Application of Contrastive Learning on ECG Data: Evaluating Performance in Japanese and Classification with Around 100 Labels

Junichiro Takahashi JingChuan Guan Masataka Sato Kaito Baba
Kazuto Haruguchi Daichi Nagashima Satoshi Kodera Norihiko Takeda
Department of Cardiovascular Medicine
The University of Tokyo Hospital, Tokyo, Japan
kodera@tke.att.ne.jp

Abstract

The electrocardiogram (ECG) is a fundamental tool in cardiovascular diagnostics due to its powerful and non-invasive nature. One of the most critical usages is to determine whether more detailed examinations are necessary, with users ranging across various levels of expertise. Given this diversity in expertise, it is essential to assist users to avoid critical errors. Recent studies in machine learning have addressed this challenge by extracting valuable information from ECG data. Utilizing language models, these studies have implemented multimodal models aimed at classifying ECGs according to labeled terms. However, the number of classes was reduced, and it remains uncertain whether the technique is effective for languages other than English. To move towards practical application, we utilized ECG data from regular patients visiting hospitals in Japan, maintaining a large number of Japanese labels obtained from actual ECG readings. Using a contrastive learning framework, we found that even with 98 labels for classification, our Japanese-based language model achieves accuracy comparable to previous research. This study extends the applicability of multimodal machine learning frameworks to broader clinical studies and non-English languages.

1 Introduction

Electrocardiograms (ECGs) provide crucial information about the electrical activity of the heart, usually obtained from 12-lead measurement device, and play a significant role in detecting various heart diseases. Due to their simplicity, ECGs have been widely used as a diagnostic tool for many years [6]. They are recorded in a wide range of facilities, from clinics to general hospitals and university hospitals. The results of these ECGs are used by professionals with varying levels of expertise, ranging from cardiologists to non-internal medicine physicians, and even nurses. ECG interpretation is complex because of many observation results, and the results of interpretation could vary significantly depending on the interpreter's level of expertise [14]. Therefore, the development of AI systems to assist in the interpretation of ECGs and bridge the gap in expertise is an important area of research.

There are already studies on medical multimodal AI such as LLaVA-Med [16], which has been developed for healthcare based on language models. This model includes data such as X-ray images but does not yet support ECG which is a type of data composed by 12 time-series. Research on multimodal machine learning models that have learned from ECG data is limited to a few models such as MedGemini [26], and further investigation is needed on how to utilize ECGs in the field of

Foundation Models for Science Workshop, 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

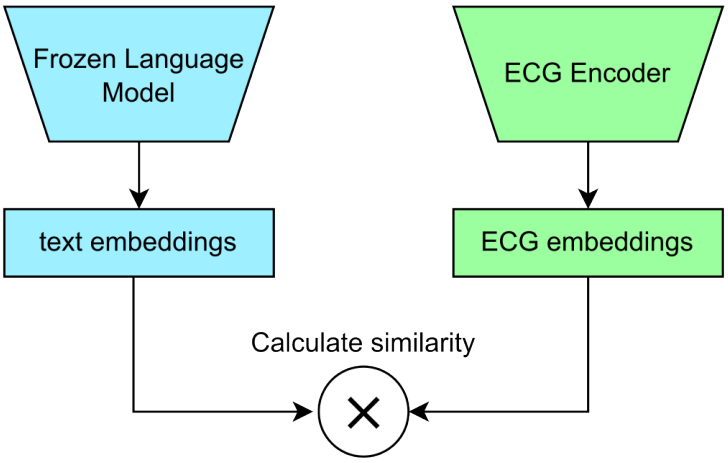

Figure 1: The overall schematics of our model. The encoder of MedLlama3-JP-v2text [28] is employed as the frozen lanuguage model. ResNet1d-18 [9] is adopted as the ECG encoder. The text embeddings and ECG embeddings are denoted as $\mathbf{t}_i$ and $\mathbf{e}_i$, respectively.

machine learning. In particular, how machine learning models can process ECG data is a crucial area of study.

