Semantics-Guided Representation Learning for Long-Tailed ECG Classification

Fengyi Guo, **Ying An***, **Jianxin Wang**
School of Computer Science and Engineering, Central South University, Changsha 410083, China
{guofengyi, anying, jxwang}@csu.edu.cn

Abstract

Deep learning has achieved remarkable success in automated electrocardiogram (ECG) diagnosis. However, the long-tailed distribution of real-world ECG data remains a critical challenge. Existing methods mainly operate at the signal feature level and fail to exploit semantic relationships among diagnostic labels, which limits their ability to effectively recognize rare categories. To address this issue, we propose a Semantics-Guided Representation Learning (SGRL) framework, designed to leverage semantic priors to guide feature learning and thereby correct feature bias under long-tailed distributions. Specifically, we propose a Label Semantic Prompt Adaptation (LSPA) strategy, which introduces ECG-calibrated learnable parameters into label prompts to generate adaptive label semantic embeddings under a semantic-consistency constraint, thereby overcoming the limitations of coarse-grained label prompts in characterizing fine-grained intra-class morphological variations. Subsequently, we propose Semantic-Guided Representation Decorrelation (SGRD) that fuses ECG embeddings with adaptive label semantics and suppresses inter-class semantic correlations to obtain discriminative class semantic anchors, thereby improving tail-class separability. Extensive experiments on multiple public datasets demonstrate the significant superiority of SGRL in long-tailed ECG classification tasks. Our code is available on https://github.com/gfywudi/SGRL.

1 Introduction

In recent years, the rising mortality of cardiovascular diseases has underscored the importance of early intervention and timely treatment. Electrocardiogram (ECG) is widely used in clinical practice for cardiac disease screening and evaluation [Ansari *et al.*, 2017]. However, the interpretation of ECG by physicians may be influenced by factors such as clinical experience and fatigue, which can lead to missed diagnoses and misdiagnoses. With the rapid advancement of deep learning, computer-aided diagnosis systems have emerged to automatically analyze ECG data and assist clinicians in identifying pathological abnormalities, thereby improving diagnostic efficiency and reducing human-related errors. Nevertheless, this technology still faces several challenges.

(1) **Existing long-tailed solutions mainly operate in the signal domain, lacking semantic guidance.** Real-world ECG data exhibit a long-tailed distribution, in which common diseases (head classes) have abundant samples, whereas rare diseases (tail classes) suffer from severe sample scarcity, as shown in Figure 1(a). This imbalance causes models to be dominated by head classes during feature learning, resulting in poor discrimination of tail classes. To mitigate this issue, prior studies have explored data resampling methods [Sharma and Sunkaria, 2020; Xia *et al.*, 2023; Li *et al.*, 2022], class-specific loss reweighting [Hammad *et al.*, 2022; Romdhane and Pr, 2020], and modifications to model architectures and training procedures [Jiang *et al.*, 2024; Chen *et al.*, 2024]. However, these approaches primarily operate within the signal feature space through empirical data-level adjustments, without explicitly addressing the structural organization of the representation space. As a result, when tail-class samples are extremely scarce, purely data-driven learning often fails to produce discriminative representations for rare categories (Figure 1(b)). Therefore, the long-tailed problem cannot be attributed solely to data scarcity or biased optimization objectives. Rather, it fundamentally arises from the structural imbalance of the learned representation space under class-imbalanced distributions. From this perspective, alleviating long-tailed imbalance requires introducing additional structural constraints to guide representation organization. Semantic space offers a promising solution by modeling semantic relationships among classes, offering tail classes additional semantic priors and structural support (Figure 1(c)).

In the vision–language domain, semantic modeling approaches have achieved significant progress. Traditional vision–language models (e.g., CLIP [Radford *et al.*, 2021]) jointly learn representations of natural language and visual features, mapping language and visual concepts to a shared semantic space. However, such semantic embedding models are not directly transferable to the ECG domain, as ECG signals exhibit strong temporal characteristics and are governed by complex physiological mechanisms, resulting in semantic

* The corresponding author.

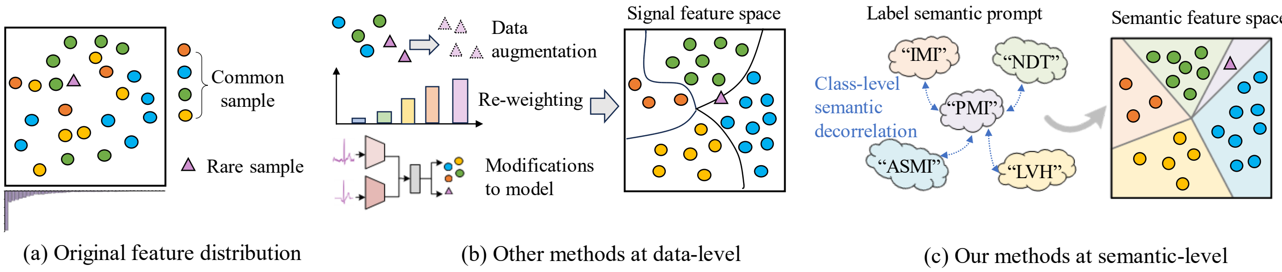

Figure 1: Comparison between data-level long-tail approaches and our semantic-level approach for ECG classification. (a) Original feature distribution of ECG samples under long-tailed label distribution. (b) Data-level long-tail processing methods that operate solely in the signal feature space partially rebalance the sample distribution but remain ineffective in recognizing rare classes. (c) Our method organizes class-specific representations in a semantic space, maintaining stable decision structures under long-tailed distributions.

representations that differ from those of visual image.

(2) **ECG-semantic alignment approaches are constrained by coarse-grained label prompts, failing to characterize the fine-grained ECG features within the same category.** Recent studies have attempted to integrate diagnostic label text and ECG signals into a shared semantic space to enhance representation [Liu *et al.*, 2024; Ye *et al.*, 2025; Seki *et al.*, 2025; Chen *et al.*, 2025]. These approaches typically rely on clinical diagnostic labels as semantic prompts. However, ECG signals within the same label category often exhibit morphological diversity. For instance, although both patients may be diagnosed with left ventricular hypertrophy (LVH), those with long-standing hypertension often present pronounced voltage increases with ST–T strain patterns, whereas patients with aortic stenosis tend to exhibit axis deviation and atypical repolarization abnormalities. In such cases, relying solely on the static textual description "LVH" results in coarse-grained semantic prompts that fail to characterize fine-grained ECG waveform features. This deviation between semantic constraints and ECG representations limits the effectiveness of semantic guidance. Therefore, we argue that constructing a learnable text prompt strategy capable of adaptively calibrating this deviation based on ECG signal features is essential.

