also driven major progress in signal-space estimation tasks such as semantic segmentation, classification, and anonymization [1]. This is aided by the increasing availability of remotely acquired physiological bio-signals, crucial for early diagnosis and predictive health analytics. However, bio-signals like electrocardiograms (ECGs) contain both diagnostic information and sensitive biometric traits such as age, sex, and individual-specific patterns. This duality creates privacy and security concerns for data sharing and distributed computation in clinical settings. Thus, anonymizing bio-signals by suppressing biometric identifiers while retaining diagnostic utility [2] is crucial for ethical AI-driven healthcare.

Recent efforts in ECG anonymization have explored privacy-preserving techniques such as differential privacy and federated learning [3]. However, these approaches often introduce indiscriminate noise, degrading signal fidelity and compromising clinical utility. Moreover, they do not effectively prevent identity leakage from raw signals, which can still be used to extract biometric templates [4]. These templates capture a holistic and identity-specific representation of the signal, unlike prior methods that focus on protecting isolated demographic attributes (modalities) such as age or sex. GAN-based anonymization methods [5] typically target specific attributes and fail to generalize across more complex identity representations. Additionally, signal-space perturbations such as Gaussian noise can distort clinically relevant regions, limiting diagnostic interpretability. To address these challenges, we propose AGILE (**A**dversarial **G**radient **I**terative **L**earning for **E**chocardiograms), an anonymization framework that operates directly on the raw signal space. The core idea behind AGILE is to perform targeted noise injection using a contrastive, gradient-sensitive adversarial strategy that distinguishes between biometric and clinically relevant signal components. It leverages a convolution-transformer hybrid (CTH) architecture, trained using a multi-task contrastive learning (MTCL) framework, which jointly models local morphological patterns and long-range dependencies. The proposed method adheres to the AAMI EC57 standard [6]for ECG signal classification ensuring clinical relevance and reusability. In summary, this paper makes the following contributions:

- We propose a novel bio-signal anonymization method called AGILE. Contrary to existing methods that operate over a singular modality for anonymization, this method operates on the raw signal space that learns contrastive distributions over the signal space for biometric and characteristic regions-of-interest and adversarially injects noise for anonymization.
- We design a unified CTH backbone that jointly learns shared representations through multi-task learning, leveraging task-specific heads to distinguish characteristic and biometric features.
- We validate the proposed approach on multiple ECG datasets, achieving a clinical classification accuracy of 94.8%, thereby demonstrating generalizability and robustness across varying data distributions. Additionally, our method attains an equal error rate (EER) of 0.125, indicating improved anonymization performance compared to baseline methods.

