《统计学习与数据科学导论》 课 程 设 计

(2023-2024 学年 暑假学期)

基于临床报告和心电图的多模态自监督学习模型实现精准诊断

学生姓名：李嘉华

提交日期：2024 年 8 月 23 日　　学生签名：李嘉华

学　　号	202030201338	座位编号	
学　　院	土木与交通学院	专业班级	交通工程 1 班
课程名称	统计学习与数据科学导论课程设计	任课教师	夏立
教师评语：			
本论文成绩评定：　　分			

Multimodal Self-Supervised Learning Leveraging Clinical Reports and ECG for Accurate Diagnosis

Jiahua Li

School of Mathematics, South China University of Technology

Abstract

Electrocardiography (ECG) is a fundamental tool for diagnosing cardiovascular diseases, but traditional analysis methods are labor-intensive and prone to errors. Recent advancements in artificial intelligence (AI), particularly deep learning, have shown promise in automating ECG analysis. However, these methods often rely solely on ECG data, limiting their interpretability and generalizability. In this paper, we introduce a novel multimodal self-supervised learning (SSL) approach that leverages both ECG data and associated clinical reports. By integrating rich medical knowledge from pre-trained language models with ECG data, our method enables zero-shot classification, reducing the dependency on large annotated datasets. Extensive experiments on the PTB-XL dataset demonstrate that our approach outperforms state-of-the-art SSL methods and even surpasses supervised learning in various classification tasks, including form, rhythm, and disease. This work highlights the potential of multimodal SSL in improving ECG diagnostic accuracy and efficiency, making it a robust and scalable solution for clinical applications.

Keywords: Multimodal Self-Supervised Learning (SSL); Electrocardiography (ECG); Clinical Reports; Zero-Shot Classification; Deep Learning

1 Introduction

Electrocardiography (ECG), as a crucial clinical diagnostic tool, plays a significant role in the diagnosis and treatment of cardiovascular diseases by recording and analyzing the electrical activity of the heart. ECG is indispensable in the diagnosis of various cardiac conditions, including myocardial infarction, arrhythmias, and cardiomyopathies. With the continuous advancement of medical technology, ECG is not only widely used in the cardiac monitoring of

emergency and hospitalized patients but is also gaining increasing attention in chronic disease management and telemedicine.

Despite the undeniable value of ECG in cardiac diagnosis, it still faces numerous challenges within the medical field. Traditional ECG analysis predominantly relies on interpretation by experienced cardiologists, which is time-consuming and susceptible to human error. Moreover, with the aging population and the rising incidence of cardiovascular diseases, there is an urgent demand for automated and intelligent ECG analysis.

In recent years, the application of artificial intelligence (AI) technologies, particularly deep learning, in ECG analysis has garnered widespread attention. By training on large-scale ECG datasets, AI models can achieve automatic recognition and classification of complex ECG signals, thereby significantly improving diagnostic accuracy and efficiency. Although deep learning has shown great potential in ECG analysis, several critical technical challenges remain unresolved. Firstly, the diversity and complexity of ECG data pose significant challenges to the generalization capabilities of models. The significant variability in ECG signals across different patients necessitates the development of robust models capable of adapting to these differences, which is a key research focus. Secondly, the need for high-quality annotated ECG data has become a bottleneck in enhancing model performance. ECG data typically require annotation by professional physicians, which is both costly and time-consuming, leading to a scarcity of high-quality labeled data. Additionally, the presence of noise and artifacts in ECG data can negatively impact the diagnostic accuracy of models.