CLIP [24] has acquired knowledge about the relationships between different modalities through pretraining on a large amount of data. Following this, there has been research that conducted pretraining using both ECG and language data [17], allowing for partially zero-shot classification about previously unseen categories [27]. There are also studies that report improved performance by enhancing clinical knowledge in LLMs through reinforcement prompt engineering, utilizing a clinically validated knowledge database created by external experts [19]. Another study reported performance improvements by generating digital twins of ECGs using GANs [7] and extracting ECG features [11]. These studies often simplify the labels to categories such as five classes. However, ECG interpretation in real clinical settings is complex, requiring the accurate reading of a greater number of labels from the ECG. Additionally, the ECG dataset [20, 30] uses English labels, and it is unclear whether the same performance can be achieved in other languages. Therefore, in practical applications, where a wide range of reading results is required and various languages are spoken, we need a machine learning model that can handle more comprehensive labeling and multiple languages.

Aiming for real-world implementation, we constructed a multimodal ECG model leveraging the data obtained from patients who visit Japanese hospitals for usual medical examination. We used enough number of Japanese labels which are utilized in normal hospital works and created by multiple cardiology specialists. The evaluation is conducted through the classification task where partially zero-shot task is included.

2 Method

2.1 Frozen pretrained language models

In the previous study [17], ClinicalBERT [1], pretrained on the MIMIC-III dataset [12] from BioBERT [15], was used as a language model with medical knowledge. In this study, we decided to select a Japanese language model based on two key criteria. The first criterion is that we should select autoregressive models. In the previous study, a BERT model [2] was used for contrastive learning with ECG data. In this study, considering future applications, we trained the ECG encoder using an autoregressive language model, such as GPT [23] or LLaMA [29], in order to integrate the created ECG model into a large multimodal model. We used the last layer of the hidden layers for the language embeddings. The second criterion is that the model should have medical knowledge in Japanese. Since the data used in this study consists of Japanese medical reports, it was essential to use a Japanese medical language model. The language model was selected from among Llama3 [4], MMed-Llama-3 [22] OpenBioLLM [20], MedAlpaca [8], Clinical GPT [31],

and MedLlama3-JP-v2 [28]. To evaluate each LLM, cardiology specialists posed questions related to ECG in Japanese and they assessed the answers. The model judged to have the best performance was MedLlama3-JP-v2. MedLlama3-JP-v2 is a merged model consisting of Llama 3-Swallow [21], OpenBioLLM, MMed-Llama-3, and Llama-3-ELYZA-JP [10]. It has also achieved an accuracy of 46.6% on IgakuQA [13], a Japanese medical QA dataset. We chose MedLlama3-JP-v2 among the 8B models available on Hugging Face due to its superior medical language knowledge in Japanese.

2.2 ECG encoder

We adopted ResNet1d-18 [9] model based on the findings from the previous study [18]. They suggested that ResNet-based models [9] outperform Vision Transformer (**ViT**) [3] in both zero-shot and linear probing tasks, and ResNet models are more effective in capturing ECG patterns.

2.3 Multimodal contrastive learning and classification

We will describe the method for calculating the contrastive loss. Let the batch size be N. The output from the last hidden layer of the language model is referred to as the text embedding $\mathbf{t}$. The output of ResNet1d is referred to as the ECG embedding $\mathbf{e}$. Both $\mathbf{t}$ and $\mathbf{e}$ are processed by linear layers respectively to ensure they have the same embedding dimensions. Under this condition, the contrastive loss is calculated by treating the same pair $(\mathbf{t}_i, \mathbf{e}_i)$ as a positive pair and the different pair $(\mathbf{t}_i, \mathbf{e}_j)$ as a negative pair. The similarity between the two vectors is measured using cosine similarity (**sim**). The cosine similarity between the two vectors is as follows:

$$\text{sim}(\mathbf{t}, \mathbf{e}) = \frac{\mathbf{t}^T \mathbf{e}}{\|\mathbf{t}\| \|\mathbf{e}\|}. \tag{1}$$

The contrastive loss consists of two loss functions. The first loss is the ECG-to-Text contrastive loss.

$$l_i{}^{(e \to t)} = -\log \frac{\exp\left(\text{sim}(\mathbf{t}_i, \mathbf{e}_i)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(\mathbf{t}_i, \mathbf{e}_j)/\tau\right)} \tag{2}$$

τ is initialized to 0.07. The second is the Text-to-ECG contrastive loss.

$$l_i{}^{(t \to e)} = -\log \frac{\exp\left(\text{sim}(\mathbf{e}_i, \mathbf{t}_i)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(\mathbf{e}_i, \mathbf{t}_j)/\tau\right)} \tag{3}$$

Finally, the contrastive loss is calculated as the average combination of the two losses for all positive pairs within a batch.

