Transfer Knowledge from Natural Language to Electrocardiography: Can We Detect Cardiovascular Disease Through Language Models?

Jielin Qiu[1]*, **William Han**[1]*, **Jiacheng Zhu**[1], **Mengdi Xu**[1],
Michael Rosenberg[3], **Emerson Liu**[2], **Douglas Weber**[1], **Ding Zhao**[1]
[1]Carnegie Mellon University, [2]Allegheny General Hospital, [3]University of Colorado

Abstract

Recent advancements in Large Language Models (LLMs) have drawn increasing attention since the learned embeddings pretrained on large-scale datasets have shown powerful ability in various downstream applications. However, whether the learned knowledge by LLMs can be transferred to clinical cardiology remains unknown. In this work, we aim to bridge this gap by transferring the knowledge of LLMs to clinical Electrocardiography (ECG). To address this problem, we propose an approach for cardiovascular disease diagnosis and automatic ECG diagnosis report generation. We also introduce an additional loss function by Optimal Transport (OT) to align the distribution between ECG and language embeddings. The learned embeddings are evaluated on two downstream tasks: (1) automatic ECG diagnosis report generation, and (2) zero-shot cardiovascular disease detection. Our approach is able to generate high-quality cardiac diagnosis reports and also achieves competitive zero-shot classification performance even compared with supervised baselines, which proves the feasibility of transferring knowledge from LLMs to the cardiac domain.

1 Introduction

Heart and cardiovascular diseases are the leading global cause of death, with 80% of cardiovascular disease-related deaths due to heart attacks and strokes. The clinical 12-lead ECG, when correctly interpreted, is the primary tool to detect cardiac abnormalities and heart-related issues. ECG provides unique information about the structure and electrical activity of the heart and systemic conditions through changes in the timing and morphology of the recorded waveforms. Achievements of ECG interpretation, such that critical and timely ECG interpretations of cardiac conditions, will lead to efficient and cost-effective intervention.

LLM starts from the Transformer model (Vaswani et al., 2017) and grows quickly with a wide range of applications (Devlin et al., 2019; Liu et al., 2019b; Brown et al., 2020). Recently, LLM has shown great potential for accelerating learning in many other domains since the learned embeddings can provide meaningful representation for downstream tasks. Examples include transferring the knowledge of LLM to, i.e., robotics control (Liang et al., 2022; Ahn et al., 2022), multimodal reasoning and interaction (Zeng et al., 2022; Zellers et al., 2021), robotics planning (Shah et al., 2022; Kant et al., 2022; Jain et al., 2022), decision-making (Li et al., 2022; Huang et al., 2022), robotics manipulation (Shridhar et al., 2022; Ren et al., 2022; Cui et al., 2022; Tam et al., 2022; Khandelwal et al., 2022), code generation (Fried et al., 2022), laws (Kaplan et al., 2020), computer vision (Radford et al., 2021), and so on.

Some previous works explored LLM and biological protein (Rives et al., 2021), or health records (Yang et al., 2022). However, the medical or healthcare domains contain so much domain knowledge that different sources preserve unique data characteristics without a unified paradigm. To the best of our knowledge, no previous work explores the knowledge transfer from LLM to cardiovascular disease with ECG signals.

In this work, we bridge the gap between LLM and clinical ECG by investigating the feasibility of transferring knowledge of LLM to the cardiology domain. Our contributions are listed as follows:

- To the best of our knowledge, our work is the first attempt to bridge the gap between LLM and clinical cardiovascular ECG by leveraging the knowledge from pretrained LLM.
- We propose a cardiovascular disease diagnosis and automatic ECG diagnosis report generation approach by transferring the knowledge from LLM to the cardiac ECG domain.
- We introduce an additional learning objective

* marked as equal contribution

based on Optimal Transport distance, which empowers the model to learn the distribution between ECG and language embedding.

- Our method can generate high-quality cardiac diagnosis reports and achieve competitive zero-shot classification performance even compared with supervised baselines, proving the feasibility of using LLM to enhance research and applications in the cardiac domain.

2 Related Work

Cardiovascular diagnosis via ECG The 12-lead ECG is derived from 10 electrodes placed on the surface of the skin (Cadogan, 2020). An ECG works by recording electrical activity corresponding to the heartbeat muscle contractions (Bonow et al., 2011). Although computerized interpretations of ECGs are widely used, automated approaches have not yet matched the quality of expert cardiologists, leading to poor patient outcomes or even fatality (Breen et al., 2019).

Deep learning in ECG Deep learning approaches have been rapidly adopted in many fields for their accuracy and flexibility, including ECG domain (Kiranyaz et al., 2015; Nonaka and Seita, 2021; Khurshid et al., 2021; Raghunath et al., 2021; Giudicessi et al., 2021; Strodthoff et al., 2021; Al-Zaiti et al., 2020; Acharya et al., 2017; Shanmugam et al., 2019; Śmigiel et al., 2021). Transformer (Vaswani et al., 2017) has recently been adopted in several ECG applications, i.e., arrhythmia classification, abnormalities detection, stress detection, etc (Yan et al., 2019; Che et al., 2021; Natarajan et al., 2020; Behinaein et al., 2021; Song et al., 2021; Weimann and Conrad, 2021).

LLM in healthcare Zhou et al. (2021) reviewed existing studies concerning NLP for smart healthcare. Yang et al. (2022) developed a large pretrained clinical language model using transformer architecture. Steinberg et al. (2021) showed that using patient representation schemes inspired by techniques in LLM can increase the accuracy of clinical prediction models. More related work can be found in Appendix B.

3 Methods

Problem Formulation We formulate the problem as generating cardiovascular diagnosis reports through pretrained LLMs. Given ECG signals $x = [x_1, x_2, ...x_t]$, our goal is to take advantage of the knowledge from LLM and learn a generated text embedding $L = [L_1, L_2, ..., L_m]$, which can then be decoded into natural language as reports or directly used for disease classification.

