ECG Semantic Integrator (ESI): A Foundation ECG Model Pretrained with LLM-Enhanced Cardiological Text

Anonymous authors
Paper under double-blind review

Abstract

The utilization of deep learning on electrocardiogram (ECG) analysis has brought the advanced accuracy and efficiency of cardiac healthcare diagnostics. In this work, we address a critical challenge in the field of ECG analysis with deep learning: learning robust representation without large-scale labeled datasets. We propose ECG Semantic Integrator (ESI), a novel multimodal contrastive pretraining framework that jointly learns from ECG signals and associated textual descriptions. ESI employs a dual objective function that comprises a contrastive loss and a captioning loss to develop representations of ECG data. To create a sufficiently large and diverse training dataset, we develop a retrieval-augmented generation (RAG)-based Large Language Model (LLM) pipeline, called Cardio Query Assistant (CQA). This pipeline is designed to generate detailed textual descriptions for ECGs from diverse databases. The generated text includes information about demographics and waveform patterns. This approach enables us to compile a large-scale multimodal dataset with over 660,000 ECG-text pairs for pretraining ESI, which then learns robust and generalizable representations of 12-lead ECG. We validate our approach through various downstream tasks, including arrhythmia detection and ECG-based subject identification. Our experimental results demonstrate substantial improvements over strong baselines in these tasks. These baselines encompass supervised and self-supervised learning methods, as well as prior multimodal pretraining approaches. Our work shows the potential of combining multimodal pretraining to improve the analysis of ECG signals.

1 Introduction

The electrocardiogram (ECG), which provides a non-invasive and comprehensive view of the heart's electrical activity, is an important tool in cardiovascular diagnostics and clinical decision-making(Kligfield et al., 2007). For example, ECG has been extensively used in various clinical scenarios, such as diagnosing cardiovascular diseases (Jain et al., 2014), obstructive sleep apnea (Faust et al., 2016), and Parkinson's disease (Haapaniemi et al., 2001), etc. On the other hand, the rapid development of deep learning has triggered general interest in ECG data analysis using data-driven approaches. These deep learning methods, recognized for their ability to learn complex representations, have been proven highly effective in enhancing the accuracy and predictive capability of ECG analysis (Hannun et al., 2019). Typically, the initial step in utilizing ECG signals involves extracting features from the raw data, either through conventional feature engineering or more recent deep learning backbones, such as 1D convolutional neural network (CNN) Zhu et al. (2020); Baloglu et al. (2019); Jing et al. (2021) and Transformer models Meng et al. (2022); Behinaein et al. (2021); Natarajan et al. (2020); Guan et al. (2021); Yan et al. (2019). However, these supervised methods often require large-scale and high-quality annotated training samples, which are costly to obtain (Mincholé & Rodriguez, 2019).

To reduce the reliance on extensive annotations, researchers have explored self-supervised learning (SSL) techniques for ECG signals (Eldele et al., 2021; Kiyasseh et al., 2021; Yu et al., 2023; Gopal et al., 2021). These methods utilize the unlabeled data when pretraining deep feature extractors. Nevertheless, the SSL strategies, which include tasks such as aligning different signal views or reconstructing masked segments, mainly focus on signals. This means these SSL methods pay attention mainly to the waveform characteristics and neglect the semantic meanings of the signals. Consequently, there is no guarantee that those methods

can effectively learn robust representations during the pretraining phase to enhance the ECG analysis in the downstream tasks.

Other studies leverage multimodal learning approaches and incorporate additional modalities, such as descriptive text, into the pretraining process. This approach has shown excellent pretraining performance by enabling a more nuanced and comprehensive understanding of the data (Radford et al., 2021; Jia et al., 2021; Yu et al., 2022b). Motivated by the success of multimodal pretraining on image-text pairs, researchers have developed similar methods for ECGs paired with other modalities including label text (Li et al., 2024), electronic health records (EHR) (Lalam et al., 2023), and clinical reports (Liu et al., 2024). However, acquiring such modalities in large quantities for ECG can also be costly, as ECG analysis would require expertise-depended semantic information compared to general computer vision tasks. Additionally, the variability in terminology and detail across different ECG datasets and sources causes a challenge when combining multiple data sources for a larger scale of pretraining data.

To address these challenges, we introduce a two-step multimodal contrastive pretraining framework to enhance the representations learned from ECG signals. We propose a retrieval-augmented generation (RAG)-based pipeline, Cardio Query Assistant (CQA), to generate standardized and enriched textual descriptions for ECGs. By leveraging the capability of RAG to retrieve relevant information from ECG textbooks, CQA transforms basic ECG conditions into enhanced text descriptions that include patient demographics and specific waveform characteristics. Further, based on the enriched textual descriptions from CQA, we introduce ECG Semantics Integrator (ESI), a contrastive learning framework with a captioning loss inspired by (Yu et al., 2022b). ESI aligns ECG signals with their corresponding text annotations, which aims to pretrain the encoders for an enhanced semantic understanding of ECG content. Our contributions are summarized as follows:

- We introduce a RAG-based ECG description generation pipeline CQA that constructs descriptive textual context for ECG samples using demographic information and diagnostic conditions.
- We develop an ESI framework with both contrastive and captioning capability in pretraining to train an ECG foundation model on approximately 650,000 12-lead ECG signals.
- Compared to strong baselines including the prior SOTA supervised and SSL methods, our evaluation demonstrates promising performances in arrhythmia detection and ECG-based user identification. For instance, we observe an improvement of 1.6% in AUC scores for diagnosing arrhythmia classes, and a 3.5% improvement in identifying subjects when compared to prior SSL methods.

2 Related Work

2.1 Multimodal Representation Learning

Recent studies have introduced foundation models for integrating image and text, which involve both visual and vision-language pretraining. Models such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), Florence (Yuan et al., 2021), and LiT (Zhai et al., 2022) demonstrate that dual-encoder architectures pretrained with contrastive objectives using image-text pairs, can develop robust representations for both modalities and improve the performances of image classification tasks. Beyond contrastive supervision, research involving encoder-decoder models trained with generative losses, such as (Yu et al., 2022b; Wang et al., 2021; 2022), has shown promising results in vision-language benchmarks, with the visual encoder remaining competitive in image classification tasks. Researchers have also applied these powerful representation learning techniques to medical imaging data, including radiography paired with clinical reports (Liu et al., 2023; You et al., 2023; Wan et al., 2024). However, compared to image-based applications, adapting multimodal pretraining methods for ECG signal processing remains relatively underexplored.

2.2 ECG Diagnosis with Deep Learning

2.2.1 Supervised Methods

Deep learning applications in ECG diagnosis have drawn significant attention (Liu et al., 2021; Pyakillya et al., 2017; Sannino & De Pietro, 2018; Wagner et al., 2020; Śmigiel et al., 2021; Mostafa et al., 2019). For instance, Śmigiel et al. (2021) proposed a CNN model with additional entropy-based features for arrhythmia classification. Their method achieves an AUC score of 0.91 across five classes. Mostafa et al. (2019) conducted a comprehensive review of deep learning applications in ECG analysis for sleep apnea detection. They highlighted the success of models such as CNNs and recurrent neural networks (RNNs), which achieve over 90% accuracy on specialized datasets. Despite the proven effectiveness of these methods, the acquisition of clinical annotations required for these methods is often expensive.