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

After pre-training, we evaluated the performance of classification tasks. First, We used the prompts similar to the labels used in training. Additionally, referring to the previous research [17], We created a superset of the labels to evaluate zero-shot performance. The correspondence between each observation and the superset is detailed in the Appendix Table 8. We do not conduct the Form test set from [17] because there were no corresponding labels available in our data.

2.4 Data

The data used in this study consists of 37285 ECG records obtained from the University of Tokyo Hospital and Mitsui Hospital. The ECGs were recorded by Fukuda Denshi (Tokyo, JAPAN) equipment. The ECG data is sequence data with a shape of 12×5000. The experiments were conducted using 98 labels contained in the data. The labels were selected by two cardiologists in our team out of the 157 ECG's labels specified by the equipment of Fukuda Denshi. We formatted these specified reports in order to make training prompts in Japanese as "この心電図から{reports}が認められます." ("This ECG shows {reports}."). To avoid data leakage, ECG data from the same patients were not present across any two data splits.At the test phase evaluating the zero-shot performance, we used the same labels described in the prior research [17], that are **Supersetdiagnosis labels**, **Rhythm labels**, and **MIT-BIT labels**.

2.5 Implementation details

In this study, we used Hugging Face library. The learning rate was set to 1×10^{-3}, weight decay was set to 1×10^{-3}, and the global batch size was 32. We trained our model over 200 epochs. Other hyperparameters related to training were set to the default values of the Hugging Face Trainer. Training was conducted using two NVIDIA A100-SXM4-80GB GPUs.

3 Result

After pretraining, we firstly evaluated the performance of the classification task by using the ECG reports in the test data as the ground truth labels. We calculated accuracy for both top-1 and top-5 predictions. The results with the top 5 scores are listed in Table 1. The overall results and results of the individual labels are detailed in the Appendix Table 9. The results for each label suggest that

Table 1: Results with the top 5 scores (excluding results with fewer than 10 labels)

Results with the top 5 scores		
Labels	Top-1 Accuracy	Top-5 Accuracy
Pacemaker Rhythm	89.41%	93.73%
Left Anterior Fascicular Block	88.00%	88.00%
Normal	78.40%	90.45%
Ventricular Couplet	77.78%	77.78%
Ventricular Bigeminy	76.92%	84.62%

our model could correctly identify normal ECGs (normal range) with high accuracy and accurately detect pacemaker rhythms (Artificial Pacemaker Rhythm). However, it struggled to interpret the reports quantified from ECG waveforms, such as Prolonged PR Interval and Prolonged QT Interval. The labels related to "Short Run of Supraventricular Premature Contractions" and "Myocardial Infarction" show a significant gap between the top-1 and top-5 accuracy. For these labels, we examined the top-5 prediction results. The result is Output 2. The outputs were originally in Japanese.

label: Short Run of Supraventricular Premature Contractions
predict: This ECG shows Ventricular Premature Contractions Couplets.
predict: This ECG shows Frequent Supraventricular Premature Contractions.
predict: This ECG shows Supraventricular Bigeminy.
predict: This ECG shows Supraventricular Premature Contractions.
predict: This ECG shows Short Run of Supraventricular Premature Contractions.

label: Suspected Inferior Wall Infarction
predict: This ECG shows Suspected Inferior Wall Infarction.
predict: This ECG shows Suspected Anterior Wall Infarction.
predict: This ECG shows Suspected Lateral Wall Infarction.
predict: This ECG shows Suspected High Posterior Wall Infarction.
predict: This ECG shows Suspected Acute Inferior Wall Infarction.