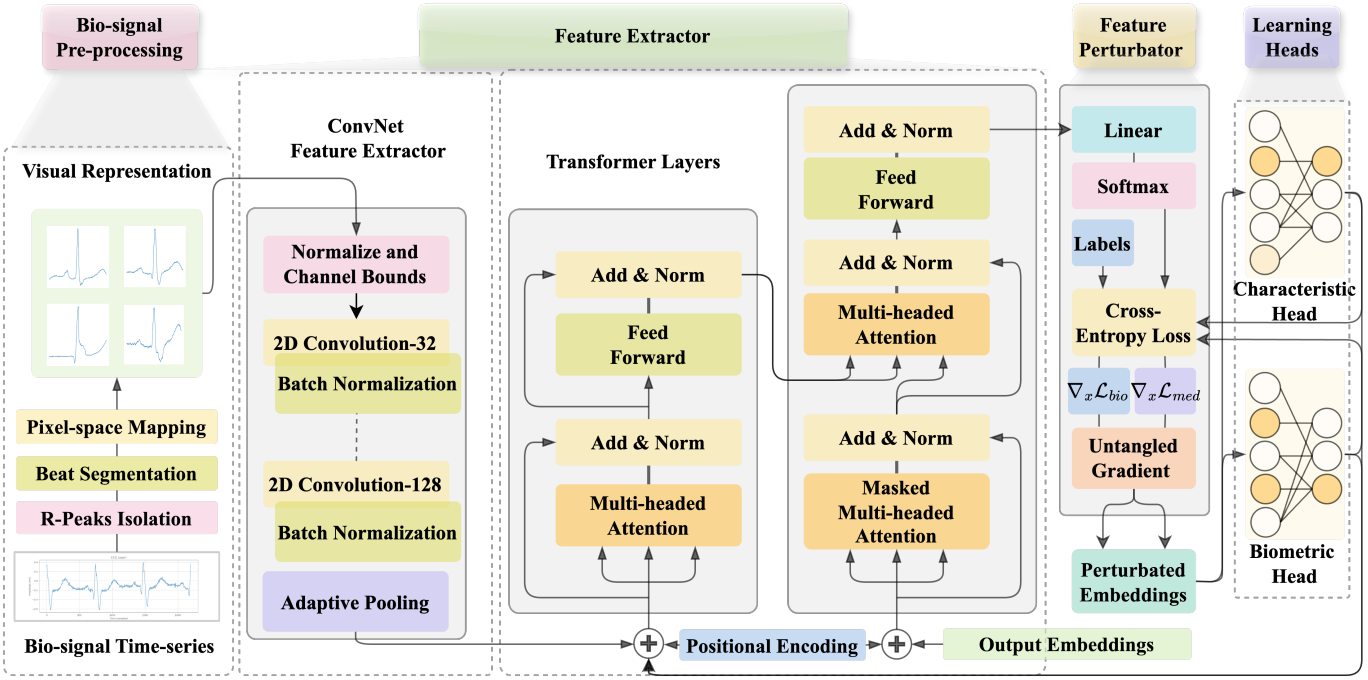

Fig. 1: System architecture of the AGILE framework with A-FGSM noise injection.

II. Proposed Methodology

A. Network Architecture

The architecture of the proposed AGILE framework is composed of four major modules as shown in Fig. 1:

1) Bio-signal Preprocessing: Our model operates on single-beat visual representations of ECG signals, extracted from single-lead records using the Pan-Tompkins algorithm [7] for R-peak detection. Around each R-peak, a fixed-length window is segmented to capture key morphological components (P, QRS, T waves), and converted into standardized 200×200 grayscale images.

2) Feature Extractor Γ*:* In feature extractor, we implement a CTH encoder to map input signals to a high-dimensional latent representation. The convolution layers extract localized spatial patterns in the signal, while the transformer layers model long-range dependencies to construct a global representation over the extracted features. The transformations of Γ can be formalized as:

$$\begin{aligned} c_i &= \sigma(BN(Conv(c_{i-1}))) \\ H_0 &= \text{Flatten}(\text{Adaptive}(c_n))w_p + b_p \\ H_1 &= \text{LN}\left(\text{Softmax}\left(\frac{H_0 w_j \cdot w_k^T H_0^T}{H_0 W_f}\right)\right) \\ F_1 &= \text{LN}\ (\text{FFN}(H_1) + H_1) \end{aligned} \tag{1}$$

where the c_{i-1} is the input to the current convolution block from the previous convolution block (c_0 is the input from the pre-processing block to the first convolution layer and $i \in [0, n]$), BN is batch normaization and σ is the sigmoid activation function. LN and FFN corresponds to layer normalization and feed-forward network operations respectively, corresponding to the various transformations.

3) Feature Perturbator $\aleph$*:* We introduce a novel adversarial fast gradient sign method (A-FGSM), based on the standard FGSM. A-FGSM introduces targeted perturbations by injecting noise aligned with the gradient of the biometric classification loss, thereby maximizing identity obfuscation. Simultaneously, it constrains perturbations along the gradient direction of the characteristic classification loss to preserve diagnostic fidelity. Noise injection is performed at a pixel-level granularity which ensures minimal perturbation while ensuring biometric obfuscation:

$$\eta = \epsilon \cdot \text{sign}\left(\nabla_x \mathcal{L}_{bio} - \lambda \nabla_x \mathcal{L}_{char}\right) \tag{2}$$

where, ϵ denotes the perturbation budget , and λ is a trade-off coefficient that balances the opposing objectives of biometric obfuscation and clinical preservation. The term $\nabla_x \mathcal{L}_{bio}$ and $\nabla_x \mathcal{L}_{char}$ represents the gradient of the biometric loss and clinical classification loss respectively.