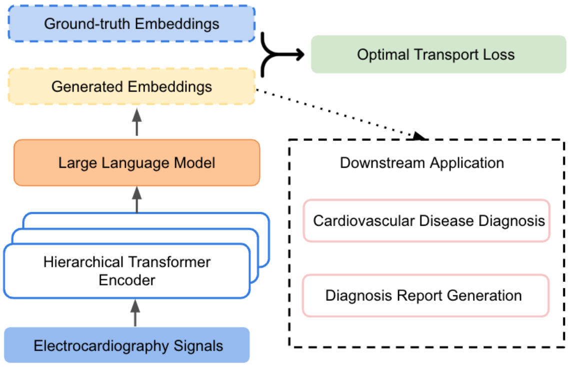
Ground-truth Embeddings
Optimal Transport Loss
Generated Embeddings
Large Language Model
Downstream Application
Cardiovascular Disease Diagnosis
Hierarchical Transformer Encoder
Diagnosis Report Generation
Electrocardiography Signals

Figure 1: The architecture of our model. The Transformer encoder takes input ECG to generate ECG features as the input to LLM, where LLM transforms it into generated embeddings. An optimal transport based loss objective is formulated on generated embeddings and ground-truth embeddings for the model update.

Model Architecture The model architecture is shown in Fig. 1, The ECG inputs are processed by hierarchical transformer encoders (Vaswani et al., 2017) to obtain transformed ECG embeddings $X = [X_1, X_2, ..., X_n]$. Then we adopt a pretrained LLM to transform the ECG embeddings into language embeddings $L = [L_1, L_2, ..., L_m]$. For the learning objective, we use expert reports to formalize the learning loss, which includes a new loss based on Optimal Transport (OT) in addition to the traditional cross-entropy loss. The learning objective is to update the transformer encoders, which can be interpreted as a sequence-to-sequence mapping from ECG embeddings X to sentence embeddings L. After the learning process, the learned embedding L should be capable of conducting downstream applications.

Downstream Applications For the downstream applications, we first consider a classification problem that uses the embeddings L for cardiovascular disease diagnosis. In addition, we consider a text generation task by decoding the output embeddings L into a cardiovascular report.

Transformer Encoders The transformer is based on the attention mechanism (Vaswani et al., 2017). The original transformer model is composed of an encoder and a decoder. The encoder maps an input sequence into a latent representation, and the decoder uses the representation with other inputs to generate a target sequence. Our model only adopts the encoder since the target is to learn the representations of ECG features. More details can be found in Appendix D.

Optimal Transport Loss OT is the problem of transporting mass between two discrete distributions supported on latent feature space $\mathcal{X}$. Let $\boldsymbol{\mu} = \{\boldsymbol{x}_i, \boldsymbol{\mu}_i\}_{i=1}^{n}$ and $\boldsymbol{v} = \{\boldsymbol{y}_j, \boldsymbol{v}_j\}_{j=1}^{m}$ be the distributions of generated embeddings and ground-truth embeddings, where $\boldsymbol{x}_i, \boldsymbol{y}_j \in \mathcal{X}$ denotes the spatial locations and μ_i, v_j, respectively, denoting the non-negative masses. Without loss of generality, we assume $\sum_i \mu_i = \sum_j v_j = 1$. $\pi \in \mathbb{R}_+^{n\times m}$ is a valid transport plan if its row and column marginals match μ and $\boldsymbol{v}$, respectively, which is $\sum_i \pi_{ij} = v_j$ and $\sum_j \pi_{ij} = \mu_i$. Intuitively, π transports π_{ij} units of mass at location $\boldsymbol{x}_i$ to new location $\boldsymbol{y}_j$. Such transport plans are not unique, and one often seeks a solution $\pi^* \in \Pi(\boldsymbol{\mu}, \boldsymbol{v})$ that is most preferable in other ways, where $\Pi(\boldsymbol{\mu}, \boldsymbol{v})$ denotes the set of all viable transport plans. OT finds a solution that is most cost-effective w.r.t. cost function $C(\boldsymbol{x}, \boldsymbol{y})$:

$$\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{v}) = \sum_{ij} \pi_{ij}^* C(\boldsymbol{x}_i, \boldsymbol{y}_j) = \inf_{\pi \in \Pi(\mu, v)} \sum_{ij} \pi_{ij} C(\boldsymbol{x}_i, \boldsymbol{y}_j) \tag{1}$$

where $\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{v})$ is known as OT distance. $\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{v})$ minimizes the transport cost from $\boldsymbol{\mu}$ to $\boldsymbol{v}$ w.r.t. $C(\boldsymbol{x}, \boldsymbol{y})$. When $C(\boldsymbol{x}, \boldsymbol{y})$ defines a distance metric on $\mathcal{X}$, and $\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{v})$ induces a distance metric on the space of probability distributions supported on $\mathcal{X}$, it becomes the Wasserstein Distance (WD). We use WD as one loss objective, in addition to the standard cross-entropy loss, for the model update.

4 Dataset and Prepossessing

Dataset We conducted the experiments on the PTB-XL dataset (Wagner et al., 2020), which contains clinical 12-lead ECG signals of 10-second length. There are five conditions in total, including Normal ECG (NORM), Myocardial Infarction (MI), ST/T Change (STTC), Conduction Disturbance (CD), and Hypertrophy (HYP). The waveform files are stored in WaveForm DataBase (WFDB) format with 16-bit precision at a resolution of 1μV/LSB and a sampling frequency of 100Hz. The ECG statements conform to the SCP-ECG standard and cover diagnostic, form, and rhythm statements.

Prepossessing The raw ECG signals are first processed by the WFDB library (Xie et al., 2022) and Fast Fourier transform (FFT) to process the time series data into the spectrum, which is shown in Fig. 2. Then we perform n-points window filtering to filter the noise within the original ECG signals and adopt notch processing to filter power frequency interference (noise frequency: 50Hz, quality factor: 30). The ECG signals are segmented by dividing the 10-second ECG signals into individual ECG beats. We first detect the R peaks of each signal by ECG detectors (Porr et al., 2022), and then slice the signal at a fixed-sized interval on both sides of the R peaks to obtain individual beats. More details can be found in Appendix C.