Output 2 The examples of the outputs of diagnosis predictions

From this output, it can be inferred that even if the top-1 prediction does not accurately identify the label, the model is still capable of detecting the ECG reports to some extent. In the first case, based on the top 5 outputs, the model appears to have the capability to detect the events at superior ventricles. In the second case, the model can detect myocardial infarction. This proposes that, although the predictions is not correct, the pretrained model seems to understand some contents of the ECG reports.

Second, referencing the prior study [17], we created the superset labels: **Superclass diagnosis**, **Rhythm**, and **MIT-BIH** and then evaluated the zero-shot performance. The results are in Table 2, Table 3, Table 4.

Table 2 Superclass diagnosis result

Superclass diagnosis result	
Labels	Accuracy
all	64.11%
Normal ECG	81.53%
Conduction Disturbance	89.15%
Mycardinal Infarction	63.55%
Hypertrophy	42.35%
ST/T change	45.39%

Table 3 Rhythm result

Rhythm result	
Labels	Accuracy
all	78.88%
Sinus rhythm	95.31%
Atrial fibrillation	67.09%
Sinus tachycardia	78.05%
Sinus arrhythmia	43.37%
Sinus bradycardia	63.46%

Table 4 MIT-BIH diagnostic result

MIT-BIH diagnostic result	
Labels	Accuracy
all	76.43%
Normal beat	94.52%
Left bundle branch block beat	87.77%
Right bundle branch block beat	58.64%
Atrial premature beat	41.30%
Premature ventricular contraction	75.26%

Table 3 and 4 show that the performance of our model approaches the previous study [17] on the Rhythm test and MIT-BIH diagnostic test set. In the previous study [17], the model recorded an accuracy of 74.60% for Rhythm test and 79.40% for MIT-BIH test. This indicates that the contrastive learning method is generally effective for Japanese clinical reports as well. From the superclass diagnosis test in Table 2, the model developed for this study shows shortcomings in the reports related to hypertrophy. One reason for this result is that diagnosing hypertrophy typically requires confirmation through echocardiography and while there are criteria for evaluating hypertrophy from an ECG, their sensitivity is relatively low [25]. This issue implies the importance of training with echocardiographic data.

Based on the results of this study, there was significant variability in accuracy depending on the labels. For the labels with low accuracy, further improvement in the ECG interpretation capabilities is essential. On the other hand, we consider that linking ECG with other information for training, such as echocardiographic data, is also important and we are planning to implement this approach in the future.

In addition, ablation study was conducted. we evaluated the classification performance of Superset labels using the Swallow model [5, 21], which has not trained for medical purposes. The results are shown in Tables 5, 6, and 7.

In Tables 5, 6, and 7, The overall accuracy is lower than the value achieved with MedLlama3-JP-v2. This suggests that medical knowledge within the language model contributes to learning the relationship between medical texts and ECG data. For some conditions, like ST/T change and atrial premature beat, accuracy drops to around 30%. Swallow, used in the ablation study, lacked knowledge about these conditions and produced hallucinations. However, for conditions like hypertrophy, the accuracy is higher than that with MedLlama3-JP-v2, which indicates the necessity for further investigation about these specific cases.

4 Conclusion

To assist physicians who read ECG data in the field of healthcare, we have built a ECG-specific CLIP model that interprets ECG data. Incorporating contrastive learning, a multimodal model has been constructed using ECG data and Japanese medical reports. During the training, we adopted a medical language model with frozen parameters and found that contrastive learning between ECG and text can effectively learn the correspondence between ECG and text in Japanese, and can also recognize detailed reports. This suggests that pretraining with ECG data and medical reports can efficiently extract semantic ECG features across multiple languages. The machine learning model

Table 5 Normal Swallow Superclass diagnosis result

Superclass diagnosis result	
Labels	Accuracy
all	60.49%
Normal ECG	66.82%
Conduction Disturbance	86.28%
Mycardinal Infarction	68.60%
Hypertrophy	51.70%
ST/T change	35.66%

Table 6 Normal Swallow Rhythm result

Rhythm result	
Labels	Accuracy
all	71.47%
Sinus rhythm	83.72%
Atrial fibrillation	60.76%
Sinus tachycardia	71.95%
Sinus arrhythmia	46.99%
Sinus bradycardia	63.46%

