Contrastive Learning for Multi-Label ECG Classification with Jaccard Score–Based Sigmoid Loss

Junichiro Takahashi Masataka Sato Satoshi Kodeta Norihiko Takeda
Department of Cardiovascular Medicine
The University of Tokyo Hospital
Tokyo, Japan
kodera@tke.att.ne.jp

Abstract

Recent advances in large language models (LLMs) have enabled the development of multimodal medical AI. While models such as MedGemini achieve high accuracy on VQA tasks like USMLE-MM, their performance on ECG-based tasks remains limited, and some models, such as MedGemma, do not support ECG data at all. Interpreting ECGs is inherently challenging, and diagnostic accuracy can vary depending on the interpreter's experience. Although echocardiography provides rich diagnostic information, it requires specialized equipment and personnel, limiting its availability.

In this study, we focus on constructing a robust ECG encoder for multimodal pretraining using real-world hospital data. We employ SigLIP, a CLIP-based model with a sigmoid-based loss function enabling multi-label prediction, and introduce a modified loss function tailored to the multi-label nature of ECG data. Experiments demonstrate that incorporating medical knowledge in the language model and applying the modified loss significantly improve multi-label ECG classification. To further enhance performance, we increase the embedding dimensionality and apply random cropping to mitigate data drift.

Finally, per-label analysis reveals which ECG findings are easier or harder to predict. Our study provides a foundational framework for developing medical models that utilize ECG data.

1 Introduction

In recent years, alongside the emergence of large language models (LLMs), multimodal medical AI has been developed. Recently, models such as MedGemini [6] and MedGemma [7] have been introduced, marking the appearance of multimodal models in the medical domain. However, while MedGemini achieves high accuracy on VQA tasks such as USMLE-MM, reaching 93.5%, its performance on ECG-QA, which involves electrocardiogram data (ECG), is considerably lower at 57.7%. In addition, MedGemma does not support an ECG at all. This discrepancy can be attributed to the inherently challenging nature of ECGs for model training.

In real-world clinical settings, interpreting ECGs is one of the more challenging tasks, and it is well known that diagnostic accuracy can vary significantly depending on the interpreter's professional background and level of experience [3]. Although transthoracic echocardiography is recommended for the diagnosis of cardiovascular diseases due to its rich informational content [2], it requires specialized technicians, and many facilities lack sufficient infrastructure to perform the examination [11]. In this context, the development of a multimodal model capable of handling electrocardiogram data and estimating echocardiographic findings from ECGs could provide substantial support in clinical settings. However, to date, no such clinically useful multimodal model exists.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

To build a high-quality multimodal model, it is essential to design ECG encoders suitable for the modality. In this study, we focus on the construction of a convincing encoder for ECGs. Previous studies have reported attempts to apply CLIP [5] has been used as a pretraining method for ECGs [1, 4, 12], but these approaches have several limitations. First, many of these studies rely on publicly available datasets such as PTB-XL [8] rather than real-world clinical data. While previous studies have made predictions using simplified PTB-XL labels, actual clinical findings are more finely categorized, and other predictive factors present in a single ECG can be overlooked. Therefore, a more detailed classification is necessary. Second, while real-world cardiovascular diseases often involve multiple abnormalities simultaneously, representing a multi-label problem, existing studies applying CLIP have been limited to single-class prediction. Therefore, CLIP-based contrastive learning for ECGs may not capture the inherent multi-label characteristics of ECG data.

In this study, we employed real-world hospital data and conducted pretraining based on SigLIP [13], assessing its performance in multi-label prediction tasks. SigLIP is a model that replaces the CrossEntropyLoss of CLIP with a sigmoid-based loss function, thereby enabling multi-label inference for each prediction. We also demonstrate that improving the loss function is necessary to enhance multi-label classification performance when training ECG data using SigLIP. Moreover, we addressed the clinically significant task of estimating echocardiographic findings from ECGs, investigating the potential of ECGs as a surrogate for echocardiography.

Overall, our study introduces two principal contributions. First, it leverages authentic clinical data for multimodal pretraining, enhancing the clinical validity of the model. Second, it adopts a sigmoid-based loss function to facilitate multi-label prediction, thereby enabling clinically meaningful inferences from ECGs that were not achievable with previous CLIP-based approaches.

2 Methods

2.1 Model architecture

In this study, we trained an ECG encoder using SigLIP and evaluated its performance in multi-label classification. The predicted findings are presented in the Appendix 5.1. We adopted a 1D ResNet-18 as an ECG Encoder as previous studies [1, 4] have reported superior performance compared with Vision Transformer (ViT) architectures. As the language model, we employed Qwen3-8B [10], which was selected based on preliminary evaluation indicating a favorable balance between model size and domain-specific knowledge regarding the target labels. For the ablation study, we utilized Gemma3-4B [9] to investigate whether ECG knowledge in language models influences pretraining effectiveness. By examining its generated outputs, we found that Gemma3-4B possesses limited ECG knowledge related to ECGs.

2.2 Dataset

The dataset consisted of 33,732 ECG data from our hospital. The ECG data consisted of 12-lead recordings sampled at 500 Hz over a duration of 10 seconds. We split the data such that the label distribution is uniform across the training, validation, and test sets. We ensured that the same patient did not appear across different splits, as this could lead to data leakage. The training text was formatted as: "This ECG shows {finding_1}, {finding_2}, ..., {finding_n}."