Feature Extraction Instead of directly using the time-series signals, we extract time domain and frequency domain features to better represent ECG signals. The time-domain features include: maximum, minimum, range, mean, median, mode, standard deviation, root mean square, mean square, k-order moment and skewness, kurtosis, kurtosis factor, waveform factor, pulse factor, and margin factor. The frequency-domain features include: FFT mean, FFT variance, FFT entropy, FFT energy, FFT skew, FFT kurt, FFT shape mean, FFT shape std, FFT shape skew, FFT kurt. More details can be found in Appendix C. An analysis of the statistics of the processed ECG data can also be found in Table 1.

Table 1: Statistics of the processed ECG data.

Category	Patients	Percentage	Beats	Percentage
NORM	9528	34.2%	28419	36.6%
MI	5486	19.7%	10959	14.1%
STTC	5250	18.9%	8906	11.5%
CD	4907	17.6%	20955	27.0%
HYP	2655	9.5%	8342	10.8%

5 Experiments

5.1 Experimental Settings

Data and Model The dimension of the processed ECG is 864, including 600 ECG signals and 264 time & frequency domain features. Experiments are conducted on two NVIDIA A6000 GPUs. All the models' parameters are listed in Appendix A.

Tasks To evaluate the learned embeddings from ECG signals, we tested the performance on two downstream applications: automatics cardiac report generation as a text generation (TG) task, and

Table 2: Comparisons of different backbones on Text generation (TG) and Disease detection (DD). (BERT as LLM)

Different backbones + BERT as LLM	Text generation (TG)						Disease detection (DD)		
	BLEU-1(%)	ROUGE-1(%)			Meteor(%)	BertScore(%)	Acc	AUCROC	F-1
		P	R	F					
MLP (Rumelhart et al., 1986)	22.24	17.68	22.63	18.11	14.27	84.68	0.71	0.89	0.57
LSTM (Hochreiter and Schmidhuber, 1997)	19.74	19.76	18.83	17.99	19.54	84.74	0.73	0.89	0.55
ResNet (He et al., 2016)	21.14	20.35	30.67	25.08	19.55	86.88	0.70	0.86	0.59
Transformer (Vaswani et al., 2017)	**26.93**	**25.35**	**35.67**	**28.08**	**21.23**	**88.90**	**0.77**	**0.92**	**0.68**

Table 3: Comparisons of different LLMs on Text generation (TG) and Disease detection (DD). (Transformer as the encoder).

Different LLMs	Text generation (TG)						Disease detection (DD)		
	BLEU-1(%)	ROUGE-1(%)			Meteor(%)	BertScore(%)	Acc	AUCROC	F-1
		P	R	F					
BERT (Devlin et al., 2019)	26.93	25.35	35.67	28.08	21.23	88.90	0.77	**0.92**	0.68
BART (Lewis et al., 2020)	27.21	**26.12**	35.71	**29.56**	**24.51**	89.61	0.75	0.88	0.68
RoBERTa (Liu et al., 2019b)	27.01	25.31	36.01	27.88	22.41	**89.72**	0.77	0.89	0.70
BioClinical BERT (Alsentzer et al., 2019)	**27.91**	25.41	**36.33**	28.42	23.54	87.21	**0.78**	0.89	**0.71**
PubMed BERT (Gu et al., 2022)	27.89	25.21	35.97	27.70	24.00	88.56	0.77	0.88	0.69
BioDischargeSummary BERT (Alsentzer et al., 2019)	26.81	25.32	35.66	28.10	21.19	88.90	0.73	0.85	0.66

Table 4: Comparisons with supervised baselines (DD).

Supervised learning baselines	Acc	AUROC	F-1
Transformer (Zhu et al., 2022)	0.75	0.843	0.575
CNN (Śmigiel et al., 2021)	0.72	0.877	0.611
SincNet (Ravanelli and Bengio, 2018)	0.73	0.84	0.6
Contrastive Learning (Lan et al., 2022)	–	0.722	–
CNN + Entropy (Śmigiel et al., 2021)	0.76	0.910	0.68
Ours$_{BERT}$	**0.77**	**0.92**	**0.68**

Table 5: Examples of comparison on generated reports (marked as Predicted-X) and ground-truth reports (marked as GT-X).

Backbone	Reports
GT-1	"sinus rhythm left type peripheral low voltage"
Predicted-1	"ventricular arrhythmia flatfar arrhythmia"
GT-2	"sinus rhythm incomplete right block otherwise normal ekg"
Predicted-2	"ventricularear extrasystole block sinus rhythm or normal."

zero-shot cardiac disease detection (DD) as a multi-class classification task.

Evaluation For text generation evaluation, we adopted the BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), Meteor (Banerjee and Lavie, 2005), and BertScore (Zhang et al., 2020) as evaluation metrics. We report the standard classification evaluation metrics for zero-shot cardiac disease detection: accuracy, AUCROC, and F-1 score.

5.2 Results

In Table 2, we showed the performance of both text generation and disease detection tasks with different backbone models as baselines. We found that the Transformer encoder outperforms other backbones, i.e., MLP, LSTM, and ResNet, showing Transformer encoder could be a good selection as the feature extractor.

In Table 4, we showed the performance of our zero-shot disease detection approach, compared with supervised baselines. Even though our method is in the zero-shot setting, we can already achieve the same performance with state-of-the-art supervised learning methods, demonstrating that the transferred ECG representation from LLM is already good for practical usage. We also showed some examples of generated reports compared with ground-truth reports in Table 5.

5.3 Ablation Study

Different LLM To further analyze the components, we conduct ablation studies on different LLMs and the number of transformer layers (with BERT as LLM). Table 3 shows the results of different LLMs for the text generation and disease detection tasks. We found that all LLMs showed good performance in both tasks, demonstrating that knowledge can be transferred from the language domain to the cardiac domain without constraints. BART shows good performance in the text generation task, while BioClinical BERT shows better performance in the disease detection task, though the variation between different LLMs is not large.

Transformer Layers To evaluate the impact of the number of transformer layers, we conducted additional experiments with different transformer layers, and the results are shown in Table 6. We

Table 6: Ablation study of different transformer layers.