2.2.2 Unimodal Representation Learning in ECG

Given the expensive nature of clinical annotations, there has been a growing interest in pretraining methods designed to reduce reliance on labeled ECG sequences (Sarkar & Etemad, 2020; Mehari & Strodthoff, 2022; Oh et al., 2022). For example, Mehari & Strodthoff (2022) applied well-known SSL frameworks such as SimCLR (Chen et al., 2020), BYOL (Grill et al., 2020), and CPC (Oord et al., 2018) to pretrain models on 12-lead ECG data. These models showed enhanced robustness, reflected in a 2% increase in AUC score for 5-class arrhythmia classification compared to purely supervised models. However, even though pretraining strategies generally provide insights into performance improvement and decrease reliance on labeled data, these methods are often limited by their focus solely on signal waveforms. The emphasis on waveform alone does not ensure the capture of clinically relevant semantic information. As a result, a multimodal approach incorporating both ECG waveforms and corresponding clinical text helps acquire more meaningful and transferable ECG representations for various downstream tasks.

2.2.3 Multimodal Representation Learning in ECG

Although few, some studies have begun to explore the alignment of ECG signals with other modalities such as textual descriptions, EHR, and clinical notes (Li et al., 2023; Lalam et al., 2023; Liu et al., 2024). For instance, Lalam et al. (2023) utilized identical encoders to extract and contrastively align embeddings from ECG, EHR, and clinical notes. The pretrained model showed promising results in clinical diagnosis. Liu et al. (2024) adopted a similar approach to couple ECG signals with clinical notes and enhanced the ECG encoder's effectiveness in zero-shot arrhythmia detection. However, the pretraining processes of these methods depend on costly annotations such as clinical notes and EHR, which are challenging to acquire on a large scale. Moreover, the variability in textual descriptions or reports associated with ECGs due to differences in detail, terminology, and style among clinicians and clinical contexts may complicate the learning of consistent mappings between ECG signals and text. The consequent variability could lead to misalignments between the ECG-text pairs. In this study, we propose to utilize a retrieval-augmented generation (RAG)-based pipeline to construct contextual ECG textual data without relying on costly notes and EHRs. Additionally, we introduce a captioning task in our model to achieve more nuanced representations.

3 Methods

In this section, we introduce our approach in three main components: (1) RAG-based ECG description pipeline, Cardio Query Assistant (CQA) and (2) the contrastive captioning pretraining framework, ECG Semantics Integrator (ESI).

3.1 Cardio Query Assistant (CQA) Framework

The Cardio Query Assistant (CQA) Framework, as shown in Figure 1, is designed to transform ECG condition labels into detailed descriptive text. The generated text incorporates demographic information, ECG conditions, and enriched waveform details. The developed CQA is outlined as follows:

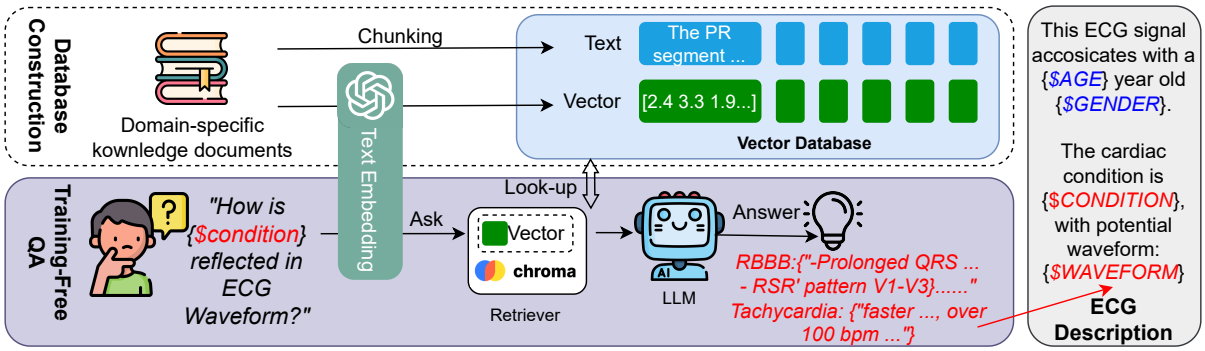

Figure 1: The Cardio Query Assistant (CQA) Framework employs a novel knowledge-based approach to generate detailed and clinically relevant textual descriptions for ECG signals, which translates ECG conditions into enriched ECG waveform patterns.

3.1.1 Establishing a Domain-Specific Knowledge Database

To leverage the enhanced interpretation of ECG conditions with domain expertise, we develop a comprehensive vector database from domain-specific literature of authoritative medical texts guided by two textbooks: (1) *ECG Workout: Exercises In Arrhythmia Interpretation* by Huff (2006), and (2) *12-Lead ECG: The Art of Interpretation* by Garcia (2015). To extract and encode this information into a usable format, we employ the *text-embedding-ada-002* API (OpenAI, 2023) because of its efficiency and performance. The resulting embeddings are then systematically organized using the *Chroma* database management tool, which was chosen for its robustness and ease of integration with the *LangChain* Python library (Mendable, 2023).

3.1.2 Enhancement of ECG Semantics

The CQA Framework enriches the ECG-associated information through a comprehensive retrieval-augmented process. With the pre-constructed domain-knowledge database, the RAG-based approach enables CQA to query related knowledge using given information such as standard clinical labels, standard communications protocol for computer-assisted ECG (SCP) statements, diagnostic interpretations, and machine-generated reports associated with ECG data. For example, the potential lead-detailed waveform descriptions cannot be found in the SCP statements and arrhythmia diagnosis; however, the CQA framework queries this piece of enriched information based on the varying annotations and generates textual descriptions for the potential waveform. Further, this approach encourages the combinative use of ECG data from different sources without relying on labeled data, which means we can combine ECG databases with different annotations as a large-scale dataset for pretraining.

The output of the CQA framework is organized utilizing a Large Language Model (LLM), e.g., GPT-3.5, to generate the potential waveform details from the queried knowledge base. Take an example of the cardiac condition of the Right Bundle Branch Block (RBBB). By executing targeted queries in our database, it retrieves and generates descriptive context for specific waveform attributes. For example, the ECG-associated information of "RBBB" is queried and converted into related waveform features, including "prolonged QRS duration" and "M-shaped RSR' pattern in leads V1-V3."

3.2 Multimodal Contrastive Captioning with ECG Semantics Integrator (ESI) Framework

The ESI framework aims to improve the quality of representation extracted from ECG signals by pretraining a specialized ECG encoder alongside a textual encoder. This dual-modality training method has been proven by cutting-edge studies in contrastive language-image pretraining (CLIP) (Radford et al., 2021) and CoCa (Yu et al., 2022b) methodology.