Table 7 Normal Swallow MIT-BIH diagnostic result

MIT-BIH diagnostic result	
Labels	Accuracy
all	73.76%
Normal beat	89.36%
Left bundle branch block beat	82.01%
Right bundle branch block beat	59.19%
Atrial premature beat	28.26%
Premature ventricular contraction	78.87%

that interprets ECG is expected to be applied in broader ways other than assisting users engaging in the field of healthcare. For example, a representative one is wearable device which measures the human electrical signals in daily lives. The device could be used by everyone to detect the signs and prevent diseases. By developing the approach used in our study, we hope the result will contribute to those downstream applications.

5 Acknowledgements

This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on " Integrated Health Care System " Grant Number JPJ012425.

References

[1] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[2] J. Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[5] K. Fujii, T. Nakamura, M. Loem, H. Iida, M. Ohi, K. Hattori, H. Shota, S. Mizuki, R. Yokota, and N. Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, page (to appear), University of Pennsylvania, USA, Oct. 2024.

[6] W. B. Fye. A history of the origin, evolution, and impact of electrocardiography. *The American journal of cardiology*, 73(13):937–949, 1994.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[8] T. Han, L. C. Adams, J.-M. Papaioannou, P. Grundmann, T. Oberhauser, A. Löser, D. Truhn, and K. K. Bressem. Medalpaca – an open-source collection of medical conversational ai models and training data, 2023.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] M. Hirakawa, S. Horie, T. Nakamura, D. Oba, S. Passaglia, and A. Sasaki. elyza/llama-3-elyza-jp-8b, 2024.

[11] Y. Hu, J. Chen, L. Hu, D. Li, J. Yan, H. Ying, H. Liang, and J. Wu. Personalized heart disease detection via ecg digital twin generation. *arXiv preprint arXiv:2404.11171*, 2024.

[12] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[13] J. Kasai, Y. Kasai, K. Sakaguchi, Y. Yamada, and D. Radev. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*, 2023.

[14] A. H. Kashou, P. A. Noseworthy, T. J. Beckman, N. S. Anavekar, M. W. Cullen, K. B. Angstman, B. J. Sandefur, B. P. Shapiro, B. W. Wiley, A. M. Kates, et al. Ecg interpretation proficiency of healthcare professionals. *Current problems in cardiology*, 48(10):101924, 2023.

[15] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[16] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] J. Li, C. Liu, S. Cheng, R. Arcucci, and S. Hong. Frozen language model helps ecg zero-shot learning. In *Medical Imaging with Deep Learning*, pages 402–415. PMLR, 2024.

[18] C. Liu, Z. Wan, S. Cheng, M. Zhang, and R. Arcucci. Etp: Learning transferable ecg representations via ecg-text pre-training. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8230–8234. IEEE, 2024.

[19] C. Liu, Z. Wan, C. Ouyang, A. Shah, W. Bai, and R. Arcucci. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. *arXiv preprint arXiv:2403.06659*, 2024.

[20] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

[21] N. Okazaki, K. Hattori, H. Shota, H. Iida, M. Ohi, K. Fujii, T. Nakamura, M. Loem, R. Yokota, and S. Mizuki. Building a large japanese web corpus for large language models. In *Proceedings of the First Conference on Language Modeling*, COLM, page (to appear), University of Pennsylvania, USA, Oct. 2024.

[22] P. Qiu, C. Wu, X. Zhang, W. Lin, H. Wang, Y. Zhang, Y. Wang, and W. Xie. Towards building multilingual language model for medicine. *arXiv preprint arXiv:2402.13963*, 2024.

[23] A. Radford. Improving language understanding by generative pre-training. 2018.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[25] N. Reichek and R. B. Devereux. Left ventricular hypertrophy: relationship of anatomic, echocardiographic and electrocardiographic findings. *Circulation*, 63(6):1391–1398, 1981.

[26] K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

[27] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26, 2013.

[28] I. Sukeda. Eques/medllama3-jp-v2, 2024.