4) Learning Heads Θ*:* The learning heads Θ are a pair of fully connected layers that estimate over a given distribution, the probability of the biometric and characteristic objectives. For the feature vector $\mathbf{F} \in \mathbb{R}^d$ extracted by Γ, predicted class probabilities over a closed-set of cardiovascular diseases C_m is given by:

$$P(y_m = c \mid \mathbf{F}) = \frac{\exp\left(\mathbf{W}_m^{(c)} \cdot \mathbf{F} + b_m^{(c)}\right)}{\sum_{j=1}^{C_m} \exp\left(\mathbf{W}_m^{(j)} \cdot \mathbf{F} + b_m^{(j)}\right)} \tag{3}$$

Algorithm 1 Adversarial Training for ECG Anonymization

Require: Training set $\xi = \{(X, Y_{\text{med}}, Y_{\text{bio}})\}$.
Require: Untrained modules: Feature extractor Γ, perturbation function $\aleph$, multi-task heads $\Theta = \{M, B\}$. $\triangleright$ M is the clinical head while B refers to the biometric head
Initialize parameters: encoder θ, medical head ϕ, biometric head ψ.
Initialize scaling factors: perturbation strength ϵ, perturbation mix factor α, adversarial weight λ.
for each epoch $e = 1$ to N **do**
  for each mini-batch $(X, Y_{\text{med}}, Y_{\text{bio}}) \sim \xi$ **do**
    $X' \leftarrow \aleph_\epsilon(\Gamma_\theta, M_\phi, B_\psi, X, Y_{\text{med}}, Y_{\text{bio}})$
    $F \leftarrow \Gamma_\theta(X')$
    $\hat{Y}_{\text{med}} \leftarrow M_\phi(F)$
    $L_{\text{med}} \leftarrow -\sum_{c=1}^{C_{\text{med}}} Y_{\text{med}}^{(c)} \log \hat{Y}_{\text{med}}^{(c)}$
    $\aleph_{\text{rev}} \leftarrow -\lambda \cdot \nabla_\aleph L_{\text{bio}} + (1-\lambda) \cdot \aleph$
    $\hat{Y}_{\text{bio}} \leftarrow B_\psi(\aleph_{\text{rev}})$
    $L_{\text{bio}} \leftarrow -\sum_{c=1}^{C_{\text{bio}}} Y_{\text{bio}}^{(c)} \log \hat{Y}_{\text{bio}}^{(c)}$
    $L \leftarrow L_{\text{med}} - \lambda L_{\text{bio}}$
    Update parameters: $(\theta, \phi, \psi) \leftarrow (\theta, \phi, \psi) - \eta \cdot \nabla L$
  end for
end for

where y_m is the disease label and $W_m^{(c)}$, $b_m^{(c)}$ are the class parameters. Similarly over a closed-set of C_b identity-classes with $W_m^{(c)}$, $b_m^{(c)}$ as the identity class parameters, the probability distribution learned by the biometric-head is given by:

$$P(y_b = c \mid \mathbf{F}) = \frac{\exp\left(\mathbf{W}_b^{(c)} \cdot \mathbf{F} + b_b^{(c)}\right)}{\sum_{j=1}^{C_b} \exp\left(\mathbf{W}_b^{(j)} \cdot \mathbf{F} + b_b^{(j)}\right)} \tag{4}$$

B. Training Framework

During the training phase of AGILE, as shown in Algorithm 1, the MTCL framework enforces learning over a closed-set distribution of multi-class characteristic C_m and biometric labels C_b. The learned distribution of the probabilistic network is approximated over a sequence of FFN learning heads Θ. This constructs the mapping of the gradients for Θ:

$$\begin{aligned} \nabla_x \mathcal{L}_{\text{char}} &= -\nabla_x \log P_{\text{char}}(y_{\text{char}} \mid x; \theta) \\ \nabla_x \mathcal{L}_{\text{bio}} &= -\nabla_x \log P_{\text{bio}}(y_{\text{bio}} \mid x; \theta) \end{aligned} \tag{5}$$

During training, the model minimizes a weighted sum of $\mathcal{L}_{\text{bio}}$ and $\mathcal{L}_{\text{char}}$. The biometric loss is computed as categorical cross-entropy: $\mathcal{L}\text{bio} = -\log P(y_{\text{bio}} \mid \tilde{\mathbf{F}})$, while $\mathcal{L}_{\text{char}}$ is estimated via cross-entropy over predicted and ground truth cardiovascular disease (CVD) labels. The A-FSGM defined in Eq. 2, implements the feedback channel from these losses thereby introducing perturbations in the signal space.