Layers	Text generation (TG)						Disease detection (DD)		
	BLEU-1(%)	ROUGE-1(%)			Meteor(%)	BertScore(%)	Acc	AUCROC	F-1
		P	R	F					
1	25.81	20.36	30.72	23.12	21.38	83.58	0.69	0.83	0.59
2	24.77	19.22	28.55	24.51	20.44	82.89	0.72	0.81	0.61
3	25.44	20.44	27.21	24.81	19.99	84.63	0.75	0.80	0.62
4	25.12	21.36	30.88	25.76	**22.68**	86.35	0.74	0.80	0.64
5	**26.93**	**25.35**	**35.67**	**28.08**	21.23	**88.90**	**0.77**	**0.92**	**0.68**

Table 7: Comparisons with different backbones on the text generation task, where BERT is used as LLM.

Backbone	BLEU-1(%)	ROUGE-1(%)			Meteor(%)	BertScore
		P	R	F		F
MLP	18.16	16.19	13.71	14.48	12.11	80.77
LSTM	19.72	19.67	18.83	17.99	19.54	84.73
Resnet	21.15	20.35	20.67	24.08	19.55	85.22
Transformer	**24.51**	**23.22**	**30.81**	**26.19**	**20.02**	**85.44**

Table 8: Comparisons with different backbones on the disease detection task, where BERT is used as LLM.

Backbone	Acc	AUCROC	F-1
MLP	0.69	0.77	0.49
LSTM	0.71	0.82	0.59
Resnet	0.70	**0.83**	0.55
Transformer	**0.75**	0.81	**0.60**

found that more layers could lead to better representations, achieving better performance for downstream applications.

ECG Time Series Signals Only For the results above, we used ECG signals along with ECG time & frequency domain features as inputs. To compare the performance, we also conducted the experiments by only using ECG signals as inputs, with no time & frequency domain features. This set of experiments can be considered an additional ablation study for the inputs. The results are shown in Tables 7, 8, 9, 10.

Compare Table 7 & 8 with Table 2, we can find that the performance of only using ECG signals as inputs is lower than combining time & frequency features as inputs in both text generation and disease detection tasks, which demonstrates that incorporating time & frequency features is useful for capturing the characteristics of ECG and can lead to better representations through LLM.

In Tables 9, 10, the transformer backbone performs the best compared to others in both disease detection and text generation tasks, which is in consistent with the findings in the paper, showing that more layers could lead to better representations, achieving better performance for downstream applications. In addition, compared with Table 6 in the paper, we can find that the performance in Tables 9 and 10 are lower than the ones in Table 6, which also proved the same findings that adding time & frequency features is useful for learning the cardiac ECGs.

Table 9: Comparisons of different number of transformer layers on the text generation task, where BERT is used as LLM.

LLM	BLEU-1(%)	ROUGE-1(%)			Meteor(%)	BertScore(%)
		P	R	F		F
1	25.52	19.10	27.65	21.43	20.11	86.52
2	24.21	20.00	28.75	23.90	20.32	84.66
3	23.44	20.44	27.21	24.81	19.99	84.63
4	23.17	20.99	28.01	24.44	20.18	87.65
5	**25.69**	**24.75**	**34.81**	**27.59**	**21.03**	**87.33**

Table 10: Comparisons of different numbers of transformer layers on the disease detection task, where BERT is used as LLM..

Num of Layers	Acc	AUCROC	F-1
1	0.62	0.79	0.51
2	0.74	0.80	0.60
3	0.71	0.82	0.59
4	0.72	0.83	0.61
5	**0.75**	**0.88**	**0.64**

6 Conclusion

In this paper, we bridge the gap between LLMs and cardiovascular ECG by transferring knowledge of LLMs into the cardiovascular domain. The transferred knowledge embeddings can be used for downstream applications, including cardiovascular disease diagnosis and automatic ECG diagnosis report generation. Our results demonstrate the effectiveness of knowledge transfer, as the proposed method shows excellent performance in both downstream tasks, where our zero-shot classification approach even achieved competitive performance with supervised learning baselines, showing the feasibility of using LLM to enhance applications in the cardiovascular domain.

7 Acknowledgements

The research is partially supported by the DARPA ADAPTER program, and partially supported by the Allegheny Health Network and Mario Lemieux Center for Innovation and Research in EP.

8 Limitations

Due to the constrain of the available datasets, we only conducted experiments on the PTB-XL dataset, which is the current largest ECG dataset that contains high-quality clinical ECG signals and cardiac reports by experienced cardiologists.

We understand that collecting high-quality clinical data is much more complicated and time-consuming than collecting other data from online resources, like images, since it requires expert domain knowledge and is limited by many privacy regulations. We are working with cardiologists, hospitals, and clinical research labs, hope we can release a new dataset to provide additional materials for this research direction.

9 Ethics Statement

In this work, the data used as experimental materials are from publicly available databases, where the patients' information is anonymized. To the best of our knowledge, we do not foresee any harmful uses of this study.

References

U. Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Muhammad Adam, Arkadiusz Gertych, and Ru San Tan. 2017. A deep convolutional neural network model to classify heartbeats. *Computers in biology and medicine*, 89:389–396.

Michael Ahn et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *ArXiv*, abs/2204.01691.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings. *ArXiv*, abs/1904.03323.

Salah Al-Zaiti, Lucas Besomi, Zeineb Bouzid, Ziad Faramand, Stephanie O. Frisch, Christian Martin-Gill, Richard E. Gregg, Samir F. Saba, Clifton Callaway, and Ervin Sejdić. 2020. Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. *Nature Communications*, 11.

Ezra A. Amsterdam, J. Douglas Kirk, David A. Bluemke, Deborah B. Diercks, Michael E. Farkouh, J. Lee Garvey, Michael C Kontos, James McCord, Todd D. Miller, Anthony P Morise, L. Kristin Newby, Frederick L. Ruberg, Kristine Anne Scordo, and Paul D. Thompson. 2010. Testing of low-risk patients presenting to the emergency department with chest pain: a scientific statement from the american heart association. *Circulation*, 122 17:1756–76.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

Lanqing Bao, Jielin Qiu, Hao Tang, Wei-Long Zheng, and Bao-Liang Lu. 2019. Investigating sex differences in classification of five emotions from eeg and eye movement signals. *EMBC*, pages 6746–6749.