3 Experiments

We conducted a series of experiments for comparison. In the first experiment, we followed the standard SigLIP training process. In the second experiment, we modified the loss function of the standard SigLIP to account for the multi-label nature specific to ECG data. While SigLIP trains by treating diagonal pairs as the correct labels, ECG datasets with a limited number of diagnostic categories may contain patients with the same ECG findings within the same batch, which can lead to label conflicts. To address this issue, we modified the loss function. The modified loss was designed to treat patients with the same condition as similar pairs, and the loss calculation was adjusted accordingly to account for this similarity. We used the Jaccard Score as a metric for this similarity. Further details are provided in the Appendix 5.2.

For all two experiments, training was conducted using the Adam optimizer with a learning rate of 1×10^{-3}. The models were trained for 250 epochs, with a warm-up phase of 5,000 steps.

The results are presented in Table 1. Evaluation was performed using the multi-label metrics: Hamming Loss, Precision (Micro), Recall (Micro), F1 Score (Micro), and Jaccard Index.

Table 1: Results of the standard SigLIP and SigLIP with the modified loss

Metric	**Standard**	**Modified loss**
Hamming Loss	0.0665 ↓	0.0451 ↓
Precision (Micro)	0.5067 ↑	0.3147 ↑
Recall (Micro)	0.0365 ↑	0.3020 ↑
F1 Score (Micro)	0.0681 ↑	0.3082 ↑
Jaccard Index	0.0373 ↑	0.0858 ↑

From Table 1, it can be observed that the Modified Loss exhibits superior performance in multi-label ECG classification, as indicated by metrics such as F1 Score (Micro), Jaccard Index, and Hamming Loss.

In the third experiment, we trained SigLIP using a language model without ECG-related knowledge to investigate how the presence or absence of domain knowledge in the language model affects pretraining performance. In all subsequent experiments, we employ our Jaccard-based sigmoid loss function instead of the original sigmoid loss of SigLIP.

Table 2: Results of SigLIP with the modified loss, and Gemma3-4b

Metric	**Modified loss (Qwen3-8B)**	**Gemma3-4b**
Hamming Loss	0.0451 ↓	0.0539 ↓
Precision (Micro)	0.3147 ↑	0.2451 ↑
Recall (Micro)	0.3020 ↑	0.2970 ↑
F1 Score (Micro)	0.3082 ↑	0.2686 ↑
Jaccard Index	0.0858 ↑	0.0736 ↑

From the results in Table 2, it can be seen that the medical knowledge of the language model affects the overall performance of multi-label classification.

Through the experiments conducted thus far, we have demonstrated that employing the Modified Sigmoid Loss, which is tailored for multi-label classification, together with a language model incorporating medical knowledge, leads to performance improvements. However, the overall F1 Score (Micro) remains low at 0.3082, which is insufficient for practical applications.

To further enhance the F1 Score (Micro), we conducted several performance improvement experiments. The first approach involved increasing the dimensionality of the embedding vector, which represents the final similarity, from 128 to 256. The reason for increasing the embedding dimensionality is that 128 dimensions may be insufficient to adequately capture the representations of ECG signals. We also experimented with 512 dimensions, but no further performance improvement was observed; therefore, those results are omitted. The second approach aimed to address the issue of data drift by randomly cropping ECG waveforms. Since real ECG signals may vary in both start and end times, this variability can degrade performance. By applying random cropping, we mitigate this issue.

In addition, to ensure that the effect of random cropping is properly reflected in the model, we set the warmup steps to 20,000, following the original SigLIP paper, and increased the number of training epochs to 600.

The results are presented in Table 3. As a result, the final F1 Score (Micro) increased to 0.5028. Although the type and amount of data differ, this result achieves an F1-score comparable to that reported in the prior CLIP-based study [4]. From these results, it can be seen that increasing the embedding dimensionality to enhance ECG representation and applying random cropping to address data drift both contribute to improved multi-label prediction performance when training ECGs with SigLIP.

Table 3: Performance comparison of baseline and proposed enhancements

Metric	Baseline	Embedding dim 256	Embedding dim 256 + random crop (250 epoch, 5k warmup)	Embedding dim 256 + random crop (600 epoch, 20k warmup)
Hamming Loss	0.0451↓	0.0769↓	0.0856↓	0.0680↓
Precision (Micro)	0.3147↑	0.4097↑	0.3824↑	0.4898↑
Recall (Micro)	0.3020↑	0.3521↑	0.4636↑	0.5165↑
F1 Score (Micro)	0.3082↑	0.3788↑	0.4191↑	0.5028↑
Jaccard Index	0.0858↑	0.2218↑	0.2827↑	0.3495↑

We will now examine the classification performance of the final model for each individual label. The Accuracy, Precision, Recall, and F1 Score for each label are presented in Appendix Table 5.

From this table, it can be seen that some labels are easier to train with SigLIP-based contrastive learning on ECGs, while others are more difficult. For example, findings such as ventricular premature contractions and myocardial infarction have low F1 scores, indicating that they are difficult to predict from ECGs. Additionally, conditions observable via echocardiography, such as left atrial enlargement and left ventricular hypertrophy, have relatively low accuracy, showing that it is challenging to predict them without any misclassification. In contrast, labels such as atrial fibrillation, ST-T abnormalities, and right and left bundle branch blocks are easier to predict from ECGs. Additionally, for lowEF, which is a condition observable via echocardiography, the model achieves a high accuracy of 0.9138 and an F1 Score of 0.5152. Furthermore, as shown in Appendix 5.4, lowEF achieved a high AUC of 0.887, confirming its strong average predictive performance. This indicates that SigLIP is capable of predicting certain conditions, such as lowEF, which are typically identified from echocardiography, directly from ECG data.

