Frozen Language Model Helps ECG Zero-Shot Learning

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

The electrocardiogram (ECG) is one of the most commonly used non-invasive, convenient medical monitoring tools that assist in the clinical diagnosis of heart diseases. Recently, deep learning (DL) techniques, particularly self-supervised learning (SSL), have demonstrated great potential in the classification of ECG. SSL pre-training has achieved competitive performance with only a small amount of annotated data after fine-tuning. However, current SSL methods rely on the availability of annotated data and are unable to predict labels not existing in fine-tuning datasets. To address this challenge, we propose **M**ultimodal **E**CG-**T**ext **S**elf-supervised pre-training (METS), **the first work** to utilize the auto-generated clinical reports to guide ECG SSL pre-training. We use a trainable ECG encoder and a frozen language model to embed paired ECG and automatically machine-generated clinical reports separately. The SSL aims to maximize the similarity between paired ECG and auto-generated report while minimize the similarity between ECG and other reports. In downstream classification tasks, METS achieves around 10% improvement in performance without using any annotated data via zero-shot classification, compared to other supervised and SSL baselines that rely on annotated data. Furthermore, METS achieves the highest recall and F1 scores on the MIT-BIH dataset, despite MIT-BIH containing different classes of ECG compared to the pre-trained dataset. The extensive experiments have demonstrated the advantages of using ECG-Text multimodal self-supervised learning in terms of generalizability, effectiveness, and efficiency.

Keywords: Multimodal self-supervised learning, Zero-shot learning, Language model, ECG, Signal processing

1. Introduction

The electrocardiogram (ECG) is a diagnostic tool that is widely used in clinical practice (Addison, 2005). In practice, the ECG is used to detect a wide range of cardiac conditions, including arrhythmias, heart attacks, and heart failure (Berkaya et al., 2018). Recently, deep learning (DL) methods have shown promising results in classifying ECG data (Ebrahimi et al., 2020; Tripathy et al., 2019). DL models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been shown to be highly accurate in classifying ECG for a variety of cardiac conditions (Baloglu et al., 2019; Singh et al., 2018; Xu et al., 2020). However, training DL models in a supervised manner (see Figure 1 (a)) often requires a large number of high-quality labels to obtain strong generalization performance (Ebrahimi et al., 2020). In addition, some forms of ECG, such as ST-elevation myocardial infarction, are difficult to detect and often require manual interpretation of the ECG by trained cardiologists (Ayer and Terkelsen, 2014). This work requires a huge effort which is costly and laborious.

1. Contributed equally

© 2023 Anonymous.