To address these challenges, we propose a Semantics-Guided Representation Learning (SGRL) framework. Inspired by the image–text alignment paradigm in CLIP, we first pretrain an ECG encoder to align ECG embeddings with their corresponding label prompts. Subsequently, to transform coarse-grained label prompts into fine-grained semantic representations that better reflect ECG characteristics, we introduce learnable parameters into the label prompts and calibrate them using ECG features to generate adaptive label semantic embeddings. These adaptive semantic embeddings are fused with ECG embeddings via multi-head attention to construct class semantic anchors. We further suppress inter-class semantic correlations to enhance anchor separability, thereby significantly enhancing the recognition performance of individual classes, particularly for tail classes. The contributions of our work can be summarized as follows:

- We propose the SGRL framework, which leverages label semantic priors to guide feature learning and correct feature bias under long-tailed distributions, offering a novel semantics-based solution for long-tailed ECG classification.
- A Label-Semantic Prompt Adaptation (LSPA) strategy is proposed. By incorporating learnable parameters calibrated by ECG into label prompts under a semantic consistency constraint, it generates adaptive label semantic embeddings that better characterize fine-grained ECG features.
- A Semantic-Guided Representation Decorrelation (SGRD) is constructed, which integrates ECG embedding with adaptive label semantic embeddings via multi-head attention, and separates class semantic anchors through label-wise decorrelation, improving discrimination for tail classes.
- Extensive experiments are conducted on multiple public long-tailed datasets and performed comparisons with baseline methods, and the experimental results demonstrate the superiority of the proposed approach.

2 Related Work

Long-Tailed ECG Classification. To alleviate the long-tailed distribution problem in ECG-based cardiac disease classification, some studies have introduced specialized loss functions to mitigate data distribution bias. For example, Hammad et al. [Hammad *et al.*, 2022] and Romdhane et al. [Romdhane and Pr, 2020] employ focal loss to mitigate class imbalance, enabling deep CNN models to emphasize hard and minority samples and achieve improved performance on underrepresented classes. However, these loss-based approaches do not introduce additional supervisory information, making it difficult to learn robust and discriminative representations for extremely rare tail classes. Several studies have employed data generation methods to augment datasets by synthesizing samples of rare classes. Sharma and Sunkaria [Sharma and Sunkaria, 2020] adopt adaptive synthetic oversampling to improve myocardial infarction detection, while Xia et al. [Xia *et al.*, 2023] and Li et al. [Li *et al.*, 2022] utilize GAN-based frameworks to augment ECG data and alleviate long-tailed distributions. Despite their effectiveness, such methods often suffer from overfitting and limited generalization when tail-class samples are severely scarce.

Multimodal ECG-Text Representation Learning. In the field of ECG-based cardiac disease diagnosis, several studies have attempted to incorporate textual prompt information to construct multimodal models, thereby enhancing the performance of ECG-based disease classification. Liu et al. [Liu *et al.*, 2024] align ECG signals with clinical text reports to enable zero-shot diagnosis. Ye et al. [Ye *et al.*, 2025] propose a dual-adapter framework to fuse medical time-series signals and clinical text from complementary perspectives. Seki et al. [Seki *et al.*, 2025] investigate multimodal large language models for ECG understanding in zero-shot visual question answering settings. Chen et al. [Chen *et al.*, 2025] integrate ECG signals with patient static information encoded as textual prompts for hierarchical disease classification. However, the textual prompts adopted in these approaches follow a fixed paradigm. Given the substantial inter-patient variability within the same diagnosis, such fix prompting can limit model generalization.

3 Methodology

In this section, we introduce the proposed SGRL method, whose overall framework is illustrated in Figure 2. First, we pretrain an ECG encoder that can produce semantic-aligned ECG embeddings. Next, we employ LSPA to generate adaptive label semantic embeddings for ECG signals. We then apply SGRD to fuse the ECG embeddings with adaptive label semantic embeddings and separate class semantic anchors. Finally, SGRL integrates the ECG classification logits and anchor-based semantic matching scores for classification, and is jointly optimized using multiple loss functions.

3.1 ECG Encoder Pretraining

During the pretraining stage, we aim to learn an ECG encoder that produces ECG embeddings aligned with the label semantic embedding. Given the ECG embedding e_i of the i-th sample produced by the ECG encoder and the corresponding label semantic embedding s_i generated by the LLM, we adopt the symmetric InfoNCE loss used in CLIP to align the ECG and semantic embeddings:

$$\mathcal{L}_{CLIP} = -\frac{1}{2B}\sum_{i=1}^{B}[\log \frac{e^{{e_i}^\top s_i/\tau}}{\sum_{j=1}^{B} e^{{e_i}^\top s_j/\tau}} + \frac{e^{{s_i}^\top e_i/\tau}}{\sum_{j=1}^{B} e^{{s_i}^\top e_j/\tau}}], \tag{1}$$

where B denotes the batch size, and τ is is the temperature coefficient. The pretrained ECG encoder serves as the backbone network in the subsequent stages, providing initial ECG embeddings for subsequent LSPA and SGRD.

3.2 Label-Semantic Prompt Adaptation

Label Semantic Prompt Construction. We first define an initial label semantic prompts for each category. Unlike traditional approaches that only use the disease name as the label semantic prompt, we augment each category prompt by incorporating synonyms of the corresponding disease. In addition, depending on the specific characteristics of each cardiac disease category, we selectively include the full name or abbreviation, as well as the names of related subtypes or supertypes of cardiac diseases.

We tokenize the initial label semantic prompts to obtain their corresponding initial prompt representations, denoted as $\mathcal{V}^I = [v_1^I, ..., v_C^I], \mathcal{V}^I \in \mathbb{R}^{C\times d}$, where C denotes the number of categories, d denotes the feature dimensionality of each category, and v_i^I represents the initial prompt representation of the i-th category.

Learnable Prompt Parameterization and Calibration. Considering that initial label semantic prompts are insufficient to represent the fine-grained semantic characteristics of ECG signals, we introduce a set of learnable prompt parameters to supplement the semantic information:

$$P = [p_1, ..., p_C], P \in \mathbb{R}^{C\times d'}, \tag{2}$$

where d' denotes the optimization dimension for each category, and p_i represents the learnable parameter associated with the i-th category.