[29] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[30] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

[31] G. Wang, G. Yang, Z. Du, L. Fan, and X. Li. Clinicalgpt: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*, 2023.

6 Appendix

Table 8 Mapping between ECG Labels and Zero-Shot Labels

Diagnosis	**Superclass Diagnosis**	**Rhythm**	**MIT-BIH**
Sinus Tachycardia			Sinus Tachycardia
Short Run of Supraventricular Premature Contractions		Sinus Arrhythmia	Atrial premature beat
Pacemaker Rhythm	Conduction Disturbance		
Short PR Interval			
Flat T Wave		ST/T change	
Severe Tachycardia			
Borderline Q Wave	Mycardinal Infarction		
Sinus Arrhythmia			Sinus Arrhythmia
Mild ST-T Abnormality		ST/T change	
Negative T Wave		ST/T change	
Prolonged PR Interval	Conduction Disturbance		
ST-T Abnormality		ST/T change	
Suspected Inferior Wall Infarction	Mycardinal Infarction		
Complete Right Bundle Branch Block	Conduction Disturbance		Right bundle branch block beat
Left Ventricular Hypertrophy with Left Atrial Enlargement	Hypertrophy		
Possible Inferior Wall Infarction	Mycardinal Infarction		
Frequent Ventricular Premature Contractions			Premature ventricular contraction
Inferior Wall Infarction	Mycardinal Infarction		
Supraventricular Premature Contractions		Sinus Arrhythmia	Atrial premature beat

Continued on next page

Continued from previous page

Diagnosis	**Superclass Diagnosis**	**Rhythm**	**MIT-BIH**
Second-degree Atrioventricular Block (Wenckebach)	Conduction Disturbance		
Bradycardia			
First-degree Atrioventricular Block	Conduction Disturbance		
Suspected Anteroseptal Infarction	Mycardinal Infarction		
Severe Bradycardia	Mycardinal Infarction		
Atrial Fibrillation			Atrial fibrillation
Poor R Wave Progression	Mycardinal Infarction		
Left Ventricular Hypertrophy	Hypertrophy		
Incomplete Right Bundle Branch Block	Conduction Disturbance		Right bundle branch block beat
Abnormal Q Wave	Mycardinal Infarction		
Intraventricular Conduction Delay	Conduction Disturbance		
Ventricular Premature Contractions			Premature ventricular contraction
Suspected Left Anterior Fascicular Block	Conduction Disturbance		Left bundle branch block beat
Anteroseptal Infarction	Mycardinal Infarction		
Sinus Bradycardia			Sinus Bradycardia
Complete Left Bundle Branch Block	Conduction Disturbance		Left bundle branch block beat
Mild Left Ventricular Hypertrophy with Left Atrial Enlargement	Hypertrophy		
Supraventricular Tachycardia		Sinus Arrhythmia	Atrial premature beat
RSR' Pattern	Conduction Disturbance		Right bundle branch block beat
Suspected Lateral Wall Infarction	Mycardinal Infarction		
Suspected Anterior Wall Infarction	Mycardinal Infarction		
Lateral Wall Infarction	Mycardinal Infarction		
Tachycardia			
Suspected Mild ST-T Abnormality		ST/T change	
Left Anterior Fascicular Block	Conduction Disturbance		Left bundle branch block beat
Atrial Flutter		Sinus Arrhythmia	
Suspected High Posterior Wall Infarction	Mycardinal Infarction		
Left Atrial Enlargement			
Suspected Acute Inferior Wall Infarction	Mycardinal Infarction		
Possible Lateral Wall Infarction	Mycardinal Infarction		