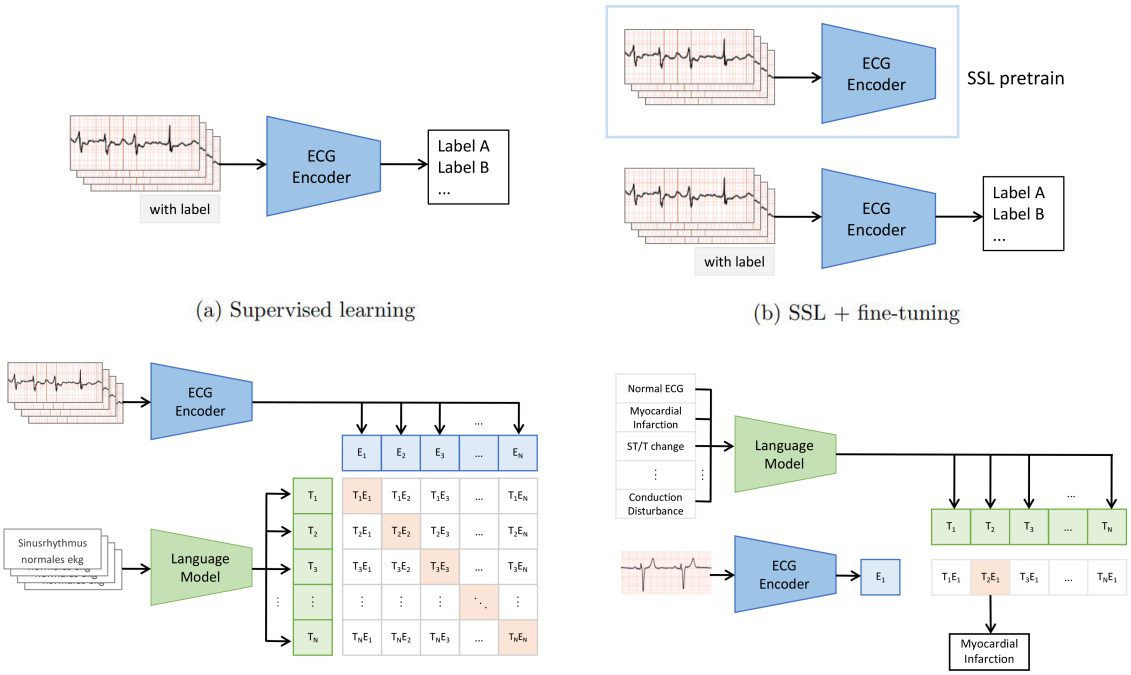

Figure 1: (a) denotes a supervised learning method for ECG. (b) denotes a common self-supervised learning method for ECG, i.e. pre-training followed by fine-tuning. (c) denotes a self-supervised learning method for multimodal ECG-Text. Zero-shot classification is performed after pre-training is completed.

Currently, self-supervised learning (SSL) has achieved impressive performance on datasets containing a small number of annotations, which provides a promising solution for unannotated ECG data (Jaiswal et al., 2020; Chou et al., 2020). It allows models to mine useful representations of ECG and can be widely used for various downstream tasks such as abnormality detection and arrhythmia classification (Lan et al., 2022; Mehari and Strodthoff, 2022). Nevertheless, existing ECG SSL methods still require a large amount of annotated data in order to fine-tune them for downstream tasks (see Figure 1 (b)). This requirement hinders the real-world application of ECG methods as some heart diseases are rare, which leads to problems with zero-shot learning. Zero-shot learning means that the model does not need any annotated samples for *unseen* categories (Socher et al., 2013). This is achieved by explicitly learning shared features from seen samples, and then generalizing them on unseen samples based on "descriptions" of the unseen categories' features (Xian et al., 2018; Pourpanah et al., 2020). Specifically, such "descriptions", are usually borrowed from external medical domain knowledge, textual ECG reports for example (see Figure 1 (c)). Zero-shot learning for ECG faces a number of challenges. The first challenge is the semantic gap, where ECG and text (automatically machine-generated ECG reports) are two heterogeneous modalities. ECG is long-term continuous numbers and text is relatively

short-term discrete clinical terminologies (Krishnan and Sowmya Kamath, 2018). They are difficult to align and characterize each other (Liang et al., 2022). The second challenge is domain adaptation. Zero-shot learning model may be sensitive to unknown domains, making it difficult to adapt to new domains or unseen categories and not performing well for downstream tasks in zero-shot learning. The third challenge is scalability. Zero-shot learning models need to learn a large number of representations and apply them to downstream tasks, which leads to a large computational cost for the model (Wang et al., 2019). Recently, (Yamaç et al., 2022) and (Bhaskarpandit et al., 2022) have reached considerable results on ECG zero-shot classification tasks. However, they pre-trained the model with supervised learning, which indicates that their methods still require large-scale annotated ECG for the pre-training stage.

To fully utilize the unannotated data, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021a) first implement multimodal SSL with two individual encoders and use zero-shot classification as the downstream task to evaluate SSL pre-trained model performance (Radford et al., 2021; Jia et al., 2021b). Florence (Yuan et al., 2021), LiT (Zhai et al., 2022), and ALBEF (Li et al., 2021) explore the potential of multimodal SSL on large-scale pre-training tasks (Yuan et al., 2021; Li et al., 2021; Zhai et al., 2022). Although recent works have achieved substantial progress on an image-text task, the medical signals-text, such as ECG, has not yet leveraged the benefits of the multimodal SSL.