For the ECG encoder, we have chosen a one-dimensional modified version of the ConvNext v2 architecture considering the sequential nature of ECG waveform data (Woo et al., 2023), which has been proven for its

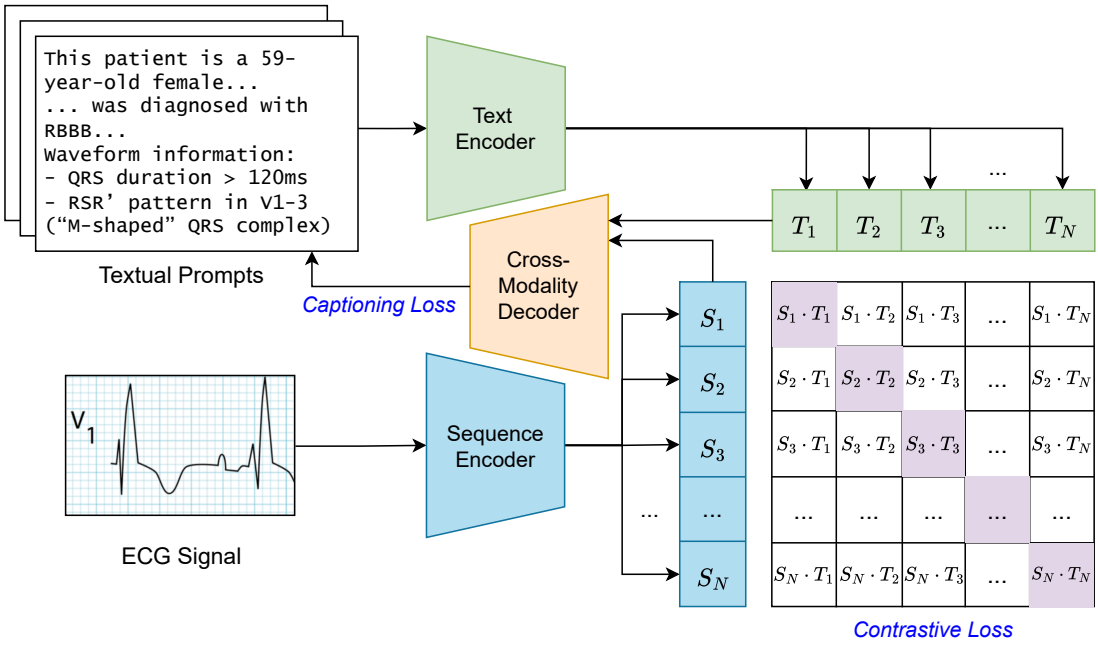

Figure 2: The ECG Semantics Integrator (ESI) is built based on an ECG signal encoder with a text encoder using captioning and contrastive losses for unified representations. This architecture learns from the alignments between detailed textual prompts and the corresponding ECG waveform data, which aims to capture nuanced clinical insights for enhanced diagnostic tasks.

capacities of extracting both local and global contexts with the designed convolutional kernels. In parallel, the textual encoder utilizes BioLinkBERT, a derivative of the BERT architecture pretrained on biomedical texts, to effectively embed medical terminologies (Yasunaga et al., 2022).

3.2.1 Multimodal Contrastive Learning

Inspired by the previous vision language pretraining approaches (Radford et al., 2021; Yu et al., 2022b), our framework uses two pretraining objectives for comprehensive learning, including contrastive loss for robust representation learning and captioning loss for semantic alignment.

Contrastive Loss: We employ the dual-encoder contrastive learning framework following the prior studies. Compared to pretraining with single-encoder as signal-focused frameworks, e.g., SimCLR (Chen et al., 2020) and BYOL (Grill et al., 2020), the dual-encoder approach in this study leverages the semantic information from the textual modality. Both encoders aim to project the inputting ECG and text into a unified embedding space. Consequently, the two encoders are jointly optimized by contrasting the paired text against others in the sampled batch:

$$\mathcal{L}_{\text{Con}} = -\frac{1}{N}\Big(\underbrace{\sum_{i}^{N}\log\frac{\exp\left(S_i^{\top}T_i/\sigma\right)}{\sum_{j=1}^{N}\exp\left(S_i^{\top}T_j/\sigma\right)}}_{\text{ecg-to-text}} + \underbrace{\sum_{i}^{N}\log\frac{\exp\left(T_i^{\top}S_i/\sigma\right)}{\sum_{j=1}^{N}\exp\left(T_i^{\top}S_j/\sigma\right)}}_{\text{text-to-ecg}}\Big),$$

with S_i and T_i representing the normalized embeddings from the ECG signal and text encoders for the i-th ECG-text pair, N is the batch size during training, and σ as the temperature scaling factor. This dual-encoder approach has been working promisingly on enabling cross-modal alignment applications such as zero-shot classification (Radford et al., 2021; Yu et al., 2022b).

Captioning Loss: While the dual-encoder approach encodes the text as an embedding for the contrastive learning purpose; the generative approach aims for detailed granularity and requires the model to predict

the exact tokenized texts with ECG and preceding texts. This approach encourages the encoders to capture the semantic information embedded in the texts actively. Inspired by the image-text multimodal pretraining study CoCa (Yu et al., 2022b), we design to align the generated textual descriptions with the corresponding ECG signals by additionally defining a captioning loss $\mathcal{L}_{cap}$ similar to that used in image captioning tasks (Vinyals et al., 2015):

$$\mathcal{L}_{Cap} = -\sum_{i}^{N} \log P(t_i | t_{<i}, S_i; \theta),$$

where t_i represents the i-th token in the textual description, $t_{<i}$ denotes all the preceding tokens, S_i is the ECG signal, and θ represents the parameters of both encoders and the cross-modality decoder.

The overall pretraining objective is the combination of both contrastive loss and captioning loss, denoted as:

$$\mathcal{L} = \lambda_{Con} \cdot \mathcal{L}_{\text{Con}} + \lambda_{Cap} \cdot \mathcal{L}_{\text{Cap}}$$

where λ_{Con} and λ_{Cap} are the loss weighting hyperparameters for the introduced objectives. We set these two weighting parameters equally to 1 in this study. By jointly optimizing these losses, the ESI Framework aims to learn a multimodal representation that enriches the semantic link between ECG waveforms and their textual explanations. This method is anticipated to improve performances in downstream tasks that leverage the waveform details and demographics, such as diagnosing arrhythmia and performing large-scale patient identification using ECG data.

4 Evaluation

In this section, we describe our evaluation settings and experimental results. We first introduce the information on the datasets used and tasks performed in this study, along with the baseline methods we used in the comparisons. Our experiments explore three settings: zero-shot learning, linear probing (frozen features), and fine-tuning.

4.1 Training Setup

Pretraining Datasets. The proposed ESI signal encoder is pretrained from scratch. Therefore, the pretraining dataset directly impacts model's generalizability. We constructed a large pretraining set combining three large-scale datasets with over 650,000 ECG-text training pairs. These datasets covers Chapman-Shaoxing (Zheng et al., 2020), PTB-XL (Wagner et al., 2020), and MIMIC-ECG (Gow et al., 2023). Each dataset contains 12-lead and 10-second ECG recordings sampled at 500 Hz. Here is a detailed breakdown of each dataset:

- PTB-XL: This dataset consists of 21,837 12-lead, 10-second ECG recordings from 18,885 participants. We followed the training and test data split guidelines outlined in the original publication (Wagner et al., 2020) and only used the training samples (17k) in the pretraining task. These samples include demographic data and SCP codes.
- Chapman-Shaoxing: This dataset offers a larger set of 45k samples with associated demographic information and arrhythmia diagnoses.
- MIMIC-IV-ECG: This is the most extensive dataset with 600k samples accompanied by demographics and machine-generated ECG reports.