We sample a matrix $\Theta \in \mathbb{R}^{C\times d'}$ from a standard normal distribution $\Theta \sim \mathcal{N}(0, \sigma^2\mathcal{I})$, where σ^2 is set to 1 in this study and $\mathcal{I}$ is the identity matrix. We then map the matrix Θ into the vocabulary index space to obtain the initialization of the learnable prompt parameter matrix P:

$$P \leftarrow \lfloor \Theta \cdot (V-1) \rceil, \tag{3}$$

where V denotes the vocabulary size, $\lfloor \rceil$ denotes rounding to the nearest integer.

To ensure that the learnable prompt parameters can be effectively calibrated based on the ECG signals, we employ a lightweight MLP (Multilayer Perceptron) to map the ECG embedding z_0 extracted by the pretrained ECG encoder into a semantic calibration vector:

$$\gamma = \mathrm{Sigmoid}(W_2 \cdot \delta(W_1 . z_0)), \tag{4}$$

where $\delta(\cdot)$ denotes the ReLU activation function, W_1, W_2 are learnable weight matrices. Subsequently, a residual modulation mechanism is applied to adjust the learnable prompt parameters P:

$$\tilde{P} = P \odot (1+\gamma), \tag{5}$$

where $\odot$ denotes the element-wise multiplication with broadcasting, and $\tilde{P} = [\tilde{p}_1, ..., \tilde{p}_C]$ represents the ECG-calibrated learnable prompt parameters.

For each category i, we concatenate the initial prompt representation v_i^I with its ECG-calibrated learnable prompt parameter $\tilde{p}_i$, thereby obtaining the optimized prompt representation:

$$v_i^O = [v_i^I, \tilde{p}_i]. \tag{6}$$

Accordingly, the optimized prompt representation for all categories can be expressed as:

$$\mathcal{V}^O = [v_1^O, ..., v_C^O], \mathcal{V}^O \in \mathbb{R}^{C\times(d+d')}. \tag{7}$$

Semantic Consistency Constraint. To prevent the optimized prompt representation from drifting excessively away from the original medical prior during optimization, we introduce a semantic consistency constraint to stabilize the semantic refinement process. Specifically, we feed the optimized

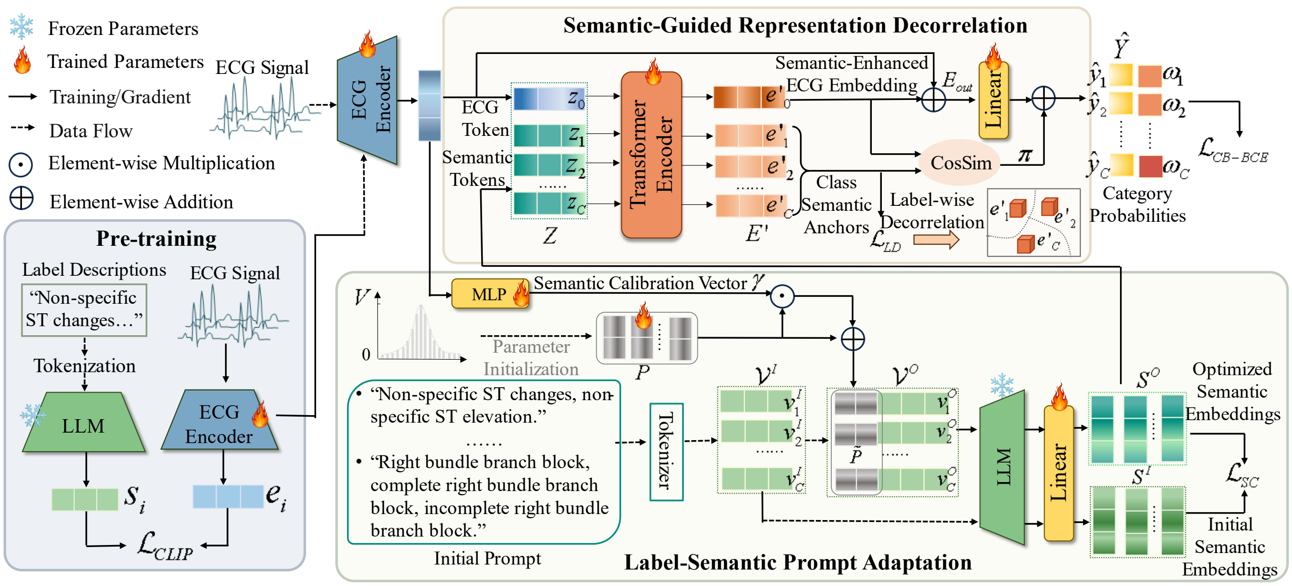

Figure 2: Framework of our proposed SGRL.

prompt representations $\mathcal{V}^O$ and the initial prompt representations $\mathcal{V}^I$ into frozen LLM to produce the corresponding semantic embeddings:

$$\begin{aligned} S^O &= LLM(\mathcal{V}^O), S^O \in \mathbb{R}^{C\times d}, \\ S^I &= LLM(\mathcal{V}^I), S^I \in \mathbb{R}^{C\times d}, \end{aligned} \tag{8}$$

where S^O and S^I denote the optimized and initial semantic embeddings, respectively. Based on these embeddings, we define Semantic Consistency (SC) loss to constrain the deviation between S^O and S^I:

$$\mathcal{L}_{SC} = ||S^O - S^I||_1 = \sum_{i=1}^{C} |s_i^O - s_i^I|, \tag{9}$$

where C denotes the number of labels, and s_i^O, s_i^I represent the optimized and initial semantic embeddings of the i-th category, respectively.

3.3 Semantic-Guided Representation Decorrelation

Based on the optimized semantic embeddings, we adopt SGRD to fuse them with ECG embeddings via multi-head attention, producing a semantic-enhanced ECG embedding and ECG-integrated class semantic anchors. We further introduce Label-wise Decorrelation to reduce semantic overlap among anchors. These class semantic anchors serve as class-matching anchors for computing similarity with the semantic-enhanced ECG embedding.