To take advantage of multimodal SSL, we propose a novel method to do **M**ultimodal **E**CG-**T**EXT **S**SL pre-training (METS). The METS model takes the ECG and the corresponding reported text as input and feeds them into a multimodal comparative learning framework. The multimodal framework contains a language component and an ECG encoder to obtain embedding representations of the text and the ECG respectively. To make full use of the a priori clinical knowledge in the report text, the report text is fed into a large frozen language model with Resnet1d-18 as the backbone of the ECG encoder, both of which have a linear projection head to embed the text and ECG into the same dimension. Then, the similarity between ECG embedding and text embedding is computed to minimize the contrast learning loss and obtain a pre-trained model with rich medical knowledge. probabilities, which can be used to classify the various categories of ECG.

The main contributions of this paper are summarised as follows:

- Our proposed METS is **the first work** to apply a large language model for ECG SSL. The apriori clinical knowledge from the language model can be fully exploited to help generate ECG medical reports.

- METS is independent of the annotated categories. Even if the external dataset categories are unseen, the classification can be done directly with zero-shot, unlike other SSL which require fine-tuning.

- Experiments demonstrate that METS can be adapted to any of the downstream tasks, e.g. form, rhythm, and superclass, without the need to fine-tune on different tasks. Besides, METS does not require any annotated data but can exceed the supervised and SSL methods of fine-tuning with small-scale annotated data.

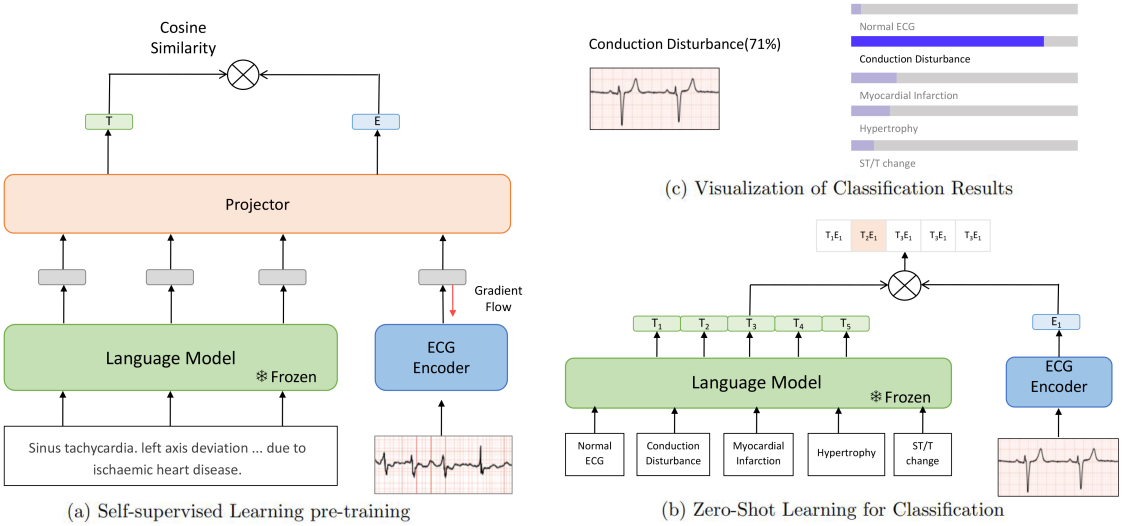

Figure 2: A framework for the METS approach. (a) shows a self-supervised pre-training approach. ECG-text pairs are fed into the model, and after comparative learning, the ECG encoder learns the parameters. (b) shows the zero-shot classification task. The corresponding labels are found by computing ECG and text similarity. (c) shows the visualization of the results of zero-shot classification.

2. Methods

In this section, we will demonstrate the details of METS. The framework is shown in Figure 2. METS consists of two components: Multimodal self-supervised pre-training (Section 2.1), and the zero-shot classification downstream task (Section 2.2).

2.1. Multimodal self-supervised pre-training

2.1.1. Frozen Pre-trained Language Models

Our approach starts with a large pre-trained language model based on the transformer architecture. In order for the large language model to fully understand the report text, we extend the report text into the language model as a complete sentence input (Radford et al., 2021). Specifically, we construct a prompt template for the report **"The report of the ECG is that {text}"**. We use a large clinical language model as the backbone of the text component. ClinicalBert has been pre-trained on all text from MIMIC III dataset (Alsentzer et al., 2019).