We investigated whether performance degradation occurs when using ECG data obtained from a different hospital. The results are presented in Appendix 5.5. Overall, the F1 score decreased only slightly to 0.4841, a reduction of approximately 0.02, indicating minimal decline in the model's inference performance. Predictions for conditions such as lowEF also maintained an AUC of 0.888. These results suggest that our training approach is capable of preserving performance even on data from a different medical institution.

Furthermore, the experiments with the ResNet1D multi-label model are presented in Appendix 5.7. In these experiments as well, our model demonstrated superior performance.

4 Conclusion

In this study, we enhanced the performance of multi-label electrocardiogram (ECG) classification by employing a SigLIP-based ECG encoder trained on real-world clinical data and a modified loss function incorporating the Jaccard similarity. By increasing the embedding dimension and applying random cropping, the F1 score improved to 0.50, revealing which findings are relatively easy or difficult to predict. These results contribute to establishing a foundation for multimodal medical AI utilizing ECG data.

References

[1] Mingsheng Cai, Jiuming Jiang, Wenhao Huang, Che Liu, and Rossella Arcucci. Supreme: A supervised pre-training framework for multimodal ecg representation learning, 2025. URL `https://arxiv.org/abs/2502.19668`.

[2] Paul A Heidenreich, Biykem Bozkurt, David Aguilar, Larry A Allen, Joni J Byun, Monica M Colvin, Anita Deswal, Mark H Drazner, Shannon M Dunlay, Linda R Evers, et al. 2022 aha/acc/hfsa guideline for the management of heart failure: a report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. *Journal of the American College of Cardiology*, 79(17):e263–e421, 2022.

[3] Anthony H Kashou, Peter A Noseworthy, Thomas J Beckman, Nandan S Anavekar, Michael W Cullen, Kurt B Angstman, Benjamin J Sandefur, Brian P Shapiro, Brandon W Wiley, Andrew M

Kates, et al. Ecg interpretation proficiency of healthcare professionals. *Current problems in cardiology*, 48(10):101924, 2023.

[4] Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ecg zero-shot learning, 2023. URL https://arxiv.org/abs/2303.12311.

[5] 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. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

[6] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. Capabilities of gemini models in medicine, 2024. URL https://arxiv.org/abs/2404.18416.

[7] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025. URL https://arxiv.org/abs/2507.05201.

[8] Nils Strodthoff, Temesgen Mehari, Claudia Nagel, Philip J Aston, Ashish Sundar, Claus Graff, Jørgen K Kanters, Wilhelm Haverkamp, Olaf Dössel, Axel Loewe, et al. Ptb-xl+, a comprehensive electrocardiographic feature dataset. *Scientific data*, 10(1):279, 2023.

[9] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik

Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL `https://arxiv.org/abs/2503.19786`.

[10] Qwen Team. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

[11] Peter W Wood, Jonathan B Choy, Navin C Nanda, and Harald Becher. Left ventricular ejection fraction and volumes: it depends on the imaging method. *Echocardiography*, 31(1):87–100, 2014.

[12] Han Yu, Peikun Guo, and Akane Sano. Ecg semantic integrator (esi): A foundation ecg model pretrained with llm-enhanced cardiological text, 2024. URL `https://arxiv.org/abs/2405.19366`.

[13] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

5 Appendix

5.1 Labels

In this study, the labels used for training is selected under the guidance of the cardiologists. These labels are listed in Table 4. Note that the ground truth for lowEF, left ventricular hypertrophy, and left atrial enlargement was obtained from echocardiography data not than from ECG.

Table 4: ECG findings used in this study

ECG Findings
Left ventricular hypertrophy
Left atrial enlargement
Low ejection fraction (lowEF)
Normal range (Normal)
Prolonged QT interval
Tall T wave
Left axis deviation
Artificial pacemaker rhythm
Intraventricular conduction delay
Complete right bundle branch block
Complete left bundle branch block
Flat T wave
Inverted T wave
ST-T abnormality
Poor R wave progression
Abnormal Q wave
Anterior wall myocardial infarction
Lateral wall myocardial infarction
Inferior wall myocardial infarction
Anterior septal myocardial infarction
Ventricular premature contraction
Frequent ventricular premature contraction
Ventricular bigeminy
Ventricular tachycardia
Couplet of ventricular premature contractions
Atrial fibrillation

5.2 Modified sigmoid loss

We improved the original loss (Listing 1) to enhance multi-label prediction performance.

```
# img_emb    : image model embedding [n, dim]
# txt_emb    : text model embedding [n, dim]
# t_prime, b : learnable temperature and bias
# n          : mini-batch size

t = exp(t_prime)
zimg = l2_normalize(img_emb)
ztxt = l2_normalize(txt_emb)
logits = dot(zimg, ztxt.T) * t + b
labels = 2 * eye(n) - ones(n) # -1 with diagonal 1
l = -sum(log_sigmoid(labels * logits)) / n
```

Listing 1: Original Sigmoid loss pseudo-implementation.