Specifically, given the optimized semantic embeddings S^O, we treat each class-wise semantic embedding in S^O as a semantic token. Given the ECG embedding z_0 extracted by the pretrained ECG encoder, we treat it as the ECG token. These tokens are concatenated to form a unified sequence representation:

$$Z = \{z_0, ..., z_C\}, Z \in \mathbb{R}^{(1+C)\times d}, \tag{10}$$

where z_0 denotes the original ECG token, and $z_i(i = 1, ..., C)$ denotes the semantic token corresponding to the i-th class. Subsequently we feed Z into a Transformer encoder to facilitate effective interaction between ECG token and semantic tokens:

$$E' = TransEnc(Z) = [e'_0, e'_1, \dots, e'_C], E' \in \mathbb{R}^{(1+C)\times d}, \tag{11}$$

where e'_0 is the semantic-enhanced ECG embedding, and $e'_i(i = 1, 2, ..., C)$ denotes the ECG-integrated class semantic anchor (class semantic anchors for brevity) for the i-th class. By leveraging the self-attention mechanism, each semantic token selectively focuses on label-relevant ECG features, producing class semantic anchors. Meanwhile, the ECG token is enhanced by aggregating semantic information from all class contexts, thereby obtaining the semantic-enhanced ECG embedding.

To mitigate semantic overlap among class semantic anchors e'_i across different classes, we introduce a Label-wise Decorrelation (LD) loss to suppress shared inter-class components and enhance class separability:

$$\mathcal{L}_{LD} = \frac{1}{C'(C'-1)} \sum_{i=1}^{C'} \sum_{j=1, j\neq i}^{C'} Corr(e'_i, e'_j)^2, \tag{12}$$

where C' denotes the number of classes that appear in the current training mini-batch, $e'_i, e'_j \in \mathbb{R}^{B\times d}$ are the class semantic anchors for the i-th and j-th classes, B is the batch size, d is the feature dimension, and $Corr(e'_i, e'_j)$ denotes the Pearson correlation coefficient between e'_i and e'_j. Minimizing $\mathcal{L}_{LD}$ encourages anchors from different classes to be decorrelated, yielding a more discriminative class-specific semantic constraint space for ECG representations.

To identify which the semantic constraint space the semantic-enhanced ECG embedding belongs to, we compute the similarity between the semantic-enhanced ECG embedding e'_0 and the class semantic anchors e'_i, thereby obtaining class-wise matching scores:

$$\pi_i = cos(e'_0, e'_i), i = 1, ..., C, \tag{13}$$

where π_i indicates the degree of matching between the ECG and the i-th diagnostic category in the semantic space.

Meanwhile, to preserve the discriminative information in the original ECG signal and facilitate stable gradient propagation, we employ a residual connection to fuse the original ECG token z_0 with the semantic-enhanced ECG embedding e'_0:

$$E_{out} = e'_0 + z_0, E_{out} \in \mathbb{R}^d, \tag{14}$$

where E_{out} is the ECG output embedding obtained by residual fusion.

3.4 Classification and Training Optimization

To obtain the final multi-label classification results, we integrate the classifier logits derived from the fused ECG embedding E_{out} with the class-wise semantic matching scores:

$$\hat{Y} = Sigmoid(WE_{out} + b + \boldsymbol{\pi}), \tag{15}$$

where W and b are the weight and bias of the classification layer, $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_C]$ is the semantic matching similarity vector, and $\hat{Y}$ denotes the predicted probability over all classes. This integration enables both the ECG feature information and the semantic information from the label matching contribute to the final classification.

We adopt the Class-Balanced Binary Cross-Entropy (CB-BCE) loss to measure the discrepancy between the predicted probabilities and the ground-truth labels in a multi-label setting. Given a training set with N samples and C classes, let $y_{ij} \in \{0, 1\}$ denote the ground-truth label indicating whether the i-th sample belongs to the j-th class, and let $\hat{y}_{ij} \in [0, 1]$ denote the corresponding predicted probability. The CB-BCE loss is defined as:

$$\mathcal{L}_{\text{CB-BCE}} = -\frac{1}{N} \sum_{i,j} \omega_j \left[y_{ij} \log \hat{y}_{ij} + (1 - y_{ij}) \log(1 - \hat{y}_{ij}) \right], \tag{16}$$

where ω_j represents the class-balanced weight associated with the j-th class, which is introduced to enhance the contribution of minority classes during training.

Finally, we combine $\mathcal{L}_{CB-BCE}$ with $\mathcal{L}_{LD}$ and $\mathcal{L}_{SC}$ to jointly optimize the LSPA and SGRD:

$$\mathcal{L} = \mathcal{L}_{CB-BCE} + \alpha\mathcal{L}_{LD} + \beta \cdot \mathcal{L}_{SC} \tag{17}$$

where α and β are hyperparameters that balance the contributions of different loss terms.

4 Experiments

4.1 Experimental Setup

Pre-training Dataset. We pre-train our model on the MIMIC-IV-ECG dataset [Gow *et al.*, 2023], which contains 800,035 paired ECG–report samples from 161,352 subjects. Each sample consists of a 10-second 12-lead ECG signal and its corresponding machine-generated diagnostic report. The original signals are sampled at 500 Hz and uniformly downsampled to 100 Hz for computational efficiency.

Downstream Datasets. PTB-XL [Wagner *et al.*, 2020]: It comprises 21,837 12-lead ECG recordings collected from 18,885 patients. All signals are sampled at 500 Hz with a fixed duration of 10 seconds. In this work, we adopt the finest-grained set of 44 diagnostic labels to formulate a multi-label classification task.

SPH [Liu *et al.*, 2022]: It is a publicly available 12-lead multi-label ECG dataset, consisting of 25,770 recordings from 24,666 patients. The signals are sampled at 500 Hz, with recording durations ranging from 10 to 60 seconds. To ensure a consistent input length, recordings longer than 10 seconds are temporally truncated. The final task is constructed using 44 finest-grained diagnostic categories, resulting in a long-tailed multi-label classification setting.

PhysioNet 2020 [Alday *et al.*, 2020]: It contains 17,497 12-lead ECG recordings, all sampled at 500 Hz with a uniform duration of 10 seconds. We utilize the 26 diagnostic categories provided by the dataset to evaluate model performance on a long-tailed multi-label classification task.

Furthermore, to assess the degree of class imbalance in each dataset, we compute the normalized entropy (NE). A smaller NE value indicates a more skewed class distribution.