2.1.2. ECG Encoder

Our ECG encoder E_{ecg} is based on ResNet1d-18, which modifies the kernel of ResNet-18 from a 2D patch to a 1D stride, in order to obtain the deep ECG embeddings $\mathbf{e}$ (He et al., 2016; Hong et al., 2020). This process can be represented as follows: $\mathbf{e} = E_{ecg}(\mathbf{y})$, where $\mathbf{y}$

is the input of ECG. Then, a linear projection head f_e maps raw embeddings to $\mathbf{e}_d \in \mathbb{R}^D$. The embedding dimension of the ECG encoder is set to be the same as the language model embedding dimension d for comparison learning.

Inspired by (Tsimpoukelli et al., 2021), we freeze the parameters of the language model (LM) and use only paired ECG-text data from the PTB-XL dataset to update the parameters of the ECG encoder during SSL pre-training. This has the advantage of allowing the ECG encoder to learn rich prior clinical knowledge from the medical corpus, thus improving the generalization ability of the model. In addition, the parameters of the language model are frozen to reduce the significant computational cost of LM parameter updates.

2.1.3. Multimodal Contrastive Learning

Following the multimodal contrast learning framework, we treat a pair of report text and ECG belonging to the same patient as a positive sample pair, while treating pairs of other patients' report text and that ECG as negative sample pairs. We maximize the contrast loss of different pairs ($\mathbf{t_i}$, $\mathbf{e_j}$) and minimize the contrast loss of the same pair ($\mathbf{t_i}$, $\mathbf{e_i}$) to improve the similarity of the same pair of samples. We first define the similarity between the representations $\mathbf{t}$ and $\mathbf{e}$ of two modalities in terms of cosine similarity, as shown in Equation 1.

$$\operatorname{sim}(\mathbf{t}, \mathbf{e}) = \frac{t^\top \cdot e}{\|t\| \, \|e\|} \tag{1}$$

Then, we need to train two contrast loss functions. The first loss function is the ECG-to-text contrast loss for the ith pair, as shown in Equation 2.

$$\ell_{\mathbf{i}}^{(e \to t)} = -\log \frac{\exp\left(\operatorname{sim}\left(\mathbf{t_i}, \mathbf{e_i}\right)/\tau\right)}{\sum_{k=1}^{N} \exp\left(\operatorname{sim}\left(\mathbf{t_i}, \mathbf{e_j}\right)/\tau\right)} \tag{2}$$

The initialization of τ is set to 0.07.Similarly, the text-to-ECG contrast loss Equation 3 is represented as follows.

$$\ell_{\mathbf{i}}^{(t \to e)} = -\log \frac{\exp\left(\operatorname{sim}\left(\mathbf{e_i}, \mathbf{t_i}\right)/\tau\right)}{\sum_{k=1}^{N} \exp\left(\operatorname{sim}\left(\mathbf{e_i}, \mathbf{t_j}\right)/\tau\right)} \tag{3}$$

Finally, our training losses are calculated as the average combination of the two losses for all positive ECG-text pairs in each minibatch, as shown in Equation 4.

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \frac{\ell_{\mathbf{i}}^{(e \to t)} + \ell_{\mathbf{i}}^{(t \to e)}}{2} \tag{4}$$

2.2. Zero-Shot ECG Classification

In zero-shot classification, a segment of the ECG is used as input. To evaluate the zero-shot performance of the model on a multi-label classification task, we extend the discrete labels into full medical diagnostic statements and feed them into the language model to obtain embedding representations. Finally, the similarity between ECG embedding and text embedding is computed to obtain probabilities, which can be used to classify the various categories of ECG.

3. Experiments

3.1. Datasets

PTB-XL We use the PTB-XL dataset to train the METS model (Wagner et al., 2020). The PTB-XL dataset contains 21,837 clinical 12-lead ECG of 10 seconds duration from 18,885 patients, where each ECG segment is paired with the corresponding ECG reports. The reports are generated by the machine and only describe the ECG without final diagnosis. The original ECG reports were written in 70.89% German, 27.9% English, and 1.21% Swedish, and were converted into structured SCP-ECG statements. The statement sets were assigned to three non-reciprocal categories: diagnosis, form, and rhythm. Specifically, the dataset consisted of 71 different statements, broken down into 44 diagnostic statements, 12 rhythmic statements, and 15 formal statements. For the diagnostic labels were divided into 5 superclasses and 24 subclasses. In the current experiments, we focused on investigating ECG-text pairs without using any other labels. Following the experimental setup in (Huang et al., 2021; Wang et al., 2022), we extracted a multiclass classification dataset, PTB-XL test set, from the test set split.