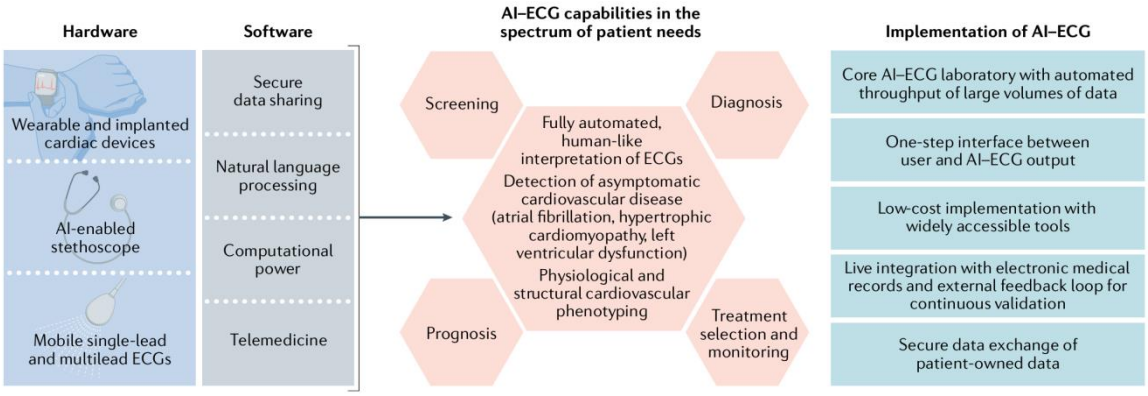


Figure 1: AI-Driven ECG Applications in Clinical Settings[1]

At the same time, most existing deep learning-based ECG analysis methods primarily rely on unimodal data, i.e., they utilize only ECG signals for analysis. From a medical perspective, unimodal data alone is often insufficient to fully reflect a patient's health status, and the interpretability of the predictions and underlying mechanisms remains limited. This limitation has long been a significant factor restricting the application of such methods in clinical practice.

In recent years, the introduction of multimodal data fusion models such as CLIP and BERT has opened a new pathway for disease diagnosis and analysis in clinical settings. By integrating multimodal data such as text, images, and other types, and simultaneously learning the associated features, these models enhance the understanding capabilities and improve the interpretability of deep learning models in clinical applications.

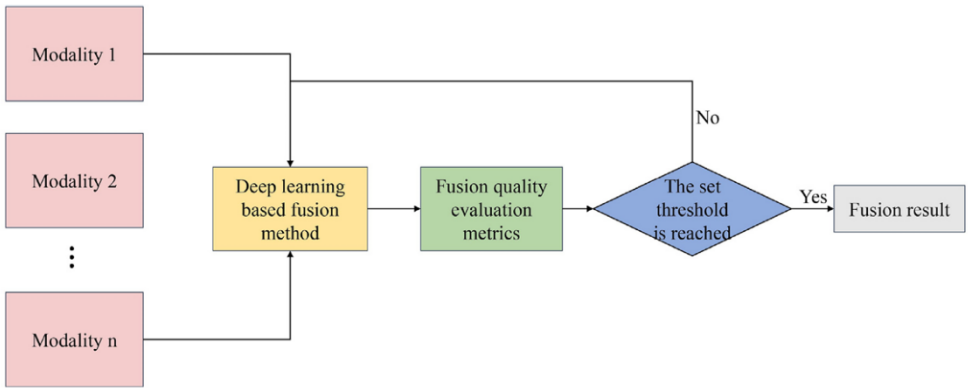


Figure 2 : Multimodal Biomedical Data Fusion Process

Given the aforementioned technical challenges, we developped a cardiac disease diagnostic method based on multimodal biomedical data and offered a more comprehensive and multidimensional assessment of a patient's heart health by integrating diverse data sources such as free-text data (e.g., electronic health records, EHR) and ECG. This approach would, in turn, improve diagnostic accuracy and interpretability.

2 Related Work

Recently, deep learning (DL) methods have shown promising results in ECG data classification[2-3]. DL models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated high accuracy in classifying ECG data associated with various cardiac conditions[4-5]. However, training DL models in a supervised manner typically requires large amounts of high-quality labeled data to achieve strong generalization

performance[3]. Additionally, certain ECG forms, such as ST-elevation myocardial infarction, are challenging to detect and often require manual interpretation by trained cardiologists, a task that is labor-intensive, costly, and time-consuming.