Implementation Details. All methods presented in this paper are implemented using the PyTorch 2.6.0 framework and trained and evaluated with Python 3.11. All experiments are conducted on a server equipped with an Intel(R) Xeon(R) Gold 6230 CPU (2.10 GHz), 251 GB of RAM, and four Tesla V100-PCIE-32GB GPUs. To improve the statistical stability of the results, five-fold cross-validation is employed, and the final performance is reported as the average over five runs. During model training, the learning rate is set to 1×10^{-4}, and the batch size is set to 256. The Adam optimizer is used for all experiments. In addition, all ECG signals originally sampled at 500 Hz are downsampled to 100 Hz.

Evaluation Metric. To comprehensively evaluate model performance while accounting for class imbalance, we adopt four metrics: Accuracy (Acc), macro-averaged F1 score (F1), macro-averaged area under the receiver operating characteristic curve (AUC), and Tail-AUC. Specifically, Tail-AUC is computed over tail classes whose sample sizes account for the lowest 1% of the overall class distribution.

4.2 Comparison with Baseline Methods

We compare SGRL with diverse baselines. ASGX, MSDNN, DenseNet, ML-ResNet, MLC-CNN, Autoformer, PatchTST, and FEDformer are conventional ECG classification models. Time-LLM is a time-text fusion method. DeepSMOTE and ADASYN are data augmentation approaches, while ASL is a loss-based reweighting strategy. In addition, ECGAD and ECGFounder are transfer-learning-based methods.

Table 1 summarizes the performance comparison between SGRL and baselines on three datasets. Overall, SGRL demonstrates stable and consistent advantages across long-tailed datasets. On the relatively balanced PhysioNet 2020 dataset, conventional imbalance-handling methods, such as ASL,ADYSN and ECGAD, remain competitive. However, on the more severely imbalanced PTB-XL and SPH datasets,

	PTB-XL, NE=0.7179				SPH, NE=0.5645				PhysioNet 2020, NE=0.8176			
	AUC(Var)	F1(Var)	Acc(Var)	Tail-AUC(Var)	AUC(Var)	F1(Var)	Acc(Var)	Tail-AUC(Var)	AUC(Var)	F1(Var)	Acc(Var)	Tail-AUC(Var)
AGSX [Kalmady *et al.*, 2024]	87.66(0.03)	26.57(0.08)	46.43(0.14)	86.49(0.17)	90.66(0.19)	34.28(0.44)	65.48(0.34)	89.03(0.26)	95.37(0.18)	81.28(0.32)	61.52(0.50)	95.11(0.12)
autoformer [Wu *et al.*, 2021]	86.20(0.36)	28.25(0.29)	43.15(0.32)	84.42(0.13)	90.76(0.11)	31.54(0.14)	58.65(0.29)	89.88(0.18)	88.27(0.13)	42.10(0.39)	21.81(0.21)	87.53(0.11)
MSDNN [Lai *et al.*, 2023]	88.41(0.09)	28.83(0.33)	46.63(0.32)	83.46(0.26)	92.01(0.33)	34.81(0.26)	65.38(0.45)	90.05(0.47)	94.00(0.29)	60.65(0.28)	35.49(0.25)	94.65(0.10)
EffNet [Holmstrom *et al.*, 2024]	80.02(0.63)	14.77(0.87)	42.83(0.95)	79.63(0.48)	78.30(0.47)	10.81(0.69)	45.40(0.54)	78.82(0.32)	90.98(0.28)	73.96(0.57)	56.38(0.76)	92.25(0.22)
Densenet [Xiong *et al.*, 2021]	88.34(0.23)	30.07(0.42)	49.37(0.39)	87.23(0.18)	86.14(0.15)	23.60(0.42)	58.60(0.49)	84.89(0.19)	94.35(0.09)	63.92(0.27)	41.10(0.35)	94.61(0.12)
ML-ResNet [Han and Shi, 2020]	86.75(0.14)	23.12(0.32)	44.65(0.24)	85.54(0.09)	89.29(0.11)	27.72(0.42)	59.51(0.39)	89.13(0.13)	92.61(0.11)	61.39(0.32)	28.91(0.48)	94.36(0.13)
MLC-CNN [Ge *et al.*, 2021]	81.33(0.23)	19.54(0.43)	41.43(0.44)	79.80(0.10)	88.00(0.17)	23.94(0.27)	49.94(0.25)	88.00(0.11)	67.78(0.46)	24.33(0.39)	12.35(0.29)	63.61(0.17)
PatchTST [Nie *et al.*, 2023]	85.30(0.47)	25.83(0.93)	47.94(0.74)	84.09(0.39)	84.79(0.63)	23.74(1.12)	45.69(1.20)	84.50(0.75)	93.82(0.27)	81.16(0.32)	57.46(0.48)	94.60(0.45)
ECGFounder [Li *et al.*, 2025]	88.05(0.11)	30.20(0.24)	46.74(0.23)	86.68(0.04)	85.45(0.13)	22.42(0.24)	46.39(0.07)	85.45(0.24)	95.39(0.32)	75.68(0.43)	54.06(0.48)	95.45(0.12)
ASL [Ridnik *et al.*, 2021]	87.94(0.07)	28.84(0.19)	48.53(0.23)	86.65(0.11)	87.73(0.21)	29.02(0.52)	48.27(0.68)	86.78(0.29)	94.70(0.31)	81.02(0.85)	62.33(0.52)	95.07(0.21)
Time-LLM [Jin *et al.*, 2024]	85.61(0.20)	26.35(0.42)	43.27(0.49)	85.26(0.24)	83.96(0.12)	21.80(0.32)	44.79(0.37)	82.11(0.21)	93.96(0.15)	81.39(0.34)	62.40(0.49)	94.78(0.21)
DeepSMOTE [Dablain *et al.*, 2022]	86.48(0.19)	30.74(0.23)	44.28(0.24)	86.33(0.13)	87.32(0.12)	28.37(0.31)	54.27(0.34)	86.84(0.13)	94.94(0.12)	80.36(0.22)	60.28(0.20)	95.05(0.04)
ADYSN [Sharma and Sunkaria, 2020]	86.83(0.08)	30.38(0.12)	46.68(0.11)	86.61(0.02)	87.19(0.04)	28.72(0.13)	55.72(0.21)	88.43(0.21)	95.06(0.03)	80.51(0.11)	62.10(0.09)	94.91(0.05)
ECGAD [Jiang *et al.*, 2024]	87.32(0.15)	31.39(0.12)	47.28(0.24)	87.37(0.40)	91.91(0.13)	34.98(0.41)	63.32(0.42)	90.26(0.20)	95.63(0.12)	81.04(0.43)	62.27(0.41)	95.59(0.14)
SGRL (ours)	**91.54(0.10)***	**38.27(0.24)***	**50.03(0.25)***	**90.72(0.13)***	**95.53(0.06)***	**44.84(0.13)***	**66.92(0.17)***	**95.32(0.09)***	**96.30(0.10)***	**81.47(0.23)**	**62.45(0.31)**	**96.19(0.12)***

NE stands for Normalized Entropy, with lower values indicating a higher degree of dataset imbalance. **Bold** denotes the optimal results and underline represents the second highest. * denotes a significant improvement is found in the T-test experiment at a significance level of 0.05 compared to all baselines.