Behnam Behinaein, Anubha Bhatti, Dirk Rodenburg, Paul C. Hungler, and Ali Etemad. 2021. A transformer architecture for stress detection from ecg. *2021 International Symposium on Wearable Computers*.

Robert O Bonow, Douglas L Mann, Douglas P Zipes, and Peter Libby. 2011. *Braunwald's heart disease e-book: A textbook of cardiovascular medicine*. Elsevier Health Sciences.

C.J. Breen, G.P. Kelly, and W.G. Kernohan. 2019. Ecg interpretation skill acquisition: A review of learning, teaching and assessment. *Journal of Electrocardiology*.

Tom B. Brown et al. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Mike Cadogan. 2020. ECG Lead positioning.

Chao Che, Peiliang Zhang, Min Zhu, Yue Qu, and Bo Jin. 2021. Constrained transformer network for ecg signal processing and arrhythmia classification. *BMC Medical Informatics and Decision Making*, 21.

Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10635–10644.

Yuchen Cui, Scott Niekum, Abhi Gupta, Vikash Kumar, and Aravind Rajeswaran. 2022. Can foundation models perform zero-shot task specification for robot manipulation? *ArXiv*, abs/2204.11134.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual encoding for video retrieval by text. *IEEE transactions on pattern analysis and machine intelligence*, PP.

Daniel Fried et al. 2022. Incoder: A generative model for code infilling and synthesis. *ArXiv*, abs/2204.05999.

John R. Giudicessi, Matthew Schram, J. Martijn Bos, Conner Galloway, Jacqueline Baras Shreibati, Patrick W. Johnson, Rickey E. Carter, Levi W Disrud, Robert B Kleiman, Zachi I. Attia, Peter A. Noseworthy, Paul A. Friedman, David E. Albert, and Michael J. Ackerman. 2021. Artificial intelligence-enabled assessment of the heart rate corrected qt interval using a mobile electrocardiogram device. *Circulation*.

Yuxian Gu, Robert Tinn, Hao Cheng, Michael R. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1 – 23.

William Han, Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Douglas Weber, Bo Li, and Ding Zhao. 2022. An empirical exploration of cross-domain alignment between language and electroencephalogram. *ArXiv*, abs/2208.06348.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Nora Hollenstein, Cédric Renggli, Benjamin James Glaus, Maria Barrett, Marius Troendle, Nicolas Langer, and Ce Zhang. 2021. Decoding eeg brain activity for multi-modal natural language processing. *Frontiers in Human Neuroscience*, 15.

Renee Y. Hsia, Zachariah Hale, and Jeffrey A. Tabas. 2016. A national study of the prevalence of life-threatening diagnoses in patients with chest pain. *JAMA internal medicine*, 176 7:1029–32.

Wenlong Huang, P. Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *ICML*.

Vidhi Jain, Yixin Lin, Eric Undersander, Yonatan Bisk, and Akshara Rai. 2022. Transformers are adaptable task planners. *ArXiv*, abs/2207.02442.

Yash Kant, Arun Ramachandran, Sriram Yenamandra, Igor Gilitschenski, Dhruv Batra, Andrew Szot, and Harsh Agrawal. 2022. Housekeep: Tidying virtual households using commonsense reasoning. *ArXiv*, abs/2205.10712.

Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Simple but effective: Clip embeddings for embodied ai. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14809–14818.

Shaan Khurshid, Samuel N. Friedman, Christopher Reeder, Paolo Di Achille, Nathaniel Diamant, Pulkit Singh, Lia X. Harrington, Xin Wang, Mostafa A. Al-Alusi, Gopal Sarma, Andrea S. Foulkes, Patrick T. Ellinor, Christopher D Anderson, Jennifer E. Ho, Anthony A. Philippakis, Puneet Batra, and Steven A. Lubitz. 2021. Electrocardiogram-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation*.

Serkan Kiranyaz, Turker Ince, Ridha Hamila, and M. Gabbouj. 2015. Convolutional neural networks for patient-specific ecg classification. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2608–2611.

Xiang Lan, Dianwen Ng, linda Qiao, and Mengling Feng. 2022. Intra-inter subject self-supervised learning for multivariate cardiac signals. In *AAAI*.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. *ArXiv*, abs/1803.08024.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Shuang Li, Xavier Puig, Yilun Du, Clinton Jia Wang, Ekin Akyürek, Antonio Torralba, Jacob Andreas, and Igor Mordatch. 2022. Pre-trained language models for interactive decision-making. *ArXiv*, abs/2202.01771.

J. Liang, Wenlong Huang, F. Xia, Peng Xu, Karol Hausman, Brian Ichter, Peter R. Florence, and Andy Zeng. 2022. Code as policies: Language model programs for embodied control. *ArXiv*, abs/2209.07753.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. 2019a. Multimodal emotion recognition using deep canonical correlation analysis. *ArXiv*, abs/1908.05349.

Wei Liu, Jielin Qiu, Wei-Long Zheng, and Bao-Liang Lu. 2021. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b.

Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264:746–748.

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

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

Naoki Nonaka and Jun Seita. 2021. In-depth benchmarking of deep neural network architectures for ecg diagnosis. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, Proceedings of Machine Learning Research, pages 414–439. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Bernd Porr, Luis Howell, Ioannis Stournaras, and Yoav Nir. 2022. Popular ecg r peak detectors written in python.

Jielin Qiu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, D. Zhao, and Hailin Jin. 2022a. Liveseg: Unsupervised multimodal temporal segmentation of long livestream videos. *ArXiv*, abs/2210.05840.

Jielin Qiu, Ge Huang, and Tai Sing Lee. 2019. Visual sequence learning in hierarchical prediction networks and primate visual cortex. *Advances in neural information processing systems*.