The variety and volume of data provide a comprehensive foundation for the pretraining of models. Table 1 summarizes the overview information of each dataset used in pretraining.

Implementation. During the pretraining phase of the ESI model, we make specific choices regarding the encoder architectures, optimizer, learning rate scheduler, training hardware, and batch size. The ECG signal encoder within ESI utilizes a 1D ConvNeXt-base (Woo et al., 2023) backbone as the default architecture. This choice allows the model to effectively capture the spatial features within the ECG signal data. For text

Table 1: Summary of datasets used in the pretraining stage

Dataset	Duration	Sampling Rate	# of Training Samples	Associated Information
PTB-XL	10 seconds	500 Hz	17K	Demographics, SCP Code
Chapman-Shaoxing	10 seconds	500 Hz	45K	Demographics, Arrhythmia Diagnosis
MIMIC-IV-ECG	10 seconds	500 Hz	600K	Demographics, Machine-generated ECG Reports

encoding, we leverage BioLinkBert (Yasunaga et al., 2022) as the default due to its proven capabilities in handling biomedical text data. The AdamW optimizer is employed for optimization during the pretraining process. We opted for an initial learning rate of 5×10^{-5} to facilitate efficient convergence. To further adjust the learning rate throughout training, a warm-up phase of 5 epochs out of the total 30 epochs is implemented. This warm-up phase allows the model to gradually adjust to the training data before applying the main learning rate. Additionally, a learning rate decay of 0.1 is introduced after every 10 epochs to prevent overfitting in the later stages of training.

The pretraining process is conducted on a server equipped with 4 Nvidia A100 GPUs. This hardware configuration provides the computational resources necessary to handle the large datasets used for pretraining efficiently. To leverage the capabilities of these GPUs effectively, a batch size of 48 samples is used on each GPU during training.

In addition to the main ESI model, we implement a parameter-efficient variant named ESI-tiny. This variant utilizes a ConvNeXt-tiny architecture as the ECG encoder, which pretrains a model with a smaller overall size. This can be beneficial in scenarios where computational resources are limited.

4.2 ECG Semantics Integrator (ESI) For Downstream Tasks

After the pretraining stage, the encoder can be applied in three different manners, including zero-shot inference, linear probing, and fine-tuning on various downstream tasks. The aim is to validate the robustness of the learned representations and the practical utility of the model in real-world clinical settings. While the zero-shot inference and linear probing can directly assess the representations of the learned framework, the fine-tuned models usually provide the best performances among these three methods, with updating parameters through downstream tasks. As shown in Figure 3, the fine-tuned encoder from our ESI method outperforms the supervised and self-supervised baselines for different downstream tasks.

4.2.1 Zero-Shot Evaluation

Zero-shot evaluation assesses the model's capacity to understand and infer information from ECG signals without any task-specific fine-tuning. Following the definition used in the context of CoCa (Yu et al., 2022b) and CLIP (Radford et al., 2021), our zero-shot evaluation strategy ensures that while the model has been exposed to a vast array of ECG and text pairs during pretraining, it has not seen any supervised examples from the downstream tasks. In the ESI Framework, each ECG signal's embedding is compared against a range of possible textual labels for different cardiac conditions without task-specific fine-tuning. The model selects the textual description that has the closest embedding distance to the ECG signal's embedding, which is determined by a similarity metric of cosine similarity. This process demonstrates the model's understanding of ECG data and its ability to correlate it with accurate clinical descriptions directly after pretraining, which aims to demonstrate the potential generalizability of the learned representations without fine-tuning the specific tasks.

4.2.2 Linear Probing

Linear probing is a strategy that makes use of the representations learned by the ESI Framework's encoders. In this setup, a linear classifier is utilized on top of the frozen encoders for different downstream tasks such as arrhythmia detection and patient identification. As the only trainable part of the model is the classifier

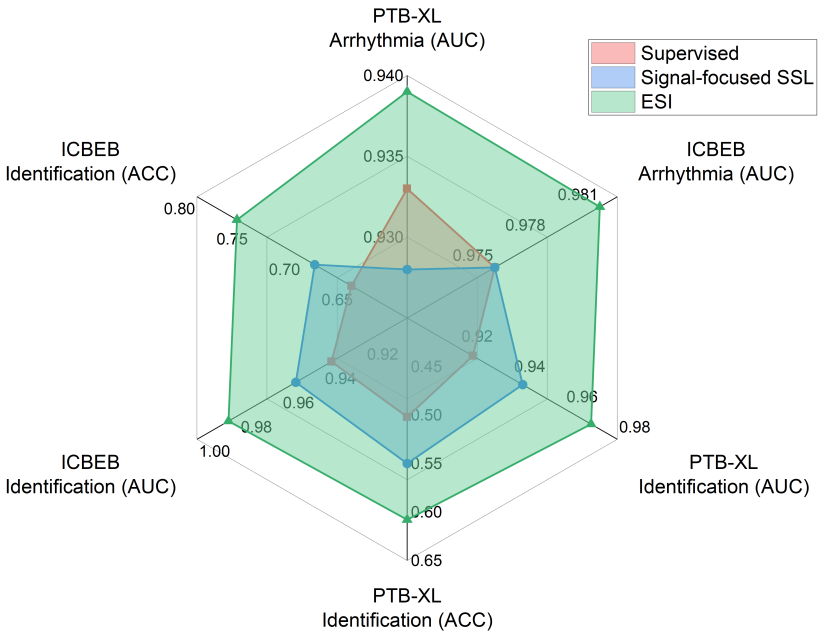

Figure 3: Comparison of the proposed ECG Semantics Integrator (ESI) with the best performances from baseline methods including the supervised models and signal-focused self-supervised learning (SSL) pretrained models. Compared to the baselines, ESI is a multimodal contrastive pretraining framework that leverages both ECG signals and corresponding textual descriptions to learn enhanced ECG representations. The evaluations of arrhythmia diagnosis and identification are conducted on datasets including PTB-XL and ICBEB, with metrics of area under the ROC curve (AUC) and accuracy (ACC).

head, thus, the quality and robustness of the representations from the pretrained encoder play an essential role in the linear probing strategy.

4.2.3 Fine-Tuning

To introduce more flexibility into the pretrained signal encoder, we can also fine-tune the entire framework on a set of downstream tasks. Similar to the linear probing strategy but with trainable encoders, this fine-tuning strategy aims to explore the full effectiveness of the structure with pretrained parameters as its initialization for downstream tasks.

4.3 Downstream Task: Arrhythmia Diagnosis

Cardiac arrhythmias are a significant contributor to cardiovascular diseases, and there is a demand for accurate and reliable detection methods for clinical use. We evaluated our proposed method on arrhythmia detection using two datasets, PTB-XL (Wagner et al., 2020) and ICBEB (Liu et al., 2018). As described in Section 4.1, we follow the training and test data split guidelines outlined in the original PTB-XL publication (Wagner et al., 2020) to divide the PTB-XL dataset. The training set is used for fine-tuning the model and the test set, which is not seen during the pretraining, is used for evaluation. The ICBEB dataset, which is not used during the pretraining, consists of 9,831 12-lead ECG signals from 9,458 patients (Liu et al., 2018). We adopt the processing settings from a prior benchmark study (Strodthoff et al., 2020), which results in 6,877 training samples and 2,954 test samples. Based on this configuration, we evaluate the model's effectiveness across three settings, including fine-tuning, linear probing, and zero-shot learning.