Table 1: Model performance comparison.

	PTB-XL	SPH	PhysioNet 2020
w/o LSPA	86.83	93.11	95.16
w/o $\mathcal{L}_{SC}$	88.47	94.97	95.36
w/o $\tilde{P}$	87.06	93.20	95.21
w/o γ	89.31	94.32	95.57
SGRL (ours)	90.72	95.32	96.19

Table 2: Tail-AUC results of the LSPA ablation study.

	PTB-XL	SPH	PhysioNet 2020
w/o SGRD	86.76	93.18	94.68
w/o $\mathcal{L}_{LD}$	90.46	93.74	95.96
w/o $\boldsymbol{\pi}$	89.83	93.38	95.21
SGRL (ours)	90.72	95.32	96.19

Table 3: Tail-AUC results of the SGRD ablation study.

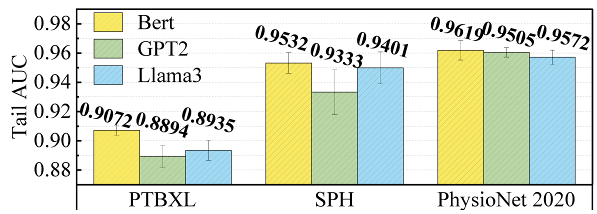

Figure 3: Analysis of different LLMs as text encoders.

SGRL exhibits substantially larger improvements, outperforming the best competing methods by 6.88% and 9.86% in F1 score, respectively. Moreover, SGRL also achieves clear gains in Tail-AUC, exceeding second-best methods by about 3.35% on PTB-XL and 5.09% on SPH. These results indicate that SGRL effectively enhances inter-class separability by imposing semantic-space constraints, enabling robust discrimination across datasets with varying imbalance severities.

4.3 Ablation Studies

Effectiveness of LSPA. To validate the effectiveness of LSPA, we individually remove the semantic consistency loss $\mathcal{L}_{SC}$, the ECG-calibrated learnable prompt parameters $\tilde{P}$, the semantic calibration vector γ, and the entire LSPA , and evaluate their impacts on model performance. As observed from the results in Table 2, all variants exhibit clear performance degradation, confirming that LSPA is critical to SGRL. Among all ablation variants, excluding w/o LSPA, removing $\tilde{P}$ results in the most significant performance degradation. This indicates that the adaptive semantic label prompts provide more informative semantic constraints, which is beneficial for improving the recognition of rare tail classes.

Effectiveness of SGRD. To assess the effectiveness of the SGRD module, we respectively remove the label-wise decorrelation loss $\mathcal{L}_{LD}$, the class-wise matching scores $\boldsymbol{\pi}$, and the entire SGRD. As summarized in Table 3 all variants lead to clear performance drops, confirming the importance of SGRD within SGRL. Among all ablation variants, excluding w/o SGRD, removing $\boldsymbol{\pi}$ lead to most performance degradation, indicating that class-wise matching in the separated semantic space is crucial for effective recognition of tail classes.

4.4 Discussion

Impact of Different ECG Encoders. To evaluate the impact of ECG encoder choices on model performance, we compare five representative architectures: ResNet1d [He *et al.*, 2016], DenseNet [Jahmunah *et al.*, 2023], MS-CNN [Wang *et al.*, 2020], ECGTransformer [El-Ghaish and Eldele, 2024], and RNN [Salloum and Kuo, 2017]. As shown in Table 4, incorporating the SGRL framework consistently improves the performance of all encoders. Among them, MS-CNN achieves the best results, benefiting from its multi-scale convolutional design that captures discriminative ECG features across diverse temporal scales. Accordingly, MS-CNN is adopted as the backbone encoder in all experiments.

Impact of Different LLMs. To investigate the impact of using different LLMs as text encoders on the results, we conduct a comparative analysis of three representative models: BERT, Llama 3, and GPT-2. The experimental results are shown in Figure 3. It can be observed that BERT achieves the best performance on the highly imbalanced PTB-XL and SPH datasets. This advantage can be attributed to its bidirectional attention mechanism, which enables more comprehensive capture of contextual semantic information and provides greater robustness on datasets with skewed class distributions. Accordingly, considering overall performance, BERT is adopted as the text encoder for all experiments.

Impact of Hyperparameter α. To study the impact of the weighting coefficient α for the label-wise decorrelation loss, we vary $alpha$ from 0.6 to 1.5, with results shown in Figure 4(a). The model achieves optimal performance at $\alpha = 0.9$.

	PTB-XL				SPH				PhysioNet 2020			
	AUC	F1	Acc	Tail-AUC	AUC	F1	Acc	Tail-AUC	AUC	F1	Acc	Tail-AUC
MS-CNN+SGRL	**91.54**	**38.27**	**50.03**	**90.72**	**95.53**	**44.84**	**66.92**	**95.32**	**96.30**	**81.47**	**62.45**	**96.19**
MS-CNN	88.90	28.98	46.33	83.59	92.43	34.14	65.56	90.39	94.37	60.76	35.38	94.22
DenseNet+SGRL	89.96	31.80	49.95	89.32	91.19	37.77	65.66	90.69	95.64	80.97	60.43	93.92
DenseNet	87.20	29.42	49.89	85.30	86.62	23.81	58.53	84.19	94.67	63.64	41.18	93.62
Resnet1d+SGRL	87.47	29.44	47.32	86.32	90.42	32.32	62.21	89.03	93.10	78.20	51.93	93.36
Resnet1d	86.71	23.50	44.54	85.28	89.74	27.97	59.72	87.80	92.28	61.39	30.95	92.80
ECGTransFormer+SGRL	88.90	28.84	47.57	87.47	89.45	31.04	54.75	87.06	95.43	79.32	53.47	94.21
ECGTransFormer	83.30	24.87	44.73	84.78	85.89	27.27	52.89	84.23	94.18	63.60	37.23	93.52
RNN+SGRL	87.46	27.14	46.59	86.75	89.01	29.39	54.75	87.06	95.43	79.32	53.47	94.21
RNN	85.40	24.72	45.35	84.27	86.42	25.71	51.66	84.32	94.18	72.87	39.65	93.80

Bold denotes the optimal results. Grey represents the results of training using the SGRL framework.