Specifically, we modified the *eye* component in Listing 1. The original *eye* is defined as a diagonal matrix

$$\text{eye} = \{E \in \{0,1\}^{n\times n} \mid E_{ii} = 1,\ E_{ij} = 0\ (i \neq j)\}, \tag{1}$$

that is, a matrix whose diagonal entries are one and off-diagonal entries are zero. The entries of one correspond to positive labels, whereas the zeros represent negative labels. This implies that the i-th ECG finding is considered positive only for the i-th label.

However, it can easily occur that the patients with the same diseases are included in the same batch. We then modified the *eye* in Eq. 1 based on the similarity of ECG findings among patients within a batch.

We employed the Jaccard similarity to represent the similarity of these ECG findings. The modified *eye* is defined as in Eq. 2, where the set of ECG findings for the i-th data is denoted by A_i and that for the j-th data is denoted by A_j.

$$\mathrm{Jaccard}(A_i, A_j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}, \quad \mathrm{eye}_{ij} = \mathrm{Jaccard}(A_i, A_j), \quad \forall i, j \in \{1, \ldots, n\}, \tag{2}$$

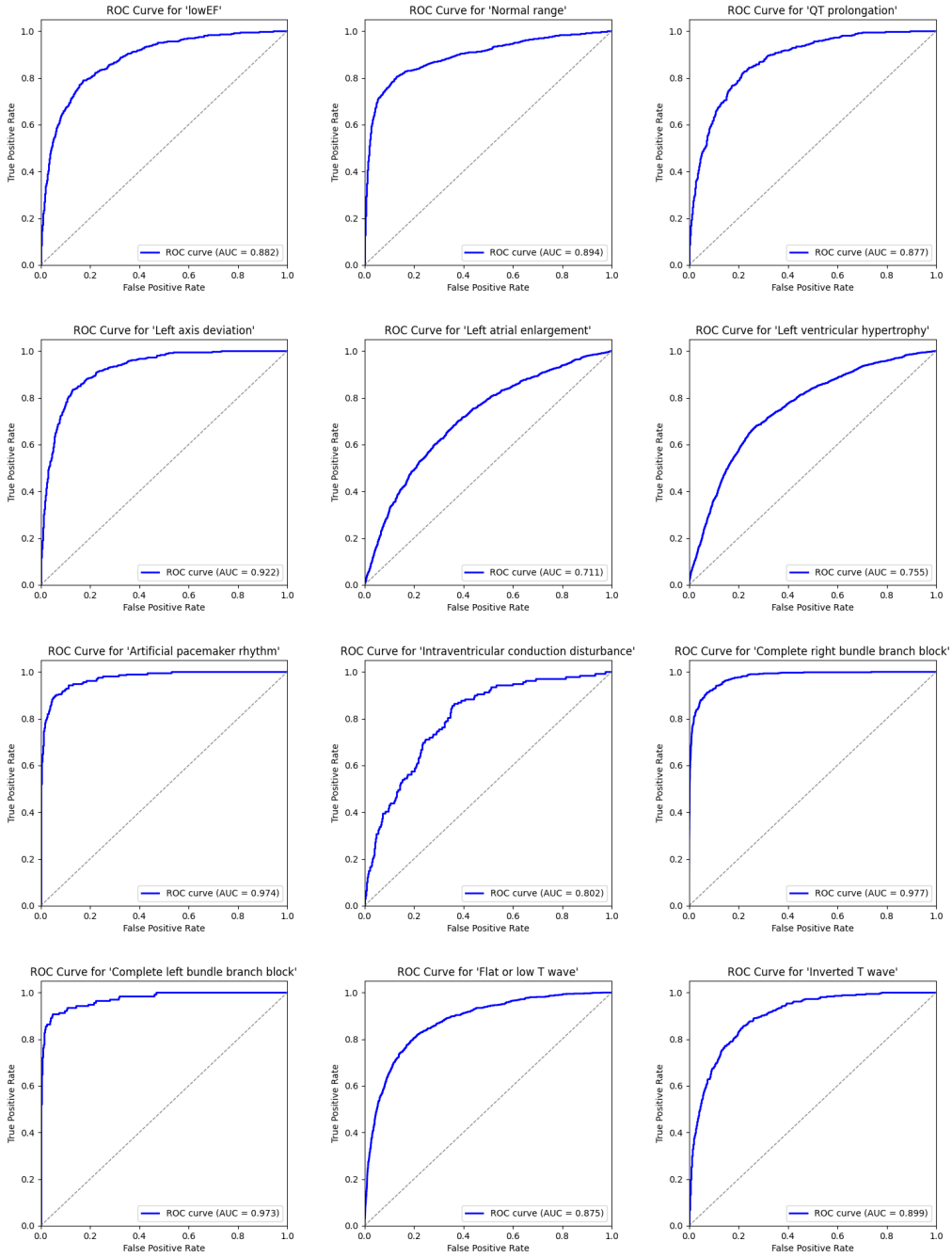
The Jaccard similarity satisfies $0 \leq \mathrm{Jaccard}(A_i, A_j) \leq 1$, $\mathrm{Jaccard}(A_i, A_j) = \mathrm{Jaccard}(A_j, A_i)$, and $\mathrm{Jaccard}(A_i, A_j) = 1$ when $i = j$. Here, a value of $\mathrm{Jaccard}(A_i, A_j)$ closer to 1 indicates that the patients have more similar diseases. Using the modified *eye* defined in Eq. 2, we conducted the experiments in this study.