Jielin Qiu, W. Liu, and Bao-Liang Lu. 2018a. Multi-view emotion recognition using deep canonical correlation analysis. *International Conference on Neural Information Processing*.

Jielin Qiu, Xin-Yi Qiu, and Kai Hu. 2018b. Emotion recognition based on gramian encoding visualization. *Brain Informatics*.

Jielin Qiu and Wei-Ye Zhao. 2018. Data encoding visualization based cognitive emotion recognition with ac-gan applied for denoising. *ICCI*CC*, pages 222–227.

Jielin Qiu, Jiacheng Zhu, Michael Rosenberg, Emerson Liu, and D. Zhao. 2022b. Optimal transport based data augmentation for heart disease diagnosis and prediction. *ArXiv*, abs/2202.00567.

Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Bo Li, Ding Zhao, and Hailin Jin. 2022c. Mhms: Multimodal hierarchical multimedia summarization. *ArXiv*, abs/2204.03734.

Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Bo Li, Ding Zhao, and Hailin Jin. 2022d. Semantics-consistent cross-domain summarization via optimal transport alignment. *ArXiv*, abs/2210.04722.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Sushravya Raghunath et al. 2021. Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ecg and help identify those at risk of atrial fibrillation–related stroke. *Circulation*, 143:1287 – 1298.

Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with sincnet. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028.

Allen Z. Ren, Bharat Govil, Tsung-Yen Yang, Karthik Narasimhan, and Anirudha Majumdar. 2022. Leveraging language for accelerated learning of tool manipulation. *ArXiv*, abs/2206.13074.

Alexander Rives, Siddharth Goyal, Joshua Meier, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 118.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning internal representations by error propagation.

Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. 2022. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. *ArXiv*, abs/2207.04429.

Divya Shanmugam, Davis Blalock, and John Guttag. 2019. Multiple instance learning for ecg risk stratification. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 124–139. PMLR.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. 2022. Perceiver-actor: A multi-task transformer for robotic manipulation. *ArXiv*, abs/2209.05451.

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

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. 2021. Transformer-based spatial-temporal feature learning for eeg decoding. *ArXiv*, abs/2106.11170.

Ethan H. Steinberg, Kenneth Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam Haresh Shah. 2021. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, page 103637.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. 2021. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE Journal of Biomedical and Health Informatics*, 25:1519–1528.

Allison C. Tam et al. 2022. Semantic exploration from language abstractions and pretrained representations. *ArXiv*, abs/2204.05080.

Kaisa Tiippana. 2014. What is the mcgurk effect? *Frontiers in Psychology*, 5.

Atousa Torabi, Niket Tandon, and Leonid Sigal. 2016. Learning language-visual embedding for movie understanding with natural-language. *ArXiv*, abs/1609.08124.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio', and Yoshua Bengio. 2018. Graph attention networks. *ArXiv*, abs/1710.10903.

Patrick Wagner, Nils Strodthoff, R. Bousseljot, D. Kreiseler, F. Lunze, W. Samek, and T. Schaeffter. 2020. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, 7.

Qinxin Wang, Haochen Tan, Sheng Shen, Michael W. Mahoney, and Zhewei Yao. 2020. An effective framework for weakly-supervised phrase grounding. *ArXiv*, abs/2010.05379.

Kuba Weimann and Tim O. F. Conrad. 2021. Transfer learning for ecg classification. *Scientific Reports*, 11.

Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. 2019. Fine-grained action retrieval through multiple parts-of-speech embeddings. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 450–459.

Hao Wu et al. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. *CVPR*, pages 6602–6611.

Chen Xie, Lucas McCullum, Alistair Johnson, Tom Pollard, Brian Gow, and Benjamin Moody. 2022. Waveform database software package (wfdb) for python (version 4.0.0). *PhysioNet*.

Genshen Yan, Shen Liang, Yanchun Zhang, and Fan Liu. 2019. Fusing transformer model with temporal features for ecg heartbeat classification. *BIBM*, pages 898–905.

Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. 2021. Taco: Token-aware cascade contrastive learning for video-text alignment. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11542–11552.

Xi Yang, Nima M. Pournejatian, Hoo-Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin B. Compas, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria P. Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *ArXiv*, abs/2203.03540.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *ECCV*.

Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-end concept word detection for video captioning, retrieval, and question answering. *CVPR*, pages 3261–3269.

B.P. Yuhas, M.H. Goldstein, and T.J. Sejnowski. 1989. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27:65–71.

Rowan Zellers, Ari Holtzman, Matthew E. Peters, Roozbeh Mottaghi, Aniruddha Kembhavi, Ali Farhadi, and Yejin Choi. 2021. Piglet: Language grounding through neuro-symbolic interaction in a 3d world. In *ACL*.

Andy Zeng, Adrian S. Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael S. Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Peter R. Florence. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *ArXiv*, abs/2204.00598.

Bowen Zhang, Hexiang Hu, and Fei Sha. 2018. Cross-modal and hierarchical modeling of video and text. In *ECCV*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Binggui Zhou, Guanghua Yang, Zheng Shi, and Shaodan Ma. 2021. Natural language processing for smart healthcare. *ArXiv*, abs/2110.15803.

Jiacheng Zhu, Jielin Qiu, Zhuolin Yang, Douglas Weber, Michael A. Rosenberg, Emerson Liu, Bo Li, and Ding Zhao. 2022. Geoecg: Data augmentation via wasserstein geodesic perturbation for robust electrocardiogram prediction. *ArXiv*, abs/2208.01220.

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

A Experiment Parameters

We provide the experimental parameters of the models in the paper in Table 11 and Table 12.

B More Related Work

Cardiovascular Disease in Current Practice Patients presenting with chest pain to the emergency department (ED) constitute a diagnostic and logistic challenge as chest pain can be caused by an extensive variety of disorders (Amsterdam et al., 2010). Diagnostic tests and decision algorithms play a critical role in speeding up the appropriate triage of chest pain patients in the ED, facilitating further (often more invasive) testing if warranted, and preventing unnecessary hospitalization of patients with non-critical disorders. In current practice, about half of the patients presenting with chest pain can be discharged from the ED, and only 5.5 percent of all ED visits lead to serious diagnoses (Hsia et al., 2016). However, research suggests the diagnosis of chest pain in the ED now costs an estimated $10 to $12 billion per year in the U.S. So a automatic cardiovascular disease diagnosis system is essential to provide cost-efficient patient care.