Table 4: Results of different ECG encoder.

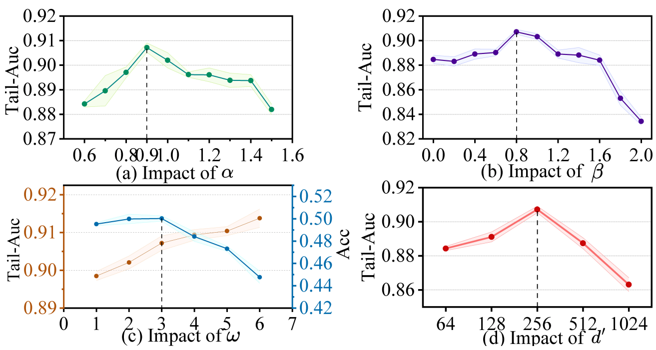

Figure 4: Hyperparameter analysis on PTB-XL.

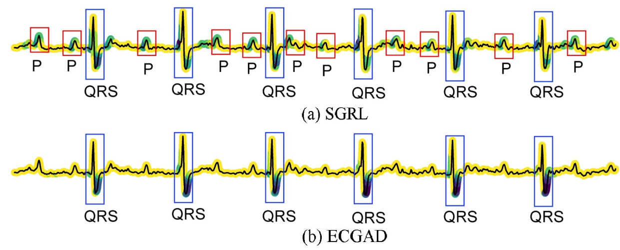

Figure 5: Grad-CAM visualization of Lead I from a patient with 3AVB (tail class) using SGRL and ECGAD. Red and blue boxes denote P waves and QRS complexes focused on by the models.

Smaller values fail to sufficiently decorrelate class semantic anchors, whereas larger values overemphasize the decorrelation constraint and interfere with discriminative feature learning. Thus, $\alpha = 0.9$ provides the best balance between regularization strength and the primary classification objective.

Impact of Hyperparameter β. To investigate the impact of the hyperparameter β on model performance, we vary its value uniformly within $[0, 2]$, with the results shown in Figure 4(b). β controls the strength of semantic consistency during training, with larger values enforcing stronger alignment between the optimized and initial semantics. The model achieves optimal performance at $\beta = 0.5$, indicating that a moderate semantic consistency constraint is most effective, while overly weak or strong constraints lead to performance degradation. Accordingly, β is set to 0.5 in all experiments.

Impact of Hyperparameter ω. To investigate the impact of the class-balanced weight ω on model performance, we vary its value from 1 to 6 , with the results shown in Figure 4(c). As ω increases, the Tail-AUC consistently improves, indicating that assigning larger weights to rare classes is beneficial for enhancing tail-class performance. However, when ω exceeds 3, the overall accuracy drops sharply, as excessive emphasis on rare classes compromises the recognition of head classes and leads to degraded overall performance. Based on this trade-off, we set $\omega = 3$ in all experiments.

Impact of the Learnable Parameter Dimensionality d'. To analyze the impact of the learnable parameter dimensionality d' in the LSPA module, we evaluate four settings: 128, 256, 512, and 1024. As shown in Figure 4(d), the best performance is achieved at $d' = 256$, while both smaller and larger values lead to performance degradation. This indicates that setting the dimensionality to 256 effectively supplements the original prompt information while avoiding the introduction of excessive redundant features. Accordingly, d' is fixed to 256 in all experiments.

Visualization. We employ Grad-CAM [Selvaraju *et al.*, 2017] to visualize the feature weights of ECG signals from a patient with third-degree atrioventricular block (3AVB, a rare class) using SGRL and ECGAD. In this 3AVB case, the P waves and QRS complexes lack a fixed conduction relationship, exhibiting independent rhythmic patterns (known as atrioventricular dissociation). As illustrated in Figure 5, SGRL concurrently focus on abnormal rhythmic features of P waves and QRS complexes. In contrast, ECGAD primarily focuses on the QRS complex, difficult to identify 3AVB.

5 Conclusion

This study proposes SGRL for long-tailed ECG classification. By exploiting the semantic correlations among class labels, SGRL organizes ECG representations into independent semantic spaces while reducing inter-class correlations. Specifically, LSPA is designed to generate patient-adaptive label text prompts, while SGRD decorrelates class semantic anchors. Extensive experiments conducted on multiple public long-tailed ECG datasets demonstrate that SGRL consistently outperforms existing baselines, validating its effectiveness in improving both overall classification performance and tail-class recognition. Despite its effectiveness, SGRL depends on predefined diagnostic label categories, whose granularity and annotation quality may affect semantic modeling. In future work, we plan to incorporate self-supervised learning to reduce label dependence and improve the robustness.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U25A20449, 62372476). This work was also carried out in part using computing resources at the High Performance Computing Center of Central South University.

References

[Alday *et al.*, 2020] Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, et al. Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. *Physiological measurement*, 41(12):124003, 2020.

[Ansari *et al.*, 2017] Sardar Ansari, Negar Farzaneh, Marlena Duda, Kelsey Horan, Hedvig Bille Andersson, Zachary D Goldberger, Brahmajee. K. Nallamothu, and Kayvan Najarian. A review of automated methods for detection of myocardial ischemia and infarction using electrocardiogram and electronic health records. *IEEE Reviews in Biomedical Engineering*, 10:264–298, 2017.

[Chen *et al.*, 2024] Yuanyuan Chen, Xiaoqing Guo, Yong Xia, and Yixuan Yuan. Disentangle then calibrate with gradient guidance: A unified framework for common and rare disease diagnosis. *IEEE Transactions on Medical Imaging*, 43(5):1816–1827, 2024.

[Chen *et al.*, 2025] Qianjiang Chen, Cheng Lian, Bingrong Xu, Quan Zhou, Yixin Su, and Zhigang Zeng. Large language model-assisted multi-scale hierarchical classification of ecg signals. *Knowledge-Based Systems*, 324:113807, 2025.

[Dablain *et al.*, 2022] Damien Dablain, Bartosz Krawczyk, and Nitesh V Chawla. Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE transactions on neural networks and learning systems*, 34(9):6390–6404, 2022.