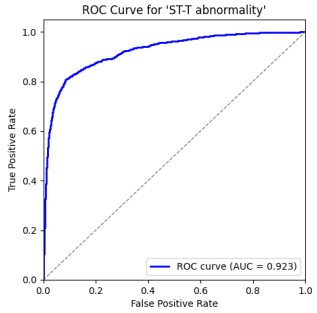
5.3 Appendix: Individual Label Metrics 5

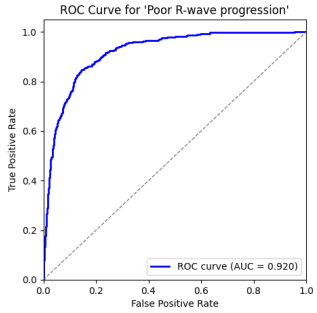
Table 5: Classification performance for each label of the final model

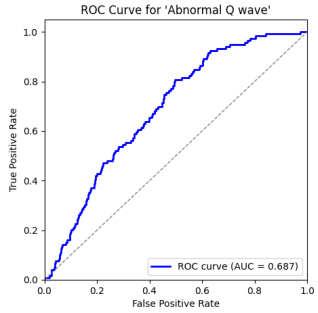
Label	Accuracy	Precision	Recall	F1-Score
lowEF	0.9138	0.5038	0.5271	0.5152
Normal	0.9091	0.8054	0.5526	0.6555
Prolonged QT	0.9368	0.5161	0.2753	0.3590
Tall T wave	0.9842	0.1471	0.1515	0.1493
Left axis deviation	0.9296	0.3872	0.5884	0.4670
Left atrial enlargement	0.7949	0.4000	0.3336	0.3638
Left ventricular hypertrophy	0.7404	0.5932	0.4410	0.5059
Artificial pacemaker rhythm	0.9804	0.6564	0.6995	0.6773
Intraventricular conduction delay	0.9578	0.1085	0.1655	0.1311
Complete right bundle branch block	0.9674	0.8351	0.7607	0.7962
Complete left bundle branch block	0.9737	0.4138	0.8571	0.5581
Flat T wave	0.8808	0.5251	0.5849	0.5534
Inverted T wave	0.9355	0.5065	0.4140	0.4556
ST-T abnormality	0.9122	0.8242	0.6239	0.7102
Poor R wave progression	0.9339	0.4179	0.6062	0.4947
Abnormal Q wave	0.9553	0.0234	0.0420	0.0300
Anterior wall myocardial infarction	0.9314	0.0422	0.3226	0.0746
Lateral wall myocardial infarction	0.9423	0.0382	0.2778	0.0671
Inferior wall myocardial infarction	0.9380	0.1887	0.4348	0.2632
Anterior septal myocardial infarction	0.9426	0.1771	0.4551	0.2549
Ventricular premature contraction	0.9163	0.3149	0.3242	0.3195
Frequent ventricular premature contraction	0.9751	0.4298	0.3190	0.3662
Ventricular bigeminy	0.9777	0.1020	0.3409	0.1571
Ventricular tachycardia	0.9665	0.0000	0.0000	0.0000
Couplet of ventricular premature contractions	0.9672	0.0314	0.1034	0.0482
Atrial fibrillation	0.9685	0.8971	0.8700	0.8833

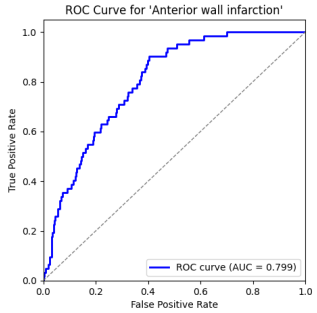
5.4 Appendix: ROC curves

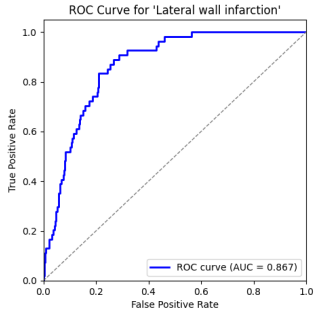


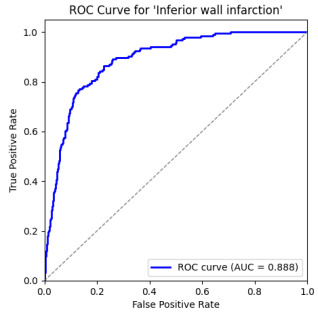

ROC Curve for 'ST-T abnormality'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.923)

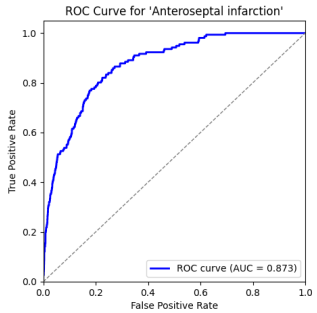

ROC Curve for 'Poor R-wave progression'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.920)

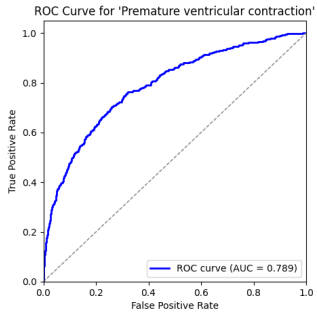

ROC Curve for 'Abnormal Q wave'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.687)