4.3.1 Fine-tuning & Linear Probing

To perform a comprehensive evaluation, we compare our proposed method with specialized supervised methods, including a long short-term memory (LSTM), XResNet101, ResNet50, ensemble methods implemented

Table 2: Evaluation results of arrhythmia diagnosis task under different settings including supervised learning, linear probing (frozen encoder), and fine-tuning. The metric used is area under the ROC curve (AUC). The best results are highlighted in **bold**.

Methods	PTB-XL ($AUC \uparrow$)	ICBEB ($AUC \uparrow$)
(Supervised)		
LSTM (Strodthoff et al., 2020)	0.907	0.964
XResNet101 (Strodthoff et al., 2020)	0.925	0.974
ResNet50 (Strodthoff et al., 2020)	0.919	0.969
Ensemble (Strodthoff et al., 2020)	0.929	0.975
MLBF-Net (Zhang et al., 2021)	0.931	-
MVMSN (Yang et al., 2023)	0.933	-
ConvNeXt-Tiny (Woo et al., 2023)	0.917	0.972
ConvNeXt-Base (Woo et al., 2023)	0.913	0.969
(Linear Probing)		
SimCLR (Chen et al., 2020)	0.764	0.788
BYOL (Grill et al., 2020)	0.771	0.804
CLOCS (Kiyasseh et al., 2021)	0.774	0.799
LEAVES (Yu et al., 2022a)	0.792	0.809
MERL (Liu et al., 2024)	0.887	-
(*Ours*) ESI-tiny	0.928	0.976
(*Ours*) ESI	0.932	0.977
(Fine-tune)		
SimCLR (Chen et al., 2020)	0.919	0.971
BYOL (Grill et al., 2020)	0.924	0.969
CLOCS (Kiyasseh et al., 2021)	0.915	0.975
CRT (Zhang et al., 2022)	0.892	-
LEAVES (Yu et al., 2022a)	0.928	0.975
(*Ours*) ESI-tiny	0.936	0.978
(*Ours*) ESI	**0.939**	**0.981**

in an ECG benchmark study (Strodthoff et al., 2020), a multi-lead-branch fusion network (MLBF-Net) (Zhang et al., 2021), and a multi-view multi-scale neural network (MVMSN) (Yang et al., 2023). Besides the supervised learning methods, we also cover the comparison between our method and signal-focused SSL methods including SimCLR (Chen et al., 2020), BYOL (Grill et al., 2020), CLOCS (Kiyasseh et al., 2021), and LEAVES (Yu et al., 2022a). The deep learning backbone used for training these methods is ConvNeXt-base, the same as the proposed ESI, to ensure fair comparisons. We perform both the linear probing and fine-tuning strategies for those pretrained methods to assess the quality of learned representations during the pretraining phase. Additionally, we benchmark against MERL (Liu et al., 2024) under a frozen encoder setting as presented in their study.

Table 2 shows the performances of the evaluated methods. In the supervised learning methods, the ConvNeXt and ConvNeXt-tiny encoders show lower performance compared to specialized methods such as MLBF-Net and MVMSN. ConvNeXt-tiny outperforms the larger ConvNeXt model potentially due to overfitting with the smaller dataset. After pretraining, the ConvNeXt-based ESI method achieves the best performance on both PTB-XL and ICBEB datasets under both the frozen encoder and fine-tuning settings. Notably, the multimodal pretraining methods (ESI and MERL) significantly outperformed the signal-focused methods such as SimCLR and BYOL. This supports our hypothesis that signal-focused approaches may have limitations in learning robust and transferable representations for downstream tasks compared to the multimodal pretraining methods. The superior performance of the pretrained ESI encoder compared to the frozen encoder in supervised learning approaches also demonstrates the robustness of the learned features for arrhythmia diagnosis.

Table 3: Evaluation results of arrhythmia diagnosis task under different settings in zero-shot learning on PTB-XL dataset. The metric used in this table is area under the ROC curve (AUC) and macro F1 score (F1-macro). $X-\%$ represents the percentage X of the training set used as the training samples in fine-tuning the pretrained encoder. The best results are highlighted in **bold**.

Method	*AUC* ↑	*F1-macro* ↑
SimCLR (Chen et al., 2020) - 5%	0.735	0.547
BYOL (Grill et al., 2020) - 5%	0.752	0.564
CLOCS (Kiyasseh et al., 2021) - 5%	0.765	0.581
LEAVES (Yu et al., 2022a) - 5%	0.760	0.577
METS (Li et al., 2023) - 0%	-	0.593
MERL (Liu et al., 2024) - 0%	0.757	-
(*Ours*) ESI - 0%	**0.812**	**0.654**

4.3.2 Zero-shot Inference

To further assess the learned representations during pretraining, we also evaluate the zero-shot learning inference assessment following METS (Li et al., 2023) and MERL (Liu et al., 2024). Other than the zero-shot evaluations, we include the few-shot setting of the existing signal-focused SSL approaches in the comparison, including SimCLR (Chen et al., 2020), BYOL (Grill et al., 2020), CLOCS (Kiyasseh et al., 2021), and LEAVES (Yu et al., 2022a), with 5% of the original training set as training samples on PTB-XL dataset.

Figure 3 demonstrates the zero-shot learning inference performance of the proposed ESI method alongside baselines (METS, MERL) and the few-shot fine-tuning results for signal-centered SSL methods. Our proposed method achieves the best performance on both AUC and macro F1 scores compared to all other methods. Additionally, ECG-text pretrained models generally outperformed signal-focused pretraining methods even without samples in fine-tuning. This highlights the improved robustness of representations from multimodal pretraining techniques.

4.4 Downstream Task: ECG-based User Identification

ECG generally shows unique patterns that can distinguish individuals, which makes them suitable for subject identification tasks and offers potential advantages over other biometric traits (Melzi et al., 2023). For example, compared to identifying persons with facial images, using ECG can further protect users' privacy during usage. In this study, we leverage the PTB-XL (Wagner et al., 2020) and ICBEB (Liu et al., 2018) datasets, also used in the arrhythmia diagnosis task, to design a one-shot learning benchmark for ECG identification. Similar to the arrhythmia diagnosis task in Section 4.3, we focus solely on the test splits of the PTB-XL and ICBEB datasets. For PTB-XL, we select 5-second signal sequences from each of the 1907 subjects for both training and testing, which results in a 1907-class classification task. Similarly, we select 4-second samples and classes from the ICBEB dataset, which forms a 689-class classification task.

In this task, we compare the performance of the proposed ESI method with various approaches under linear probing and fine-tuning settings. For the specialized supervised methods, we compared our method with established supervised learners, including LSTM, XResNet101, ResNet50, and ensemble methods (Strodthoff et al., 2020). Besides the supervised learning methods, we also evaluated the ESI method against signal-focused SSL approaches including SimCLR (Chen et al., 2020), BYOL (Grill et al., 2020), CLOCS (Kiyasseh et al., 2021), and LEAVES (Yu et al., 2022a) under the linear probing and fine-tuning settings. Due to the patient de-identification during pretraining, we cannot perform zero-shot learning for this identification task, as the model has not been exposed to identifiable patient information.