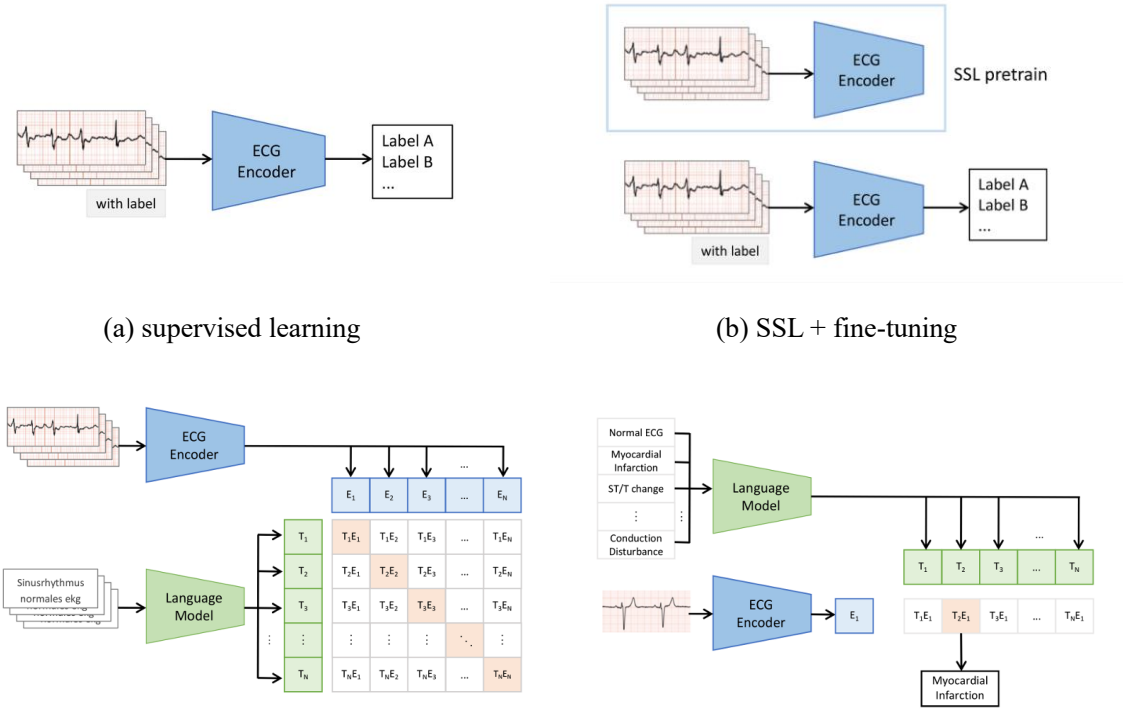
Currently, self-supervised learning (SSL) has achieved impressive performance on datasets with limited annotations, offering a promising solution for unlabeled ECG data[6-7]. SSL enables models to learn useful representations from ECG data, which can be widely applied to various downstream tasks such as anomaly detection and arrhythmia classification[8-9]. Nonetheless, existing ECG SSL methods still require a substantial amount of labeled data for fine-tuning in downstream tasks. This requirement hampers the practical application of ECG methods, particularly for certain rare cardiac conditions, leading to the zero-shot learning problem. Zero-shot learning enables models to generalize to unseen categories without requiring labeled samples from those categories. This is achieved by explicitly learning shared features from seen samples and then generalizing based on the "descriptions" of unseen categories[10-11]. Specifically, these "descriptions" are often derived from external medical domain knowledge, such as textual ECG reports.

Zero-shot learning in ECG faces several challenges. The first challenge is the semantic gap, where ECG and text (automatically generated ECG reports) are heterogeneous modalities. ECG signals are continuous over long periods, while text is composed of relatively short-term discrete clinical terms[12]. Aligning and representing these two modalities is difficult[13]. The second challenge is domain adaptability. Zero-shot learning models may be sensitive to unknown domains, making it challenging to adapt to new domains or unseen categories, and often resulting in poor performance in downstream tasks within 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 increases computational costs[14]. Recently, Yamaç and Bhaskarpandit et al.[15-16]achieved significant results in zero-shot ECG classification tasks. However, they utilized pre-trained models based on supervised learning, indicating that their methods still require large-scale labeled ECG data in the pre-training phase.

To fully exploit unlabeled data, CLIP[17] and ALIGN[18]were the first to implement multimodal SSL using two independent encoders and evaluated the performance of SSL pre-trained models using zero-shot classification as a downstream task. Florence, LiT, and ALBEF explored the

potential of multimodal SSL in large-scale pre-training tasks[19-21]. Despite recent advances in image-text tasks, the advantages of multimodal SSL have not yet been leveraged in medical signal-text scenarios, such as ECG.

To harness multimodal self-supervised learning (SSL), this paper proposes a novel approach for multimodal ECG-TEXT self-supervised learning.