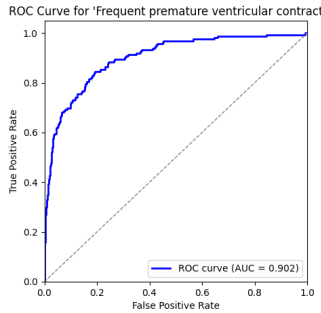

ROC Curve for 'Anterior wall infarction'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.799)

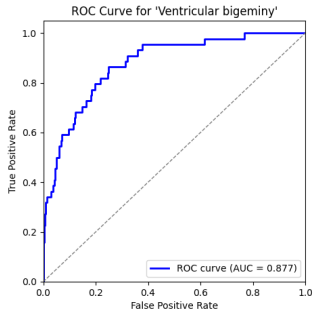

ROC Curve for 'Lateral wall infarction'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.867)

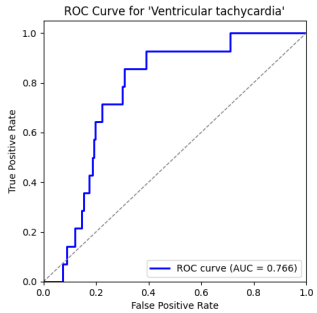

ROC Curve for 'Inferior wall infarction'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.888)

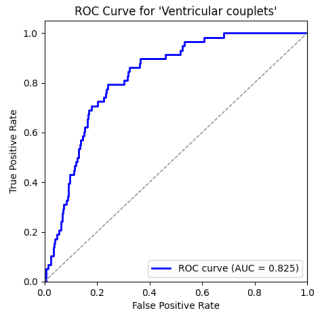

ROC Curve for 'Anteroseptal infarction'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.873)

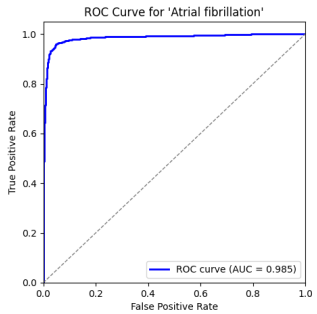

ROC Curve for 'Premature ventricular contraction'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.789)


ROC Curve for 'Frequent premature ventricular contraction
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.902)


ROC Curve for 'Ventricular bigeminy'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.877)


ROC Curve for 'Ventricular tachycardia'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.766)


ROC Curve for 'Ventricular couplets'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.825)


ROC Curve for 'Atrial fibrillation'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.985)

5.5 Evaluation on data from a different medical institution

We performed inference using data from a medical institution different from the one used for training in the paper, in order to examine the degradation in performance caused by differences in data distribution. Note that the dataset from this institution did not include any positive cases for Left Atrial enlargement or Frequent ventricular premature contractions.

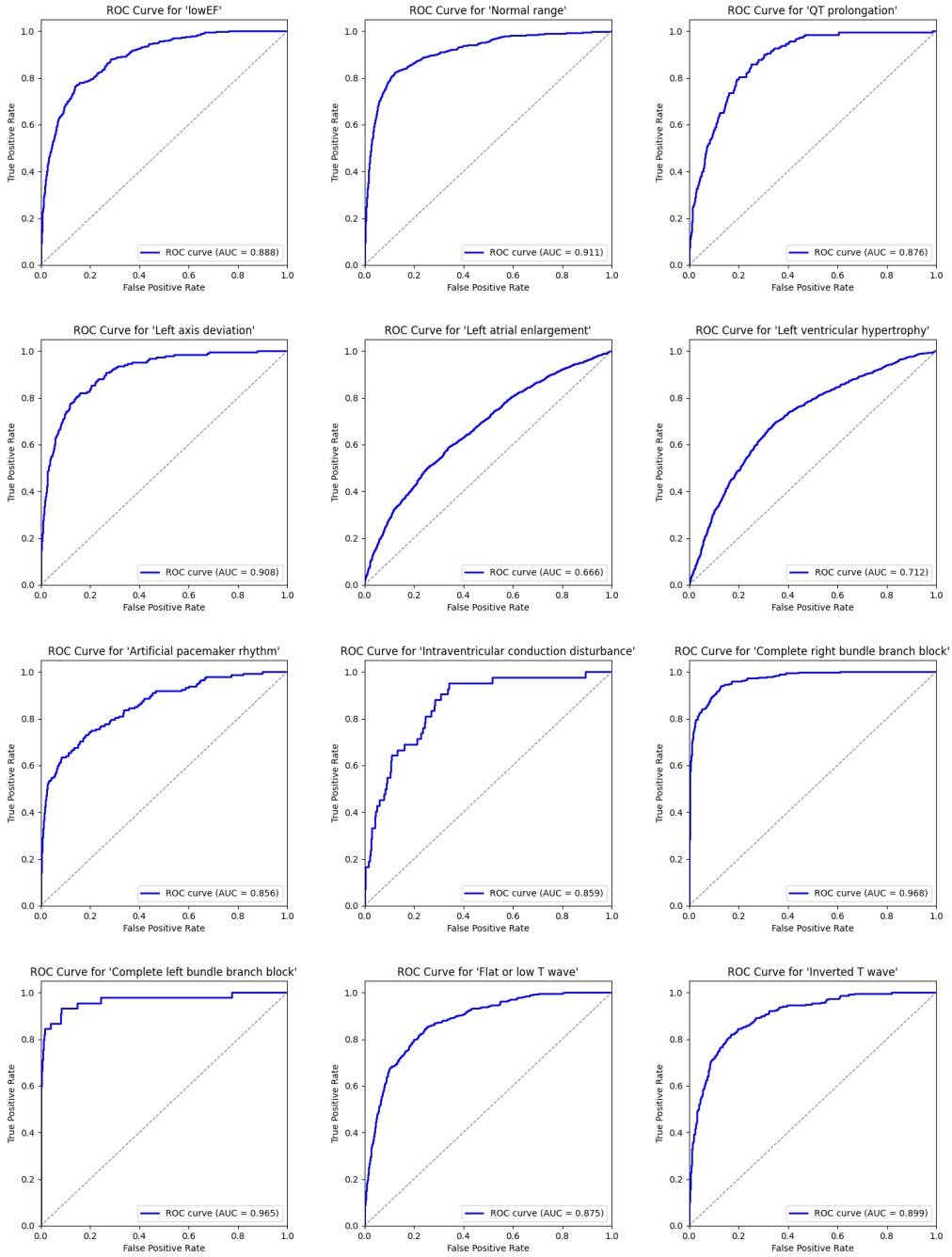
Table 6: Results of the different institutions