Table 4 summarizes the findings. The results demonstrate that the ESI method significantly outperforms the baselines in both linear probing and fine-tuning settings. Notably, the fine-tuned ESI method achieves a substantial accuracy improvement (12.0% and 12.2% on PTB-XL and ICBEB, respectively) compared to the supervised ConvNeXt baselines. Additionally, in linear probing, ESI surpasses the signal-focused SSL

Table 4: Evaluation results of 1-shot ECG-based user identification task under settings including supervised learning, linear probing, and fine-tuning. The metric used is area under the ROC curve (AUC) and accuracy score (ACC). The best results are highlighted in **bold**.

Method	PTB-XL		ICBEB	
	AUC ↑	*ACC* ↑	*AUC* ↑	*ACC* ↑
(Supervised)				
LSTM (Strodthoff et al., 2020)	0.905	0.446	0.916	0.608
XResNet101 (Strodthoff et al., 2020)	0.917	0.477	0.922	0.624
ResNet50 (Strodthoff et al., 2020)	0.922	0.493	0.930	0.636
Ensemble (Strodthoff et al., 2020)	0.925	0.502	0.936	0.653
ConvNeXt-Tiny (Woo et al., 2023)	0.919	0.481	0.934	0.647
ConvNeXt-Base (Woo et al., 2023)	0.920	0.488	0.932	0.640
(Linear Probing)				
SimCLR (Chen et al., 2020)	0.802	0.174	0.842	0.254
BYOL (Grill et al., 2020)	0.835	0.236	0.857	0.281
CLOCS (Kiyasseh et al., 2021)	0.814	0.199	0.835	0.237
LEAVES (Yu et al., 2022a)	0.841	0.248	0.855	0.276
(*Ours*) ESI-tiny	0.923	0.511	0.937	0.654
(*Ours*) ESI	0.926	0.518	0.942	0.663
(Fine-tune)				
SimCLR (Chen et al., 2020)	0.934	0.522	0.945	0.673
BYOL (Grill et al., 2020)	0.942	0.547	0.951	0.686
CLOCS (Kiyasseh et al., 2021)	0.929	0.516	0.940	0.666
LEAVES (Yu et al., 2022a)	0.944	0.550	0.953	0.688
(*Ours*) ESI-tiny	0.966	0.591	0.980	0.747
(*Ours*) ESI	**0.970**	**0.608**	**0.985**	**0.762**

Table 5: Ablation experiments. The evaluation performances on the PTB-XL dataset for both arrhythmia diagnosis and identification tasks. The experiments are under the linear probing setting with frozen encoders, and the evaluation metrics used are the AUC scores. In the following tables, the performances from the most contributing components are highlighted in **bold**.

(a) Components of the proposed framework

w/o	Arrhythmia	Identification
-	0.928	0.923
CQA	0.901	0.902
$\mathcal{L}_{\text{Cap}}$	0.913	0.905
$\mathcal{L}_{\text{Con}}$	**0.850**	**0.877**

(b) Selection of the signal encoder

Backbone	Params	Arrhythmia	Identification
ViT-1D	84.92 M	0.897	0.910
XResNet101-1D	1.80 M	0.894	0.901
ConvNeXt-tiny	26.81 M	0.928	0.923
ConvNeXt-base	85.56 M	**0.932**	**0.926**

methods by a significant margin. These findings highlight the effectiveness of multimodal pretraining with ECG and text data in learning transferable representations for ECG identification.

5 Discussion

In this section, we first discuss the ablations on the designed components, including the effectiveness of CQA, the selections of the signal encoder, as well as the contributions from captioning loss and contrastive loss to the learned representations. The experiments for ablations are mostly conducted on a tiny model variant as ESI-tiny with the ConvNeXt-tiny signal backbone. We also explore the impact of the size of pretraining data as well as the potential misalignment to the proposed method.

5.1 Ablation Study: Component Analysis and Impact

We conduct ablation experiments to assess the contributions of individual components within the pretraining framework. We evaluate the pretrained models' performance on both arrhythmia diagnosis and ECG-based user identification tasks using the PTB-XL dataset with the ESI-tiny variant featuring a ConvNeXt-tiny signal encoder. The model is evaluated under a linear probing setting with frozen encoders, and the AUC score serves as the primary metric.

The results of this ablation study are summarized in Table 5(a). Removing the Contrastive Question Answering (CQA) module results in a decrease in AUC scores for both tasks. This indicates that CQA contributes to aligning ECG signals with their enriched text annotations. By aligning these modalities, CQA helps the model learn more robust representations that capture the inherent relationships between ECG patterns and associated diagnoses. Ablating the captioning loss $\mathcal{L}_{\text{Cap}}$ also leads to a performance decline on both arrhythmia diagnosis and ECG identification, which suggests that the model benefits from explicitly generating captions during pretraining. The most substantial impact on performance is observed when removing the contrastive loss $\mathcal{L}_{\text{Con}}$. By contrasting ECG with different textual representations, the model is encouraged to identify subtle variations that are relevant to downstream tasks.

5.2 Ablation Study: Selection of Backbone for Signal Encoder

We investigate the impact of various backbone architectures for the signal encoder on the model's performance. Prior research on vision-text pretraining informed our selection of candidate backbones, including ViT (Dosovitskiy et al., 2020) and ConvNeXt (Woo et al., 2023) architectures. Additionally, considering the strong performance of XResNet1D101 in supervised arrhythmia diagnosis (Strodthoff et al., 2020) as shown in Table 2, we examine its potential as a foundation encoder after multimodal pretraining.

Table 5(b) summarizes the results. ConvNeXt-based models (ConvNeXt-tiny and ConvNeXt-base) achieved the highest AUC scores on both arrhythmia diagnosis and ECG identification tasks. Notably, ConvNeXt-base, the largest model with the most parameters (85.56M), provides the best overall performance (AUC of 0.932 for arrhythmia diagnosis and 0.926 for ECG identification). ConvNeXt-tiny, a more parameter-efficient option (26.81M), shows a slightly lower performance compared to the base version, but also achieves competitive AUC scores with a significantly lower parameter footprint. With the same-level size as ConvNeXt models in both tasks, the ViT model shows substantially lower performances, which might indicate the convolutional kernel could be more suitable in processing ECG signals. Moreover, the XResNet1D101 backbone displays lower AUC scores compared to ConvNeXt architectures. This can due to the significantly smaller parameter size of the XResNet model.

5.3 Ablation Study: Impact of Pretraining Data Sizes and Distribution Shifts

In this ablation study, we also investigate the impact of pretraining dataset size on the performance of our ESI-tiny model in two arrhythmia classification datasets including PTB-XL and ICBEB. We artificially change the size of the unlabeled pretraining dataset from 0 to all available samples by randomly sampling and selecting from all the pretraining samples. After ESI-tiny encoders are trained, the models are then evaluated on both datasets using a linear probing approach with frozen encoder weights, and the AUC score is employed as the performance metric.

The results presented in Figure 4 demonstrate a consistent performance improvement as the pretraining dataset size increases for both PTB-XL and ICBEB datasets. The model's performance improves most substantially in the early stages of increasing the pretraining dataset size. This indicates that a modest amount of pretraining data can help substantial gains in learning informative representations.