(a) supervised learning

(b) SSL + fine-tuning

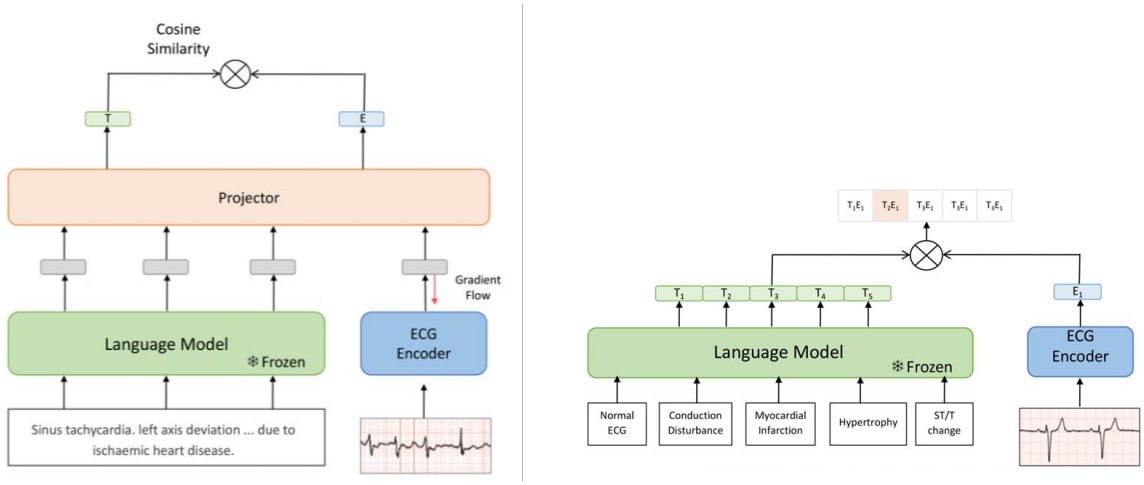
(c) ECG-Text SSL Pre-train and + Zero-shot

Figure3 (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.

3 Methods

To leverage multimodal SSL, we introduce a novel approach for pre-training that uses both ECG data and associated text reports as inputs. This method incorporates a comparative learning framework, which includes a language processing component and an ECG encoder to generate embedded representations for both modalities. The clinical knowledge contained in the report text is utilized by feeding it into a large, frozen language model, while the ECG data

is processed using an encoder based on Resnet1d-18. Both components have linear projection heads that map the text and ECG data into the same dimensional space. The similarity between the embeddings is then calculated to minimize the contrastive learning loss, resulting in a pre-trained model enriched with medical knowledge. This model can subsequently be applied to classify different ECG categories.



(a) Self-supervised Learning pre-training (b) Zero-shot Learning for Classification

Figure 4: A framework for the 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.

3.1 Multimodal self-supervised pre-training

3.1.1. Frozen Pre-trained Language Models

Our method utilizes a large language model based on the transformer architecture. To ensure the model comprehends the report text, we input the text into the model as a complete sentence following the format "The report of the ECG is that {text}". We employ a clinical language model as the foundation for the text component. ClinicalBert, which serves as this foundation, has been pre-trained on the entire text from the MIMIC III dataset.

3.1.2. ECG Encoder

Our ECG encoder is built on ResNet1d-18, adapting the ResNet-18 architecture by modifying its 2D kernel to a 1D stride to generate deep ECG embeddings. This process is expressed as

$e = E_{ecg}(y)$, where y represents the input ECG signal. Then, a linear projection head f_e maps raw embeddings to $e_d \in R^D$.The embedding dimension of the ECG encoder is set to be the same as the language model embedding dimension d for comparison learning.

We freeze the language model's parameters and only update the ECG encoder using paired ECG-text data from the PTB-XL dataset during SSL pre-training. This allows the ECG encoder to absorb rich clinical knowledge from the text corpus, enhancing the model's generalization capability. Freezing the language model's parameters also significantly reduces the computational burden associated with updating the language model.

3.1.3. Multimodal Contrastive Learning

Following the multimodal contrastive learning framework, we consider the report text and ECG from the same patient as a positive sample pair, while treating combinations of report texts from other patients and that ECG as negative sample pairs. To enhance the similarity of the same sample pairs, we maximize the contrastive loss of different pairs $(t_i e_j)$ and minimize the contrastive loss of the same pair $(t_i e_j)$. We first define the similarity between the representations t and e of the two modalities using cosine similarity, as shown in Equation 1.