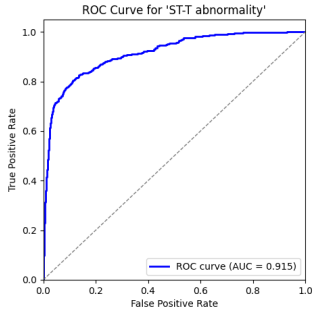
Metric	**Original dataset**	**Different dataset**
Hamming Loss	0.0680↓	0.0536↓
Precision (Micro)	0.4898↑	0.4601↑
Recall (Micro)	0.5165↑	0.5107↑
F1 Score (Micro)	0.5028↑	0.4841↑
Jaccard Index	0.3495↑	0.3360↑

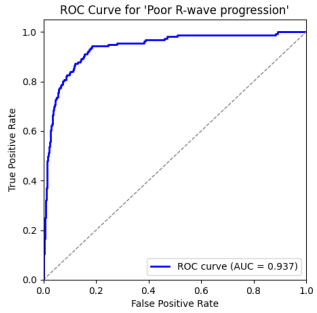
Table 7: Classification performance for each label of different dataset

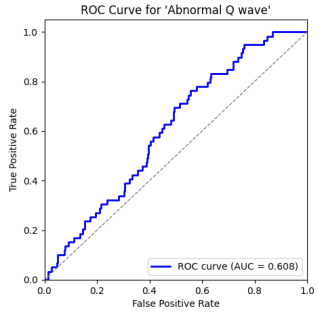
Label	Accuracy	Precision	Recall	F1-score
lowEF	0.9264	0.5483	0.4504	0.4946
Normal	0.8793	0.8056	0.6121	0.6956
Prolonged QT	0.9531	0.3803	0.1421	0.2069
Tall T wave	0.9878	0.1471	0.1667	0.1563
Left axis deviation	0.9443	0.3804	0.5243	0.4409
Left atrial enlargement	0.9110	0.0000	0.0000	0.0000
Left ventricular hypertrophy	0.7525	0.4535	0.3595	0.4011
Artificial pacemaker rhythm	0.9660	0.4909	0.3649	0.4186
Intraventricular conduction delay	0.9724	0.0833	0.1905	0.1159
Complete right bundle branch block	0.9649	0.8281	0.6901	0.7528
Complete left bundle branch block	0.9812	0.3333	0.8444	0.4780
Flat T wave	0.8922	0.5204	0.5141	0.5172
Inverted T wave	0.9420	0.4643	0.4333	0.4483
ST-T abnormality	0.9257	0.7807	0.5848	0.6687
Poor R wave progression	0.9527	0.4007	0.6859	0.5059
Abnormal Q wave	0.9740	0.0172	0.0169	0.0171
Anterior wall myocardial infarction	0.9570	0.0407	0.2188	0.0686
Lateral wall myocardial infarction	0.9581	0.1043	0.3036	0.1553
Inferior wall myocardial infarction	0.9570	0.1477	0.3939	0.2149
Anterior septal myocardial infarction	0.9663	0.1489	0.4200	0.2199
Ventricular premature contraction	0.9567	0.4213	0.4601	0.4399
Frequent ventricular premature contraction	0.9798	0.0000	0.0000	0.0000
Ventricular bigeminy	0.9835	0.0909	0.3158	0.1412
Ventricular tachycardia	0.9703	0.0000	0.0000	0.0000
Coupled ventricular premature contraction	0.9740	0.0364	0.3077	0.0650
Atrial fibrillation	0.9783	0.8721	0.8766	0.8743

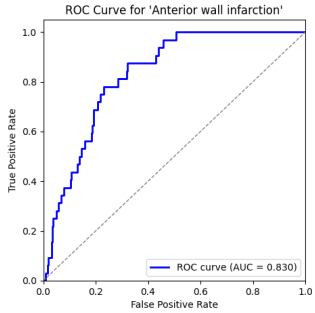
5.6 Appendix: ROC curves of different data