Additionally, to understand the effect of pretraining on distribution shift, we measure the Maximum Mean Discrepancy (MMD) between the pretraining dataset and each test set (PTB-XL and ICBEB) as the pretraining size increases. As shown in Figure 5, the distribution shift between the pretraining and test sets decreases as the pretraining dataset size increases. This suggests that larger pretraining datasets help the model learn representations that are more generalizable to the target datasets. The decrease in MMD corre-

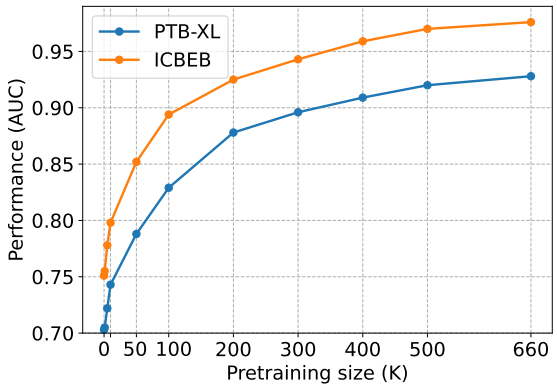

Figure 4: Performances of linear probing inference in arrhythmia diagnosis (AUC) on PTB-XL and ICBEB data using the pretrained encoders with varying training samples.

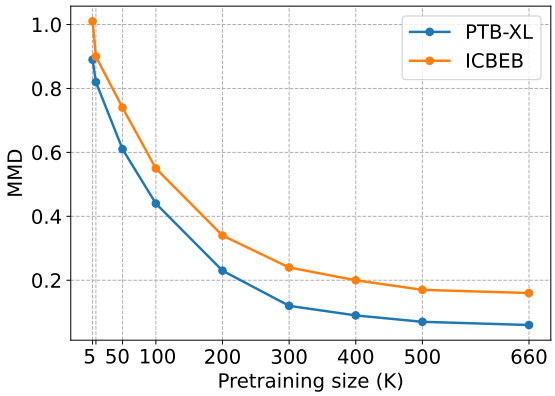

Figure 5: Distribution difference measured by maximum mean discrepancy (MMD) between the pretraining set and test set using encoders with varying training samples.

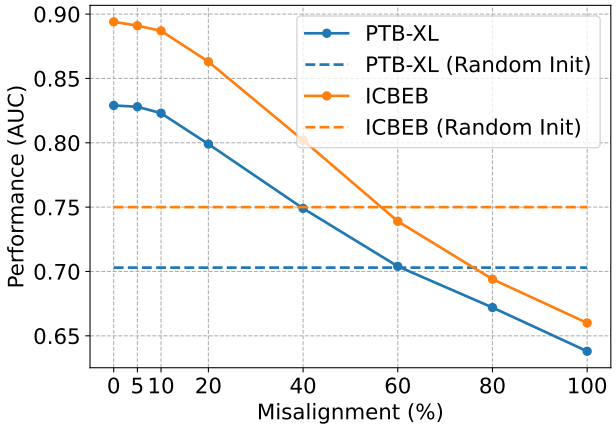

Figure 6: Performances of linear probing inference in arrhythmia diagnosis (AUC) on PTB-XL and ICBEB data using pretrained encoders with varying ECG-text misalignment ratios. Models trained with randomly initialized weights are included as baselines.

lates with the improvements in AUC scores observed in Figure 4, which potentially indicates that reducing distribution shift through larger pretraining sets contributes to better performance on the downstream tasks.

5.4 Ablation Study: Impact of ECG-text Misalignment

To investigate the impact of potential misalignments between ECG signals and their corresponding text descriptions on the proposed ESI method, we conduct an experiment on ESI-tiny with a sub-training set of 100K ECG-text pairs. We introduce various degrees of misalignment by randomly shuffling a percentage of ECG-text pairs in our pretraining dataset. The model was then evaluated on two arrhythmia classification datasets with a linear probing approach with frozen encoder weights.

Figure 6 shows the results of the experiment. As the percentage of misaligned pairs increases, the performance of ESI-tiny decreases substantially on both datasets. This indicates that the alignment between the pretraining sample pairs plays an essential role in learning robust representations. Notably, increasing misalignment can lead to performance that is worse than using un-pretrained models with random initialization, which also highlights the critical importance of accurate ECG-text pairing in pretraining.

6 Conclusion

This study introduces a novel multimodal contrastive pretraining framework to enhance the quality and robustness of representations learned from ECG signals. To address the lack of descriptive text associated with ECGs, we propose a retrieval-augmented generation (RAG) pipeline called the Cardio Query Assistant (CQA). This pipeline generates detailed textual descriptions for ECG data with demographic information, potential conditions, and waveform patterns. Inspired by the success of multimodal pretraining strategies in vision-language tasks, we develop the ECG Semantic Integrator (ESI). This framework integrates both contrastive and captioning capabilities to foster a deeper semantic understanding of ECG signals. Our evaluation validates the effectiveness of the proposed approach across several downstream tasks. The ESI method demonstrates improvement in arrhythmia diagnosis and ECG-based user identification tasks by outperforming strong baselines that cover supervised learning and SSL approaches. These results with the ablation studies highlight the benefits of multimodal learning for ECG analysis and the value of integrating captioning loss with contrastive pretraining. Beyond ECG, we believe the proposed CQA and ESI frameworks hold the potential for applications to other types of biomedical time series data, where contextual information can be leveraged to enhance the representation learning and downstream analysis.

On the other hand, this study is limited by the use of 10-second ECG signals only in pretraining. While our approach demonstrates effectiveness on this data, real-world ECG recordings can vary significantly in length and may contain more diverse features. In future work, we plan to investigate the impact of using more diverse ECG signals on the performance of the proposed framework.

References

Ulas Baran Baloglu, Muhammed Talo, Ozal Yildirim, Ru San Tan, and U Rajendra Acharya. Classification of myocardial infarction with multi-lead ecg signals and deep cnn. *Pattern Recognition Letters*, 122:23–30, 2019.

Behnam Behinaein, Anubhav Bhatti, Dirk Rodenburg, Paul Hungler, and Ali Etemad. A transformer architecture for stress detection from ecg. In *2021 International Symposium on Wearable Computers*, pp. 132–134, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

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

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021.

Oliver Faust, U Rajendra Acharya, EYK Ng, and Hamido Fujita. A review of ecg-based diagnosis support systems for obstructive sleep apnea. *Journal of Mechanics in Medicine and Biology*, 16(01):1640004, 2016.

Tomas B Garcia. *12-lead ECG: The art of interpretation*. Jones & Bartlett Learning, 2015.

Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Machine Learning for Health*, pp. 156–167. PMLR, 2021.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Seth Berkowitz, Dana Moukheiber, Parastou Eslami, et al. Mimic-iv-ecg-diagnostic electrocardiogram matched subset. *Type: dataset*, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Jian Guan, Wenbo Wang, Pengming Feng, Xinxin Wang, and Wenwu Wang. Low-dimensional denoising embedding transformer for ecg classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1285–1289. IEEE, 2021.

TH Haapaniemi, Ville Pursiainen, JT Korpelainen, HV Huikuri, KA Sotaniemi, and VV Myllylä. Ambulatory ecg and analysis of heart rate variability in parkinson's disease. *Journal of neurology, neurosurgery & psychiatry*, 70(3):305–310, 2001.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019.