[El-Ghaish and Eldele, 2024] Hany El-Ghaish and Emadeldeen Eldele. Ecgtransform: Empowering adaptive ecg arrhythmia classification framework with bidirectional transformer. *Biomedical Signal Processing and Control*, 89:105714, 2024.

[Ge *et al.*, 2021] Zhaoyang Ge, Xiaoheng Jiang, Zhuang Tong, Panpan Feng, Bing Zhou, Mingliang Xu, Zongmin Wang, and Yanwei Pang. Multi-label correlation guided feature fusion network for abnormal ecg diagnosis. *Knowledge-Based Systems*, 233:107508, 2021.

[Gow *et al.*, 2023] Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. *Type: dataset*, 2023.

[Hammad *et al.*, 2022] Mohamed Hammad, Monagi H Alkinani, Brij B Gupta, and Ahmed A Abd El-Latif. Myocardial infarction detection based on deep neural network on imbalanced data. *Multimedia Systems*, 28(4):1373–1385, 2022.

[Han and Shi, 2020] Chuang Han and Li Shi. Ml–resnet: A novel network to detect and locate myocardial infarction using 12 leads ecg. *Computer methods and programs in biomedicine*, 185:105138, 2020.

[He *et al.*, 2016] 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*, pages 770–778, 2016.

[Holmstrom *et al.*, 2024] Lauri Holmstrom, Harpriya Chugh, Kotoka Nakamura, Ziana Bhanji, Madison Seifer, Audrey Uy-Evanado, Kyndaron Reinier, David Ouyang, and Sumeet S Chugh. An ecg-based artificial intelligence model for assessment of sudden cardiac death risk. *Communications Medicine*, 4(1):17, 2024.

[Jahmunah *et al.*, 2023] V Jahmunah, Eddie Yin Kwee Ng, Ru-San Tan, Shu Lih Oh, and U Rajendra Acharya. Uncertainty quantification in densenet model using myocardial infarction ecg signals. *Computer Methods and Programs in Biomedicine*, 229:107308, 2023.

[Jiang *et al.*, 2024] Aofan Jiang, Chaoqin Huang, Qing Cao, Yuchen Xu, Zi Zeng, Kang Chen, Ya Zhang, and Yanfeng Wang. Self-supervised anomaly detection pretraining enhances long-tail ecg diagnosis. *arXiv preprint arXiv:2408.17154*, 2024.

[Jin *et al.*, 2024] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, pages 1–24, 2024.

[Kalmady *et al.*, 2024] Sunil Vasu Kalmady, Amir Salimi, Weijie Sun, Nariman Sepehrvand, Yousef Nademi, Kevin Bainey, Justin Ezekowitz, Abram Hindle, Finlay McAlister, Russel Greiner, et al. Development and validation of machine learning algorithms based on electrocardiograms for cardiovascular diagnoses at the population level. *npj Digital Medicine*, 7(1):133, 2024.

[Lai *et al.*, 2023] Jiewei Lai, Huixin Tan, Jinliang Wang, Lei Ji, Jun Guo, Baoshi Han, Yajun Shi, Qianjin Feng, and Wei Yang. Practical intelligent diagnostic algorithm for wearable 12-lead ecg via self-supervised learning on large-scale dataset. *Nature Communications*, 14(1):3741, 2023.

[Li *et al.*, 2022] Wenqiang Li, Yuk Ming Tang, Kai Ming Yu, and Suet To. Slc-gan: An automated myocardial infarction detection model based on generative adversarial networks and convolutional neural networks with single-lead electrocardiogram synthesis. *Information Sciences*, 589:738–750, 2022.

[Li *et al.*, 2025] Jun Li, Aaron D Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7), 2025.

[Liu *et al.*, 2022] Hui Liu, Dan Chen, Da Chen, Xiyu Zhang, Huijie Li, Lipan Bian, Minglei Shu, and Yinglong Wang. A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. *Scientific data*, 9(1):272, 2022.

[Liu *et al.*, 2024] Che Liu, Zhongwei Wan, Ouyang Cheng, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. In *International Conference on Machine Learning*, pages 1–13, 2024.

[Nie *et al.*, 2023] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, pages 1–24, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021.

[Ridnik *et al.*, 2021] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 82–91, 2021.

[Romdhane and Pr, 2020] Taissir Fekih Romdhane and Mohamed Atri Pr. Electrocardiogram heartbeat classification based on a deep convolutional neural network and focal loss. *Computers in Biology and Medicine*, 123:103866, 2020.

[Salloum and Kuo, 2017] Ronald Salloum and C-C Jay Kuo. Ecg-based biometrics using recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2062–2066, 2017.

[Seki *et al.*, 2025] Tomohisa Seki, Yoshimasa Kawazoe, Hiromasa Ito, Yu Akagi, Toru Takiguchi, and Kazuhiko Ohe. Assessing the performance of zero-shot visual question answering in multimodal large language models for 12-lead ecg image interpretation. *Frontiers in cardiovascular medicine*, 12:1458289, 2025.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Sharma and Sunkaria, 2020] LD Sharma and RK Sunkaria. Myocardial infarction detection and localization using optimal features based lead specific approach. *Irbm*, 41(1):58–70, 2020.

[Wagner *et al.*, 2020] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

[Wang *et al.*, 2020] Ruxin Wang, Jianping Fan, and Ye Li. Deep multi-scale fusion neural network for multi-class arrhythmia detection. *IEEE journal of biomedical and health informatics*, 24(9):2461–2472, 2020.

[Wu *et al.*, 2021] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

[Xia *et al.*, 2023] Yi Xia, Yangyang Xu, Peng Chen, Jun Zhang, and Yongliang Zhang. Generative adversarial network with transformer generator for boosting ecg classification. *Biomedical Signal Processing and Control*, 80:104276, 2023.

[Xiong *et al.*, 2021] Peng Xiong, Yanping Xue, Jieshuo Zhang, Ming Liu, Haiman Du, Hong Zhang, Zengguang Hou, Hongrui Wang, and Xiuling Liu. Localization of myocardial infarction with multi-lead ecg based on densenet. *Computer Methods and Programs in Biomedicine*, 203:106024, 2021.

[Ye *et al.*, 2025] Jiexia Ye, Weiqi Zhang, Ziyue Li, Jia Li, Meng Zhao, and Fugee Tsung. Medualtime: A dual-adapter language model for medical time series-text multimodal learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 7913–7921, 2025.