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

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.

$$\ell_i^{(e\to t)} = -\log \frac{\exp(\mathrm{sim}(t_i, e_i)/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(t_i, e_j)/\tau)} \tag{2}$$

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

$$\ell_i^{(t\to e)} = -\log \frac{\exp(\mathrm{sim}(e_i, t_i)/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(e_i, t_j)/\tau)} \tag{3}$$

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.

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

3.2 Zero-Shot ECG Classification

In zero-shot classification, an ECG segment is used as input. To assess the model's zero-shot performance on a multi-label classification task, discrete labels are expanded into complete medical diagnostic statements, which are then processed by the language model to generate embeddings. The similarity between the ECG and text embeddings is calculated to determine probabilities, which are then used to classify the different ECG categories.

4 Experiment

4.1 Datasets

PTB-XL. We utilize the PTB-XL dataset for training the model. This dataset comprises 21,837 clinical 12-lead ECG recordings, each 10 seconds long, collected from 18,885 patients. Each ECG segment is paired with a corresponding machine-generated report, which describes the ECG but does not include a final diagnosis. The original reports were written in 70.89% German, 27.9% English, and 1.21% Swedish, and were later converted into structured SCP-ECG statements. These statements are categorized into three non-overlapping groups: diagnosis, form, and rhythm. Specifically, the dataset contains 71 unique statements, divided into 44 diagnostic, 12 rhythmic, and 15 form-related statements. Diagnostic labels are further grouped into 5 superclasses and 24 subclasses. In our experiments, we focus on exploring ECG-text pairs without incorporating additional labels. In addition, we extracted a multiclass classification dataset, referred to as the PTB-XL test set, from the test split.

PTB-XL Test Set. The original PTB-XL dataset includes multi-label annotations for diagnosis, form, and rhythm. In the context of zero-shot downstream task classification, it is necessary to compute the similarity between ECG and text embeddings to identify the closest match, as multiple labels for a single target can complicate categorization. To address this, we created separate test sets for diagnostic superclasses, form, and rhythm to perform the corresponding zero-shot downstream tasks. Each test set contains 1,000 samples.

MIT-BIH Test Set. We use the MIT-BIH dataset for testing to assess the performance of our pre-trained representation framework on external datasets. It is important to note that we did not pre-train the model on the MIT-BIH dataset. Similar to the approach used for PTB-XL, we created an MIT-BIH test set by following the same segmentation method.

4.2. Implementation Details

The transformer models were sourced from the transformer library. A linear projection head with an output dimension of 128 was used, and the temperature τ was set to an initial value of 0.07. The ECG encoder was optimized with the Adam optimizer, utilizing a learning rate of 1e-3 and a weight decay of 1e-3. We conducted pre-training and downstream tasks over 50 epochs with a batch size of 32. The experiments were performed using PyTorch 1.7 on an NVIDIA GeForce RTX-3090 GPU, and each experiment took approximately 8 hours to complete.

4.3 Baselines

To evaluate the effectiveness of our method, we compare it against the following baselines: We use ResNet-18[22] to demonstrate the performance when fine-tuning with a small fraction of data. SimCLR[23] is a self-supervised contrastive learning model known for its strong performance in SSL, which we compare against our ECG SSL method. The temperature parameter for SimCLR is set to 0.1. For all SSL methods mentioned, we use 5% of the data for fine-tuning. We train ResNet1d-18[24], a supervised learing model, in a fully supervised manner to compare its performance with our method.

4.4 Results and Discussion

In this experiment, we evaluated ECG classification using common metrics: Accuracy, Precision, Recall, and F1. We first conducted zero-shot classification on the PTB-XL Test set for diagnostic superclasses, with the results shown in Table 1. Our method outperforms all other SSL approaches and achieves performance comparable to supervised training. For instance, compared to SimCLR, our method improves accuracy by 11% and F1 by 4%. In form classification, as presented in Table 2, our approach significantly surpasses other SSL methods and even outperforms supervised learning in accuracy, precision, and F1. Overall, the PTB-XL results indicate that the representations learned by our approach are more informative than those

from other leading SSL methods, demonstrating that incorporating reports with prior knowledge can enhance performance.