PTB-XL Test Set The original ECG in the PTB-XL dataset is multi-labeled with diagnostic, form, and rhythm. In zero-shot downstream task classification, we need to calculate the similarity of ECG and text to find the most similar target, and multiple labels for a target can confuse the categories. Therefore, we produce diagnostic superclass, form, and rhythm test sets to complete the corresponding zero-shot downstream tasks. There are 1,000 samples on each test set. Details of the specific split test set are shown in figure 4.

MIT-BIH Test Set We use the MIT-BIH dataset for testing to evaluate the performance of our pre-trained representation framework for classification on external datasets (Moody and Mark, 2001). Please note that we do not pre-train on the MIT-BIH dataset. Similarly, we produced an MIT-BIH test set following the segmentation method above. Details of the specific split test set are also shown in figure 4.

Table 1: *PTB-XL* result on superclass. % refers to fractions of labels used in the training data.

Methods	Accuracy	Precision	Recall	F1
Self-supervised				
random - 5%	0.581	0.438	0.421	0.429
SimCLR - 5%	0.648	0.545	0.443	0.485
METS - 0%	**0.842**	**0.694**	**0.626**	**0.657**
Supervised				
Resnet18 - 100%	0.894	0.811	0.745	0.776

3.2. Implementation Details

The models for the transformer were taken from the transformer library (Wolf et al., 2020). We took a linear projection head with an output dimension of 128 and a temperature τ

initialized to 0.07. The ECG encoder is optimized using Adam optimizer with a learning rate of 1e-3 and weight decay of 1e-3. We use 50 epochs and a batch size of 32 for pre-training and downstream tasks. The experiments were conducted using PyTorch 1.7 on NVIDIA GeForce RTX-3090 GPU, which took about 8 hours.

3.3. Baselines

To demonstrate the performance of the METS method, our approach is compared with the following baselines. (1) **ResNet-18** (He et al., 2016): We choose ResNet-18 for showing the performance of small data fraction fine-tune. (2) **SimCLR** (Chen et al., 2020): A self-supervised contrast learning model that achieves good performance in SSL. We compare it with ECG SSL. The temperature parameter of SimCLR is set to 0.1. For all SSL methods above, we use 5% data for fine-tuning. (3) **Supervised** (He et al., 2016): We train ResNet1d-18 in a supervised manner, in order to compare the learning performance of our method with that of a fully supervised.

Table 2: *PTB-XL* result on form. % refers to fractions of labels used in the training data.

Methods	Accuracy	Precision	Recall	F1
Self-supervised				
random - 5%	0.603	0.364	0.342	0.351
SimCLR - 5%	0.660	0.446	0.471	0.456
METS - 0%	**0.734**	**0.537**	**0.503**	**0.518**
Supervised				
Resnet18 - 100%	0.724	0.520	0.508	0.509

3.4. Results and Discussion

In this experiment, we assessed the ECG classification using the commonly used metrics: Accuracy, Precision, Recall, F1. We first performed a zero-shot classification of the PTB-XL Test set for the diagnostic superclass. We illustrate the classification results for the diagnostic superclasses in Table 1. It can be found that our method outperforms all other SSL methods with comparable performance to supervised training. For example, compared to SimCLR, accuracy and F1 are improved by 11% and 4%, respectively. In contrast, METS far outperforms other SSL methods in the form classification results presented in Table 2 and achieves better performance than supervised learning in accuracy, precision, and F1. As shown in Table 3, METS also achieves good performance for the rhythm classification task. Overall, our results in PTB-XL show that the representations learned by METS are more informative than those of other state-of-the-art SSL methods. This also demonstrates that reports containing a priori knowledge can improve performance on the metrics.