Deep learning in ECG Deep learning approaches have been rapidly adopted across a wide range of fields due to their accuracy and flexibility but require large labeled training sets. With the development in machine learning, many models have been applied to ECG disease detection (Kiranyaz et al., 2015; Nonaka and Seita, 2021; Khurshid et al., 2021; Raghunath et al., 2021; Giudicessi et al., 2021; Strodthoff et al., 2021; Qiu et al., 2022b; Zhu et al., 2022). Al-Zaiti et al. (2020) predicted acute myocardial ischemia in patients with chest pain with a fusion voting method. Acharya et al. (2017); Moody and Mark (2001) proposed a nine-layer deep convolutional neural network (CNN) to classify heartbeats in the MIT-BIH Arrhythmia database. Shanmugam et al. (2019) estimate a patient's risk of cardiovascular death after an acute coronary syndrome by a multiple instance learning framework. Recently, Śmigiel et al. (2021) proposed models based on SincNet (Ravanelli and Bengio, 2018) and used entropy-based features for cardiovascular diseases classification. The transformer model has also recently been adopted in several ECG applications, i.e., arrhythmia classification, abnormalities detection, stress detection, etc (Yan et al., 2019; Che et al., 2021; Natarajan et al., 2020; Behinaein et al., 2021; Song et al., 2021; Weimann and Conrad, 2021).

Multimodal Learning Formalized multimodal learning research dates back to 1989, when Yuhas et al. (1989) conducted an experiment that built off the McGurk Effect for audio-visual speech recognition using neural networks (Tiippana, 2014; McGurk and MacDonald, 1976). Aligning representations from different modalities is an important step in multimodal learning. With the recent advancement in computer vision and natural language processing, multimodal learning, which aims to explore the explicit relationship between vision and language, has drawn significant attention (Wang et al., 2020). There are many methods proposed for exploring the multimodal alignment objective. Torabi et al. (2016); Yu et al. (2017) adopted attention mechanisms, Dong et al. (2021); Qiu et al. (2022a,d,c) composed pairwise joint representation, Chen et al. (2020); Wray et al. (2019); Zhang et al. (2018) learned fine-grained or hierarchical alignment, Lee et al. (2018); Wu et al. (2019) decomposed the images and texts into sub-tokens, Velickovic et al. (2018); Yao et al. (2018) adopted graph attention for reasoning, and Yang et al. (2021) applied contrastive learning algorithms for video-text alignment.

Multimodal Learning in Healthcare Applications Many previous works have explored multimodal learning to boost performance in clinical healthcare applications, i.e., affective computing for depression disease detection and so on (Liu et al., 2021; Qiu et al., 2018a; Liu et al., 2019a; Qiu and Zhao, 2018; Qiu et al., 2018b, 2019; Han et al., 2022). Liu et al. (2021); Qiu et al. (2018a); Liu et al. (2019a); Qiu and Zhao (2018); Qiu et al. (2018b) explored the inner correlation between different modalities. Bao et al. (2019) investigated the demographics, showing that the subject's individual characteristics can also be involved in robustness and personalized design. Qiu et al. (2019) investigated the relationship between computational vision models and computational neuroscience. Hollenstein et al. (2021); Han et al. (2022) explored the connectivity between natural language and EEG signals.

C Prepossessing

The raw ECG signals are first processed by the WFDB library (Xie et al., 2022) and Fast Fourier

Table 11: Experiment parameters (best ones marked in bold).

Task	Batch Size	Encoder Layers	Att. Heads	Dropout	Epochs	Warmup Steps
Text Generation	[8, **16**, 32, 64]	[1, 2, 3, 4, **5**]	[1, 2, 3, 4, **5**]	[0.1, 0.2, **0.3**]	[10, 20, **50**, 100, 200]	[1000, **2000**]
Disease Detection	[8, **16**, 32, 64]	[1, 2, 3, 4, **5**]	[1, 2, 3, 4, **5**]	[0.1, 0.2, **0.3**]	[10, 20, **50**, 100, 200]	[1000, **2000**]

Table 12: Baseline parameters (best ones marked in bold).

Models	Batch Size	Layers	In Channel Size	Kernel Sizes	Dropout	Epochs	Warmup Steps
MLP	[8, **16**, 32, 64]	[**2**, 3, 4]	[**128**, 256, 512, 1024]	[**1**,3]	[0.1, 0.2, **0.3**]	[10, 20, **50**, 100, 200]	[1000, **2000**]
LSTM	[8, **16**, 32, 64]	[1, **2**, 3, 4]	[128, **256**, 512, 1024]	[**1**,3]	[0.1, 0.2, **0.3**]	[10, 20, **50**, 100, 200]	[1000, **2000**]
Resnet	[8, **16**, 32, 64]	[1, **2**, 3, 4]	[128, **256**, 512, 1024]	[**1**,3]	[0.1, 0.2, **0.3**]	[10, 20, **50**, 100, 200]	[1000, **2000**]
Transformer	[8, **16**, 32, 64]	[1, 2, 3, 4, **5**]	[128, **256**, 512, 1024]	[**1**,3]	[0.1, 0.2, **0.3**]	[10, 20, **50**, 100, 200]	[1000, **2000**]

transform (FFT) to process the time series data into the spectrum, which is shown in Fig. 2. Then we perform n-points window filtering to filter the noise within the original ECG signals and adopt notch processing to filter power frequency interference (noise frequency: 50Hz, quality factor: 30). The ECG signals are segmented by dividing the 10-second ECG signals into individual ECG beats. We first detect the R peaks of each signal by ECG detectors (Porr et al., 2022), and then slice the signal at a fixed-sized interval on both sides of the R peaks to obtain individual beats. Examples of the filtered ECG signal results after n-points window filtering, notch processing, R peak detection, and segmented ECG beats are shown in Figures. 3,4,5.