Jane Huff. *ECG workout: Exercises in arrhythmia interpretation*. Lippincott Williams & Wilkins, 2006.

Rahul Jain, Robin Singh, Sundermurthy Yamini, and Mithilesh K Das. Fragmented ecg as a risk marker in cardiovascular diseases. *Current Cardiology Reviews*, 10(3):277–286, 2014.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Enbiao Jing, Haiyang Zhang, ZhiGang Li, Yazhi Liu, Zhanlin Ji, and Ivan Ganchev. Ecg heartbeat classification based on an improved resnet-18 model. *Computational and Mathematical Methods in Medicine*, 2021, 2021.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pp. 5606–5615. PMLR, 2021.

Paul Kligfield, Leonard S Gettes, James J Bailey, Rory Childers, Barbara J Deal, E William Hancock, Gerard Van Herpen, Jan A Kors, Peter Macfarlane, David M Mirvis, et al. Recommendations for the standardization and interpretation of the electrocardiogram: part i: the electrocardiogram and its technology: a scientific statement from the american heart association electrocardiography and arrhythmias committee, council on clinical cardiology; the american college of cardiology foundation; and the heart rhythm society endorsed by the international society for computerized electrocardiology. *Circulation*, 115(10):1306–1324, 2007.

Sravan Kumar Lalam, Hari Krishna Kunderu, Shayan Ghosh, Harish Kumar, Samir Awasthi, Ashim Prasad, Francisco Lopez-Jimenez, Zachi I Attia, Samuel Asirvatham, Paul Friedman, et al. Ecg representation learning with multi-modal ehr data. *Transactions on Machine Learning Research*, 2023.

Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ecg zero-shot learning. *arXiv preprint arXiv:2303.12311*, 2023.

Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ecg zero-shot learning. In *Medical Imaging with Deep Learning*, pp. 402–415. PMLR, 2024.

Che Liu, Sibo Cheng, Miaojing Shi, Anand Shah, Wenjia Bai, and Rossella Arcucci. Imitate: Clinical prior guided hierarchical vision-language pre-training. *arXiv preprint arXiv:2310.07355*, 2023.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. *arXiv preprint arXiv:2403.06659*, 2024.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

Xinwen Liu, Huan Wang, Zongjin Li, and Lang Qin. Deep learning in ecg diagnosis: A review. *Knowledge-Based Systems*, 227:107187, 2021.

Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ecg data. *Computers in biology and medicine*, 141:105114, 2022.

Pietro Melzi, Ruben Tolosana, and Ruben Vera-Rodriguez. Ecg biometric recognition: Review, system proposal, and benchmark evaluation. *IEEE Access*, 2023.

Mendable. Langchain: A framework for developing applications powered by language models. Software framework, 2023. Retrieved from `https://github.com/mendable/langchain`.

Lingxiao Meng, Wenjun Tan, Jiangang Ma, Ruofei Wang, Xiaoxia Yin, and Yanchun Zhang. Enhancing dynamic ecg heartbeat classification with lightweight transformer model. *Artificial Intelligence in Medicine*, 124:102236, 2022.

Ana Mincholé and Blanca Rodriguez. Artificial intelligence for the electrocardiogram. *Nature medicine*, 25 (1):22–23, 2019.

Sheikh Shanawaz Mostafa, Fábio Mendonça, Antonio G. Ravelo-García, and Fernando Morgado-Dias. A systematic review of detecting sleep apnea using deep learning. *Sensors*, 19(22):4934, 2019.

Annamalai Natarajan, Yale Chang, Sara Mariani, Asif Rahman, Gregory Boverman, Shruti Vij, and Jonathan Rubin. A wide and deep transformer neural network for 12-lead ecg classification. In *2020 Computing in Cardiology*, pp. 1–4. IEEE, 2020.

Jungwoo Oh, Hyunseung Chung, Joon-myoung Kwon, Dong-gyun Hong, and Edward Choi. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In *Conference on Health, Inference, and Learning*, pp. 338–353. PMLR, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenAI. Embedding - openai. Software API, 2023. Retrieved from `https://platform.openai.com/docs/guides/embeddings`.

Boris Pyakillya, Natasha Kazachenko, and Nikolay Mikhailovsky. Deep learning for ecg classification. In *Journal of physics: conference series*, volume 913, pp. 012004. IOP Publishing, 2017.

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*, pp. 8748–8763. PMLR, 2021.

Giovanna Sannino and Giuseppe De Pietro. A deep learning approach for ecg-based heartbeat classification for arrhythmia detection. *Future Generation Computer Systems*, 86:446–455, 2018.

Pritam Sarkar and Ali Etemad. Self-supervised ecg representation learning for emotion recognition. *IEEE Transactions on Affective Computing*, 13(3):1541–1554, 2020.

Sandra Śmigiel, Krzysztof Pałczyński, and Damian Ledziński. Ecg signal classification using deep learning techniques based on the ptb-xl dataset. *Entropy*, 23(9):1121, 2021.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE journal of biomedical and health informatics*, 25(5):1519–1528, 2020.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

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): 154, 2020.

Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *Advances in Neural Information Processing Systems*, 36, 2024.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pp. 23318–23340. PMLR, 2022.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.

Genshen Yan, Shen Liang, Yanchun Zhang, and Fan Liu. Fusing transformer model with temporal features for ecg heartbeat classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 898–905. IEEE, 2019.

Shunxiang Yang, Cheng Lian, Zhigang Zeng, Bingrong Xu, Junbin Zang, and Zhidong Zhang. A multi-view multi-scale neural network for multi-label ecg classification. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*, 2022.

Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 101–111. Springer, 2023.

Han Yu, Huiyuan Yang, and Akane Sano. Leaves: Learning views for time-series data in contrastive learning. *arXiv preprint arXiv:2210.07340*, 2022a.

Han Yu, Huiyuan Yang, and Akane Sano. Ecg-sl: Electrocardiogram (ecg) segment learning, a deep learning method for ecg signal. *arXiv preprint arXiv:2310.00818*, 2023.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022b.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133, 2022.

Jing Zhang, Deng Liang, Aiping Liu, Min Gao, Xiang Chen, Xu Zhang, and Xun Chen. Mlbf-net: A multi-lead-branch fusion network for multi-class arrhythmia classification using 12-lead ecg. *IEEE journal of translational engineering in health and medicine*, 9:1–11, 2021.

Wenrui Zhang, Ling Yang, Shijia Geng, and Shenda Hong. Self-supervised time series representation learning via cross reconstruction transformer. *arXiv preprint arXiv:2205.09928*, 2022.

Jianwei Zheng, Huimin Chu, Daniele Struppa, Jianming Zhang, Sir Magdi Yacoub, Hesham El-Askary, Anthony Chang, Louis Ehwerhemuepha, Islam Abudayyeh, Alexander Barrett, et al. Optimal multi-stage arrhythmia classification approach. *Scientific reports*, 10(1):2898, 2020.

Zhaowei Zhu, Han Wang, Tingting Zhao, Yangming Guo, Zhuoyang Xu, Zhuo Liu, Siqi Liu, Xiang Lan, Xingzhi Sun, and Mengling Feng. Classification of cardiac abnormalities from ecg signals using se-resnet. In *2020 Computing in Cardiology*, pp. 1–4. IEEE, 2020.