We evaluated the performance of METS in migration learning. Table 4 shows the performance comparison under cross-dataset testing. In general, METS outperforms other state-of-the-art methods and even outperforms supervised learning. Compared to Table 1, there is a significant improvement in F1 for METS. This suggests that the features learned by METS are robust and have the potential to be generalized to other data sources.

Table 3: *PTB-XL* result on rhythm. % refers to fractions of labels used in the training data.

Methods	Accuracy	Precision	Recall	F1
Self-supervised				
random - 5%	0.627	0.435	0.442	0.438
SimCLR - 5%	0.697	0.516	0.565	0.549
METS - 0%	**0.746**	**0.576**	**0.612**	**0.593**
Supervised				
Resnet18 - 100%	0.790	0.664	0.607	0.633

Table 4: *MIT-BIH* result. % refers to fractions of labels used in the training data.

Methods	Accuracy	Precision	Recall	F1
Self-supervised				
random - 5%	0.565	0.468	0.499	0.483
SimCLR - 5%	0.749	0.642	0.610	0.624
METS - 0%	**0.794**	**0.680**	**0.735**	**0.706**
Supervised				
Resnet18 - 100%	0.836	0.697	0.712	0.704

4. Conclusion

In this paper, we present METS that uses automatically generated clinical reports to guide ECG pre-training. We pre-train the ECG encoder by applying the rich medical knowledge from the frozen large language model to the report text. As a result, our approach is independent of the class of annotated data and can be directly migrated to any unseen database. We can also do classification directly with zero-shot, unlike other SSL methods that require fine-tuning. Our experiments demonstrate that METS can be adapted to various downstream tasks, e.g. form, rhythm, disease, and abnormality classification. This means that the METS approach is more effective and efficient.

References

Paul S Addison. Wavelet transforms and the ecg: a review. *Physiological measurement*, 26 (5):R155, 2005.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Antoine Ayer and Christian Juhl Terkelsen. Difficult ecgs in stemi:: Lessons learned from serial sampling of pre-and in-hospital ecgs. *Journal of electrocardiology*, 47(4):448–458, 2014.

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.

Selcan Kaplan Berkaya, Alper Kursat Uysal, Efnan Sora Gunal, Semih Ergin, Serkan Gunal, and M Bilginer Gulmezoglu. A survey on ecg analysis. *Biomedical Signal Processing and Control*, 43:216–235, 2018.

Sathvik Bhaskarpandit, Priyanka Gupta, and Manik Gupta. Lets-gzsl: A latent embedding model for time series generalized zero shot learning. *arXiv preprint arXiv:2207.12007*, 2022.

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

Yen-hsiu Chou, Shenda Hong, Yuxi Zhou, Junyuan Shang, Moxian Song, and Hongyan Li. Knowledge-shot learning: An interpretable deep model for classifying imbalanced electrocardiography data. *Neurocomputing*, 417:64–73, 2020.

Zahra Ebrahimi, Mohammad Loni, Masoud Daneshtalab, and Arash Gharehbaghi. A review on deep learning methods for ecg arrhythmia classification. *Expert Systems with Applications: X*, 7:100033, 2020.

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.

Shenda Hong, Yanbo Xu, Alind Khare, Satria Priambada, Kevin Maher, Alaa Aljiffry, Jimeng Sun, and Alexey Tumanov. Holmes: health online model ensemble serving for deep learning models in intensive care units. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1614–1624, 2020.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

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*, pages 4904–4916. PMLR, 2021a.

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*, pages 4904–4916. PMLR, 2021b.

Gokul S Krishnan and S Sowmya Kamath. A supervised learning approach for icu mortality prediction based on unstructured electrocardiogram text reports. In *International Conference on Applications of Natural Language to Information Systems*, pages 126–134. Springer, 2018.

Xiang Lan, Dianwen Ng, Shenda Hong, and Mengling Feng. Intra-inter subject self-supervised learning for multivariate cardiac signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4532–4540, 2022.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705, 2021.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022.

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

George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001.

Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, and Xi-Zhao Wang. A review of generalized zero-shot learning methods. *arXiv preprint arXiv:2011.08641*, 2020.

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. PMLR, 2021.

Shraddha Singh, Saroj Kumar Pandey, Urja Pawar, and Rekh Ram Janghel. Classification of ecg arrhythmia using recurrent neural networks. *Procedia computer science*, 132:1290–1297, 2018.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26, 2013.