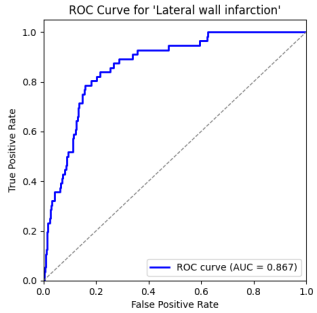


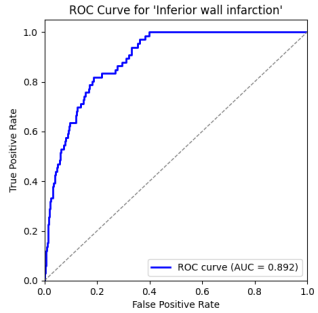

ROC Curve for 'ST-T abnormality'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.915)

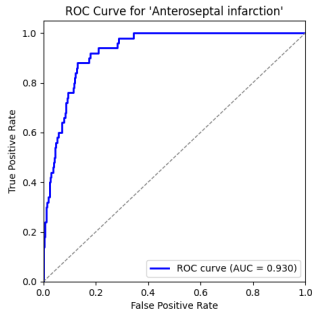

ROC Curve for 'Poor R-wave progression'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.937)

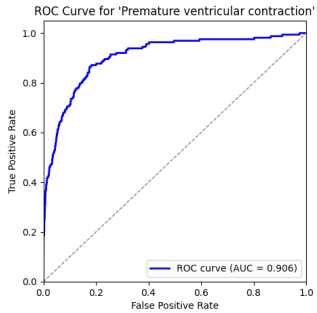

ROC Curve for 'Abnormal Q wave'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.608)

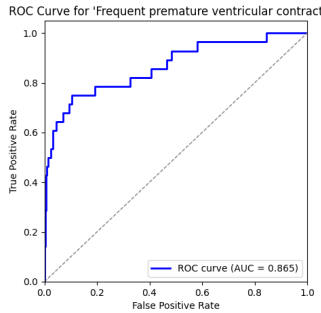

ROC Curve for 'Anterior wall infarction'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.830)

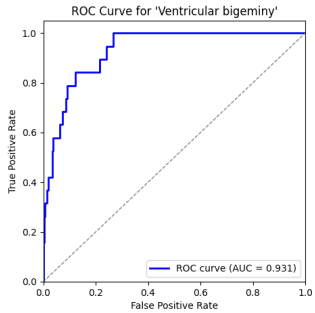

ROC Curve for 'Lateral wall infarction'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.867)

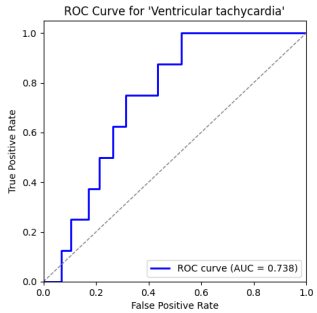

ROC Curve for 'Inferior wall infarction'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.892)

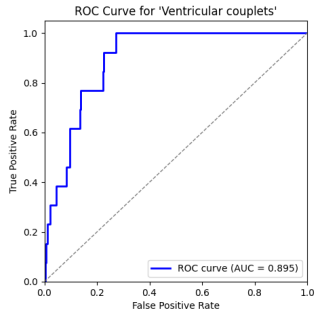

ROC Curve for 'Anteroseptal infarction'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.930)

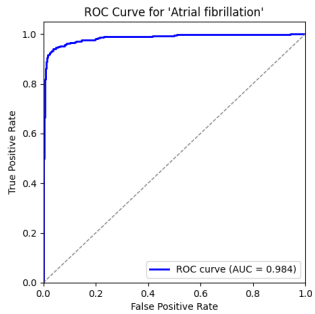

ROC Curve for 'Premature ventricular contraction'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.906)


ROC Curve for 'Frequent premature ventricular contraction
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.865)


ROC Curve for 'Ventricular bigeminy'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.931)


ROC Curve for 'Ventricular tachycardia'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.738)


ROC Curve for 'Ventricular couplets'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.895)


ROC Curve for 'Atrial fibrillation'
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.984)

5.7 Ablation Study: Comparison with a ResNet-Based Multilabel Classification Model

As an ablation study, we used a ResNet-1D model for multilabel prediction. The model was trained for 600 epochs with a learning rate of 1e-4 and a batch size of 32. As described, we applied data augmentation to the training set using random cropping. Since training can become unstable when only a small number of samples are available for certain labels, we weighted the loss according to the label distribution.

Table 8: Performance comparison of baseline and a ResNet-based multi-label model

Metric	**SigLIP Embedding dim 256 + random crop (600 epoch, 20k warmup)**	**ResNet-based multi-label model + random crop**
Hamming Loss	0.0680 ↓	0.1854 ↓
Precision (Micro)	0.4898 ↑	0.2444 ↑
Recall (Micro)	0.5165 ↑	0.8526 ↑
F1 Score (Micro)	0.5028 ↑	0.3799 ↑
Jaccard Index	0.3495 ↑	0.2641 ↑

As shown in Table 8, the proposed SigLIP-based method achieved a higher F1 score than a ResNet-based multi-label classification. This improvement is likely because multimodal training allows the model to leverage features embedded in the text, enabling more accurate inference even with a smaller number of ECG findings.