### Improving ECG Diagnostics Through Multimodal Self-Supervised Learning

Jiahua, Li

### Abstract

...

(Currently incomplete, expected completion date: August 23.)

### 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.

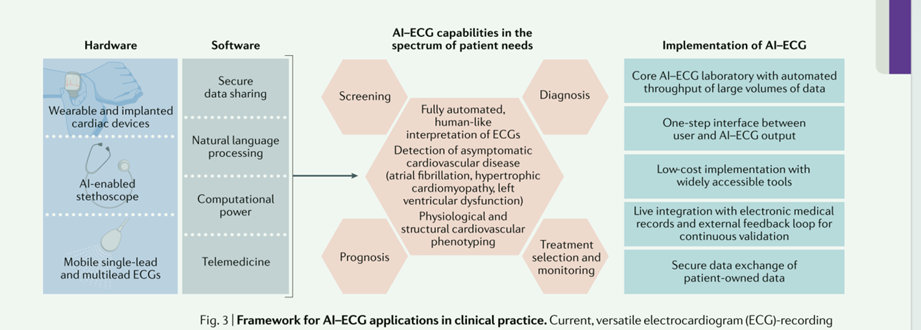
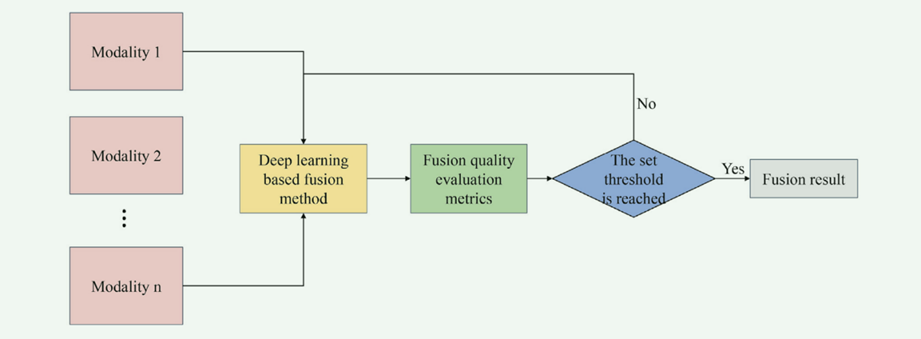


Figure 1: AI-Driven ECG Applications in Clinical Settings

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, developing a cardiac disease diagnostic method based on multimodal medical data could offer 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), ECG, and imaging data (e.g., echocardiography). 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[1-2]. 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[3-4]. However, training DL models in a supervised manner typically requires large amounts of high-quality labeled data to achieve strong generalization performance[1]. 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[5-6]. 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[7-8]. 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[9-10]. 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[11]. Aligning and representing these two modalities is difficult[12]. 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[13]. Recently, Yamaç and Bhaskarpandit et al.[14-15]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[16] and ALIGN[17]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[18-20]. 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.

### 3 Methods

Common multimodal biomedical data include image data, biomarker data, biomedical sensor data, diagnostic data, and others, as classified in the following figure:

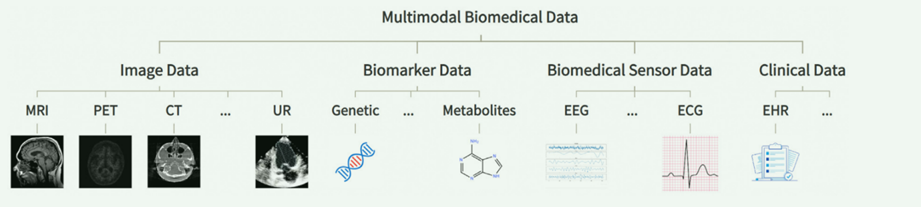


Figure 3: Common Multimodal Biomedical Data

In general, multimodal fusion can be conducted at three main levels.

First is data-level fusion, which typically involves summarizing biomarker indicators from multimodal biological information data, such as blood pressure and cell concentration.

Next is feature-level fusion, which is the mainstream approach to modality fusion. The goal here is to extract meaningful features from different physiological modes, integrate them at the feature level, create a more comprehensive representation of physiological states, and capture complex relationships between different signal modalities. This fusion strategy not only improves the accuracy and reliability of physiological assessments but also helps uncover intricate interactions that might be overlooked when analyzing individual signal modalities in isolation. It is generally applied in scenarios like disease classification and sleep state monitoring.

Lastly, decision-level fusion involves integrating decision outcomes from different modalities or feature layers to generate final decisions or predictions. This can be achieved by training multiple models, each responsible for processing different physiological signal modalities or features, and then fusing their decision outcomes.

The following section will introduce the primary methods used in this paper:

(Due to the ongoing experiments, the details of the methodology cannot be finalized at this time. Therefore, the remaining sections have not yet been written. These sections will be completed once the experiments are concluded, with an expected completion date of August 16.)

### 4 Experiment

4.1 Datasets

In this experiment, four datasets were used: CPSC2018, Chapman, MIMIC-IV-ECG, and UK Biobank-ECG.

(The experiments are still in progress, and this section has not been completed yet, with an expected completion date of August 16.)

### 5 Ablation Studies

The experiments are still in progress, and this section has not been completed yet. The expected completion date is August 16.

### 6 Conclusion

The experiments are still ongoing, and this section has not been completed yet. The expected completion date is August 23.

### Broader Impact

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] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.