Rajesh Kumar Tripathy, Abhijit Bhattacharyya, and Ram Bilas Pachori. A novel approach for detection of myocardial infarction from ecg signals of multiple electrodes. *IEEE Sensors Journal*, 19(12):4509–4517, 2019.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

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.

Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

Xue Xu, Sohyun Jeong, and Jianqiang Li. Interpretation of electrocardiogram (ecg) rhythm by combined cnn and bilstm. *Ieee Access*, 8:125380–125388, 2020.

Mehmet Yamaç, Mert Duman, İlke Adalıoğlu, Serkan Kiranyaz, and Moncef Gabbouj. A personalized zero-shot ecg arrhythmia monitoring system: From sparse representation based domain adaption to energy efficient abnormal beat detection for practical ecg surveillance. *arXiv preprint arXiv:2207.07089*, 2022.

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*, pages 18123–18133, 2022.

Appendix A. Frozen Language Model Details

The ClinicalBert language model trained on all MIMIC notes and was initialized from BioBERT. The model was trained using code from Google's BERT repository on a GeForce GTX TITAN X 12 GB GPU. Model parameters were initialized with BioBERT. The model was trained using a batch size of 32, a maximum sequence length of 128, and a learning rate of 5e-5 for pre-training language models. The models trained on all MIMIC notes were trained for 150,000 steps. The dup factor for duplicating input data with different masks was set to 5. Specifically, masked language model probability = 0.15 and max predictions per sequence = 20 (Alsentzer et al., 2019).

Appendix B. Example texts of inputted Frozen Language Model

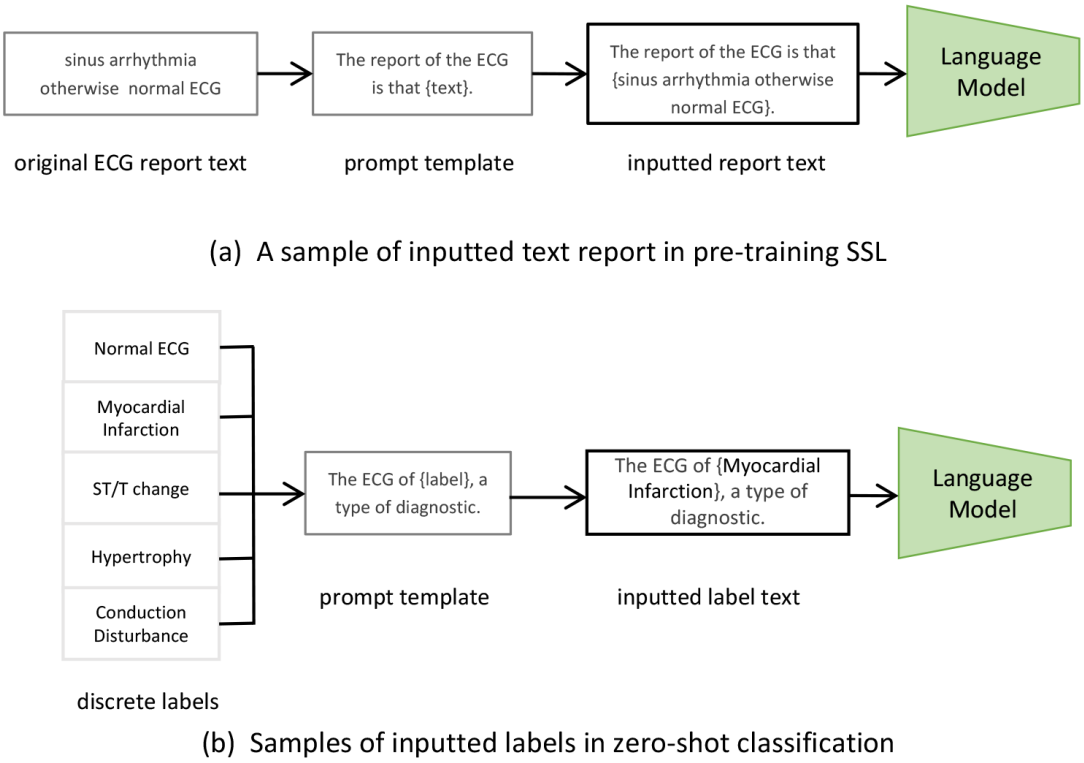

Figure 3: Example texts of inputted Frozen Language Model.