We also tested our method in transfer learning scenarios, as shown in Table 3, comparing performance across different datasets. Our method consistently outperforms other advanced methods and even surpasses supervised learning, with notable improvements in F1 compared to Table 1. This suggests that the features learned by our approach are robust and generalizable to other datasets.

Table1 PTB-XL result on superclass. % refers to fractions of label used in the training data.

Method	Accucacy	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
Ours – 0%	**0.842**	**0.694**	**0.626**	**0.657**
Supervised				
Resnet18- 100%	0.894	0.811	0.745	0.776

Table2 PTB-XL result on form. % refers to fractions of label used in the training data.

Method	Accucacy	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
Ours – 0%	**0.734**	**0.537**	**0.503**	**0.518**
Supervised				
Resnet18- 100%	0.724	0.520	0.508	0.509

Table3 MIT-BIH result. % refers to fractions of label used in the training data.

Method	Accucacy	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
Ours – 0%	**0.794**	**0.680**	**0.735**	**0.706**
Supervised				
Resnet18- 100%	0.836	0.697	0.712	0.704

5 Conclusion

In this paper, we introduce a method which leverages automatically generated clinical reports to guide ECG pre-training. By utilizing the extensive medical knowledge embedded in a frozen large language model, we pre-train the ECG encoder on report text. This approach is not dependent on annotated data and can be easily transferred to any new database. Unlike other SSL methods that require fine-tuning, our method enables direct zero-shot classification. Our experiments show that this mehtod is versatile, adapting well to various downstream tasks such as form, rhythm, disease, and abnormality classification, making it a more effective and efficient approach.

With the rapid development of artificial intelligence and deep learning technologies in the medical field, the application of multimodal self-supervised learning (SSL) methods in the fusion of electrocardiogram (ECG) and textual data holds significant potential. The multimodal ECG-TEXT self-supervised pretraining method proposed in this study not only offers an innovative solution for the automated diagnosis of cardiac diseases from a technical perspective but may also have far-reaching impacts on medical research, clinical practice, and public health, as highlighted in the following three aspects:

1) Enhanced Accuracy and Robustness in ECG Classification: By integrating information from ECG signals and textual reports, this study significantly improves the accuracy and robustness of ECG classification. Traditional ECG analysis methods typically rely on unimodal data, which may lead to incomplete information and risks of misdiagnosis. The multimodal learning

approach proposed in this study effectively leverages prior clinical knowledge embedded in textual reports to complement the limitations of ECG data, thereby enhancing diagnostic precision and reliability. This technological advancement is expected to promote the development of automated ECG analysis systems, facilitating their broader application in clinical practice, alleviating the burden on physicians, and particularly aiding in resource-limited settings to provide timely and accurate diagnoses to more patients.

2) Demonstrating the Potential of SSL in Medical Applications: The multimodal diagnostic method proposed in this study showcases the immense potential of self-supervised learning in the medical domain. Unlike traditional supervised learning methods, self-supervised learning can effectively pretrain models using unlabeled data, thereby reducing the cost and time associated with model development. This is particularly significant in areas like medical imaging and biosignals, where data annotation is challenging and expensive. The successful implementation of this study will serve as a reference for the future application of self-supervised learning in other medical contexts, fostering the emergence of more innovative approaches.

3) Positive Impact on Public Health: By increasing the level of automation in ECG classification and diagnosis, the methods proposed in this study contribute to the early detection and prevention of cardiac diseases, reducing misdiagnosis and missed diagnoses, and ultimately lowering the incidence and mortality rates of heart diseases. Furthermore, the widespread adoption of this method may enhance the accessibility and equity of healthcare services globally, particularly in regions with limited medical resources, thereby promoting an overall improvement in health standards.

In summary, this study is of significant innovative importance both technically and in terms of application, and it is expected to have a broad and profound impact on medical research, clinical practice, and public health in the future.

Reference

[1] Siontis K C, Noseworthy P A, Attia Z I, et al. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management[J]. Nature Reviews Cardiology, 2021, 18(7): 465-478.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

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

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

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

[11] 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.

[12] 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.

[13] 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.

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

[15] Mehmet Yama¸c, Mert Duman, Ilk˙e Adalıo˘glu, 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[23] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. PMLR, 2020: 1597-1607.

[24] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.