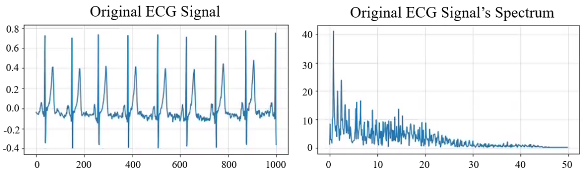

Figure 2: ECG data in the format of time series and spectrum.

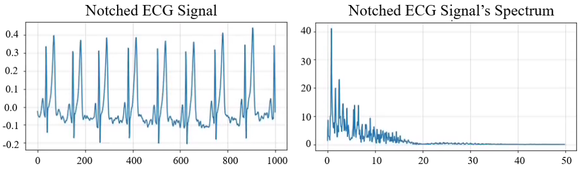

Figure 3: Filtered ECG data in the format of time series and spectrum.

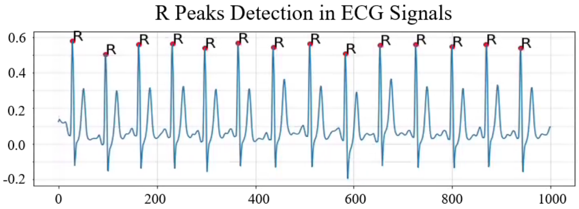

Figure 4: Detecting R peaks in the ECG signals.

Table 13: ECG statistical features in the frequency domain.

Feature Symbol	Formula
Z_1	$\frac{1}{N}\sum_{k=1}^{N} F(k)$
Z_2	$\frac{1}{N-1}\sum_{k=1}^{N} (F(k) - Z_1)^2$
Z_3	$-1 \times \sum_{k=1}^{N} \left(\frac{F(k)}{Z_1 N} \log_2 \frac{F(k)}{Z_1 N}\right)$
Z_4	$\frac{1}{N}\sum_{k=1}^{N}(F(k))^2$
Z_5	$\frac{1}{N}\sum_{k=1}^{N} \left(\frac{F(k)-Z_1}{\sqrt{Z_2}}\right)^3$
Z_6	$\frac{1}{N}\sum_{k=1}^{N} \left(\frac{F(k)-Z_1}{\sqrt{Z_2}}\right)^4$
Z_7	$\frac{\sum_{k=1}^{N}(f(k)-F(k))}{\sum_{k=1}^{N} F(k)}$
Z_8	$\sqrt{\frac{\sum_{k=1}^{N}\left[(f(k)-Z_6)^2 F(k)\right]}{\sum_{k=1}^{N} F(k)}}$
Z_9	$\frac{\sum_{k=1}^{N}\left[(f(k)-F(k))^3 F(k)\right]}{\sum_{k=1}^{N} F(k)}$
Z_{10}	$\frac{\sum_{k=1}^{N}\left[(f(k)-F(k))^4 F(k)\right]}{\sum_{k=1}^{N} F(k)}$

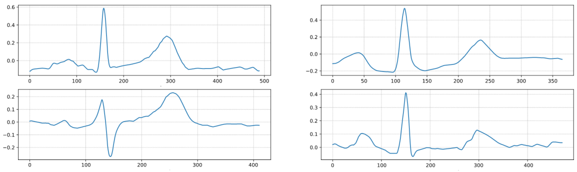

Figure 5: Extracted ECG beats divided by R peaks.

Feature Extraction Instead of directly using the time-series signals, we extract time domain and frequency domain features to better represent ECG signals. The time-domain features include: maximum, minimum, range, mean, median, mode, standard deviation, root mean square, mean square, k-order moment and skewness, kurtosis, kurtosis factor, waveform factor, pulse factor, and margin factor. The frequency-domain features include: FFT mean, FFT variance, FFT entropy, FFT energy, FFT skew, FFT kurt, FFT shape mean, FFT shape std, FFT shape skew, FFT kurt. The function of each component is shown in Table 13. An analysis of the statistics of the processed ECG data can also be found in Table 1.

D Transformer Encoders

The input for the Transformer is the ECG signal. First, we feed out the input into an embedding layer, which is a learned vector representation of each ECG feature by mapping each ECG feature to a vector with continuous values. Then we inject positional information into the embeddings by:

$$
\begin{aligned}
PE_{(pos,2i)} &= \sin\left(pos/10000^{2i/d_{\text{model}}}\right) \\
PE_{(pos,2i+1)} &= \cos\left(pos/10000^{2i/d_{\text{model}}}\right)
\end{aligned}
\tag{2}
$$

The attention model contains two sub-modules, a multi-headed attention model and a fully connected network. The multi-headed attention computes the attention weights for the input and produces an output vector with encoded information on how each feature should attend to all other features in the sequence. There are residual connections around each of the two sub-layers followed by a layer normalization, where the residual connection means adding the multi-headed attention output vector to the original positional input embedding, which helps the network train by allowing gradients to flow through the networks directly.

In our model, our attention model contains N same layers, and each layer contains two sub-layers, which are a multi-head self-attention model and a fully connected feed-forward network. Residual connection and normalization are added in each sub-layer. So the output of the sub-layer can be expressed as: Output $=$ LayerNorm$(x + (\text{SubLayer}(x)))$ For the Multi-head self-attention module, the attention can be expressed as: attention $=$ Attention(Q, K, V), where multi-head attention uses h different linear transformations to project query, key, and value, which are Q, K, and V, respectively, and finally concatenate different attention results:

$$\text{MultiHead(Q,K,V)} = \text{Concat}(head_1, ..., head_h)W^O \tag{3}$$

$$head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{4}$$

where the projections are parameter matrices:

$$
\begin{aligned}
W_i^Q \in \mathbb{R}^{d_{\text{model}}\ d_k}, \quad & W_i^K \in \mathbb{R}^{d_{\text{model}}\ d_k} \\
W_i^V \in \mathbb{R}^{d_{\text{model}}\ d_v}, \quad & W_i^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}
\end{aligned}
\tag{5}
$$

where the computation of attention adopted scaled dot-product:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

For the output, we use a 1D convolutional layer and softmax layer to calculate the final output.