Example texts of inputted Frozen Language Model are shown in figure 4.

First, during the pre-training process of self-supervised learning, in order for the large language model to understand the report text, we extend the report text into the language model as a complete sentence input. A sample inputted text report is shown below.

"The report of the ECG is that {sinus arrhythmia otherwise normal ECG}".

Second, in zero-shot classification, we expand the labels into complete sentences for input into the frozen language model. This may be useful for specifying categories. Here are a few examples.

In the diagnostic classification task: **"The ECG of {X}, a type of diagnostic."** where X represents the different superclass diagnostic labels, e.g. Normal ECG, Conduction Disturbance, Myocardial Infarction, etc.

In the form classification task: **"The ECG of {Y}, a type of form."** where Y represents a different form label, such as Abnormal QRS, Non-diagnostic abnormalities, etc.

Appendix C. Description of test sets.

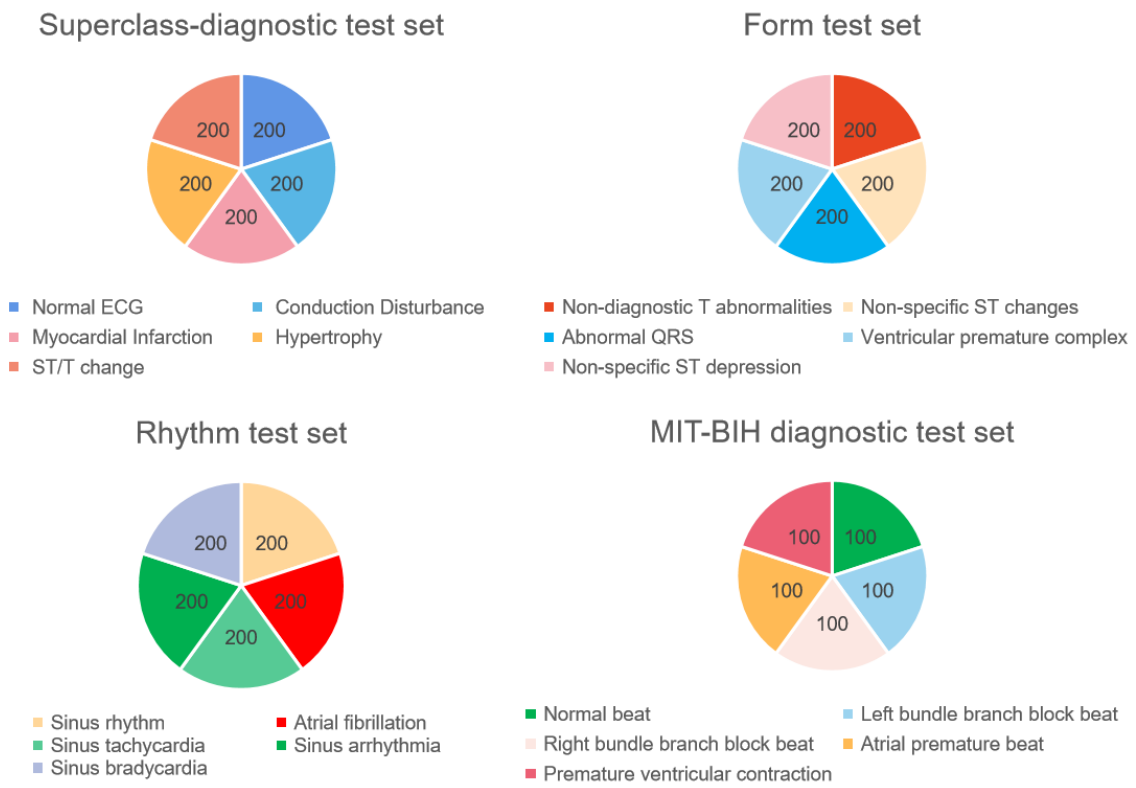

Figure 4: Description of test sets.