III. Experimental Results

A. Experimental Setup

We train our proposed framework using single-lead ECG, time-series data from four benchmark datasets: the Physikalisch-Technischen Bundesanstalt (PTB) [8], the MIT-BIH Arrhythmia [9], the MIMIC-IV-ECG [10], and the CODE-15% dataset [11]. To asses anonymization effectiveness and utility preservation, we compare AGILE against several SOTA generative anonymization baselines, including **PrivECG-GAN** [12], **REACT** [13], and **ODE-GAN** [14]. These methods typically leverage conditional GAN architectures to synthesize realistic ECG signals while targeting specific modalities for biometric obfuscation. To evaluate the anonymization performance during inference, we assess both privacy preservation and diagnostic fidelity using task-specific networks trained independently of the anonymization pipeline. Privacy is measured using a simulated adversarial identification model by computing the Kullback–Leibler divergence between identity distributions before and after anonymization while we assess characteristic preservation by employing a diagnostic classifier trained to predict clinical labels from the anonymized signals:

$$\begin{aligned} \mathcal{M}_{\text{privacy}} &= \mathbb{E}_{x\sim\xi}\left[D_{\text{KL}}\left(P_{\text{bio}}(y_b \mid x) \parallel P_{\text{bio}}(y_b \mid \tilde{x})\right)\right] \\ \mathcal{F}_{\text{char}} &= \mathbb{E}_{x\sim\xi}\left[\text{CE}(y_m, P_{\text{char}}(y_m \mid \tilde{x}))\right] \end{aligned} \tag{6}$$

B. Evaluation Metrics

The performance of the AGILE framework is assessed along two primary dimensions: **biometric obfuscation** and **clinical fidelity**, corresponding to the dual objectives of anonymization. In evaluating biometric identifiability, we use: 1) **EER**, which corresponds to the point on the receiver operating characteristic (ROC) curve where the false acceptance rate (FAR) equals the false rejection rate (FRR). An increase in EER after anonymization indicates enhanced privacy preservation by reducing identifiability, 2) **Accuracy** of the biometric classifier for evaluating baselines involving non-convergent or adversarial training dynamics. To assess the clinical fidelity of anonymized signals, we evaluate performance on disease classification tasks. Specifically, we compute the **classification accuracy** and **F1-score** using models trained on the original signal distribution and tested on anonymized data.

C. Performance Evaluation

1) Biometric Evaluation: We evaluate the anonymization strength of AGILE by applying identification networks trained on original ECG signals and testing them on anonymized outputs. As shown in Table I, AGILE reduces the identification accuracy from 95.00% (pre-anonymization) to 15.74% (post-anonymization), indicating a significant drop in subject traceability. Here we present performance metrics across varying sample sizes (e.g., 20, 40, 60) and subject categories, including both healthy controls (samples from individuals without CVDs) and mixed samples (individuals with varying CVD classes). This evaluation offers a more comprehensive and generalizable assessment of the framework's anonymization effectiveness. These results are further validated through comparisons with SOTA GAN-based anonymization baselines, as shown in Table II. Each baseline targets a different attribute or modality for identity inference (e.g., age, sex) serving as a

TABLE I: Attacker accuracy measure over pre- and post- anonymization samples using SOTA biometric systems. The siamese twin network relies on a CNN-based approach hence the model performance is unreliable at lower sample sizes.

	Model	**Healthy Control**				**Mixture Sample**			
		20	**40**	**60**	**80**	**50**	**100**	**150**	**268**
Un-anonymized Samples	**SVD [4]**	0.9545	0.9323	0.9310	0.9267	0.9565	0.9189	0.9001	0.5268
	Siamese Twin*et al.* [15]	-	-	-	-	-	-	-	0.9500
Anonymized Samples	**SVD [4]**	0.2333	0.2134	0.2465	0.9323	0.2500	0.1974	0.2091	0.1765
	Siamese Twin [15]	-	-	-	-	-	-	-	0.1574

TABLE II: Comparison of characteristic and identification performance on the PTB dataset for various SOTA models.

Model	**Characteristic Analysis**		**Identification (Biometrics)**		
	Accuracy	**F1 Score**	**Accuracy**	**EER**	**Modality/Attribute**
Priv-ECG GAN [12]	0.8850	0.8425	0.5290	0.098 ± 0.005	Sex-prediction
ODE-based GAN [14]	0.9608	–	0.5403	–	Age
REACT [13]	0.8946	–	0.4125	Non-convergent	Age and Sex
AGILE	0.9480	0.9672	0.1574	0.125	Template

proxy for re-identification risk. In contrast, AGILE evaluates identifiability based on a biometric template, approximatiing identity leakage in a real-world scenario. Despite this, AGILE consistently outperforms all baselines across evaluation scenarios.

2) Clinical Evaluation: As shown in Table II, to validate clinical utility, we measure the performance of diagnostic classification for the models trained on original data and evaluated on anonymized signals. This assesses the degree to which clinically salient features are retained post-anonymization. AGILE achieves a classification accuracy of 94.80% across a diverse sample set spanning four ECG datasets, demonstrating strong generalizability and robustness. Compared to original classification accuracy of 97.40% on non-anonymized data, samples anonymized through AGILE show minimal drop in accuracy to 94.8%. This highlights AGILE's ability to preserve clinical fidelity while effectively obfuscating re-identifiable components of the signal space.

IV. Conclusion

In this work, we proposed a methodology for anonymizing bio-signals, with a focus on ECG-specific anonymization. Unlike prior approaches that target isolated attributes such as age or sex, our framework obfuscates the underlying biometric template, which captures a comprehensive and identity-rich representation of the signal space. We achieve a clinical classification of 94.7% over anonymized samples while degrading the biometric identification accuracy to 15.74%. The method demonstrates strong generalization across diverse subjects and cardiac conditions, offering a robust foundation for privacy-preserving biomedical analytics.

References

[1] K. Packhäuser, S. Gündel, F. Thamm, F. Denzinger, and A. Maier, "Deep learning-based anonymization of chest radiographs: a utility-preserving measure for patient privacy," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 262–272, Springer, 2023.

[2] R. Lazzeretti, J. Guajardo, and M. Barni, "Privacy preserving ecg quality evaluation," in *Proceedings of the on Multimedia and Security*, pp. 165–174, 2012.

[3] Y. Zhao and J. Chen, "A survey on differential privacy for unstructured data content," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–28, 2022.

[4] P. Melzi, R. Tolosana, and R. Vera-Rodriguez, "Ecg biometric recognition: Review, system proposal, and benchmark evaluation," *IEEE Access*, vol. 11, pp. 15555–15566, 2023.

[5] Y. Kang, G. Yang, H. Eom, S. Han, S. Baek, S. Noh, Y. Shin, and C. Park, "Gan-based patient information hiding for an ecg authentication system," *Biomedical Engineering Letters*, vol. 13, no. 2, pp. 197–207, 2023.

[6] *Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms*.

[7] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230–236, 1985.

[8] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet," 1995.

[9] G. B. Moody and R. G. Mark, "Mit-bih arrhythmia database," *(No Title)*, 1992.

[10] B. Gow, T. Pollard, L. A. Nathanson, A. Johnson, B. Moody, C. Fernandes, N. Greenbaum, J. W. Waks, P. Eslami, T. Carbonati, *et al.*, "Mimic-iv-ecg: Diagnostic electrocardiogram matched subset," *Type: dataset*, vol. 6, pp. 13–14, 2023.

[11] A. H. Ribeiro, G. Paixao, E. M. Lima, M. H. Ribeiro, M. M. Pinto Filho, P. R. Gomes, D. M. Oliveira, W. Meira Jr, T. B. Schon, and A. L. P. Ribeiro, "Code-15%: A large scale annotated dataset of 12-lead ecgs," *Zenodo, Jun*, vol. 9, pp. 10–5281, 2021.

[12] A. Nolin-Lapalme, R. Avram, and H. Julie, "Privecg: Generating private ecg for end-to-end anonymization," in *Machine Learning for Healthcare Conference*, pp. 509–528, PMLR, 2023.

[13] A. Datta, T. Bhattacharyya, E. Khatibi, A. Seth, Z. Wang, S. R. Mousavi, A. M. Rahmani, F. Firouzi, and K. Chakrabarty, "React: Reinforcement learning-based adaptive ecg anonymization and privacy threat mitigation," in *2025 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pp. 1–8, IEEE, 2025.

[14] S. Jafarlou, A. M. Rahmani, N. Dutt, and S. R. Mousavi, "Ecg biosignal deidentification using conditional generative adversarial networks," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1366–1370, IEEE, 2022.

[15] R. D. Labati, E. Muñoz, V. Piuri, R. Sassi, and F. Scotti, "Deep-ecg: Convolutional neural networks for ecg biometric recognition," *Pattern Recognition Letters*, vol. 126, pp. 78–85, 2019.