Continued on next page

Continued from previous page

Diagnosis	**Superclass Diagnosis**	**Rhythm**	**MIT-BIH**
Anterior Wall Infarction	Mycardinal Infarction		
Mild Left Axis Deviation			
High Voltage (Leads Corresponding to Left Ventricle)	Hypertrophy		
Frequent Supraventricular Premature Contractions		Sinus Arrhythmia	Premature ventricular contraction
Right Axis Deviation			
Left Axis Deviation			
Possible Anteroseptal Infarction	Mycardinal Infarction		
Left Posterior Fascicular Block	Conduction Disturbance		Left bundle branch block beat
Supraventricular Trigeminy		Sinus Arrhythmia	Atrial premature beat
Biventricular Hypertrophy	Hypertrophy		
Prolonged QT Interval			
Mild Left Ventricular Hypertrophy	Hypertrophy		
Acute Anterior Wall Infarction	Mycardinal Infarction		
Low Voltage (Limb Leads)			
Severe Right Axis Deviation			
Ventricular Couplet			Premature ventricular contraction
Subacute Anteroseptal Infarction	Mycardinal Infarction		
Right Atrial Enlargement			
Mild Right Ventricular Hypertrophy	Hypertrophy		
Normal	Normal ECG		Normal Beat
Clockwise Rotation			
Counterclockwise Rotation			
Right Ventricular Hypertrophy	Hypertrophy		
Ventricular Rhythm			
T-wave Elevation	ST/T change		
S1, S2, S3 Pattern			
Mild ST Elevation	ST/T change		
Ventricular Bigeminy			Premature ventricular contraction
Possible Anterior Wall Infarction	Mycardinal Infarction		
Ventricular Tachycardia			
Sinoatrial Block	Mycardinal Infarction		
Indeterminate Arrhythmia			
Subacute Anterior Wall Infarction	Mycardinal Infarction		
Subacute Lateral Wall Infarction	Mycardinal Infarction		
Subacute Inferior Wall Infarction	Mycardinal Infarction		

Continued on next page

Continued from previous page

Diagnosis	**Superclass Diagnosis**	**Rhythm**	**MIT-BIH**
Mild Right Ventricular Hypertrophy with Left Atrial Enlargement	Hypertrophy		
Right Ventricular Hypertrophy with Right Atrial Enlargement	Hypertrophy		
Low Voltage (Chest Leads)			
Second-degree Atrioventricular Block (Mobitz)	Conduction Disturbance		
Mild Right Ventricular Hypertrophy with Right Atrial Enlargement	Hypertrophy		
Right Ventricular Hypertrophy with Left Atrial Enlargement	Hypertrophy		
Suspected Acute Lateral Wall Infarction	Mycardinal Infarction		
Ventricular Premature Contractions Couplets		Sinus Arrhythmia	Atrial premature beat
Ventricular Trigeminy			Premature ventricular contraction
Supraventricular Bigeminy		Sinus Arrhythmia	Premature ventricular contraction
Complete Atrioventricular Block	Conduction Disturbance		
Possible High Posterior Wall Infarction	Mycardinal Infarction		
Acute Lateral Wall Infarction	Mycardinal Infarction		
Suspected Acute Anterior Wall Infarction	Mycardinal Infarction		

Table 9 Top-1 and Top-5 Accuracy for Various Diagnoses

Diagnosis (Data Counts)	Top-1 Accuracy	Top-5 Accuracy
all labels (7710)	35.91%	44.80%
Sinus Tachycardia (82)	73.17%	75.61%
Short Run of Supraventricular Premature Contractions (4)	0.00%	75.00%
Pacemaker Rhythm (255)	89.41%	93.73%
Short PR Interval (62)	38.71%	53.23%
Flat T Wave (444)	22.75%	32.66%
Severe Tachycardia (42)	61.90%	61.90%
Borderline Q Wave (123)	18.70%	26.83%
Sinus Arrhythmia (63)	1.59%	4.76%
Mild ST-T Abnormality (278)	16.91%	28.06%
Negative T Wave (258)	24.81%	34.11%
Prolonged PR Interval (150)	40.67%	47.33%
ST-T Abnormality (501)	36.13%	44.31%
Suspected Inferior Wall Infarction (50)	16.00%	42.00%
Complete Right Bundle Branch Block (278)	58.27%	63.31%
Left Ventricular Hypertrophy with Left Atrial Enlargement (22)	18.18%	36.36%
Possible Inferior Wall Infarction (77)	23.38%	35.06%
Frequent Ventricular Premature Contractions (22)	68.18%	86.36%
Inferior Wall Infarction (61)	42.62%	50.82%

Continued on next page

Continued from previous page

Diagnosis (Data Counts)	Top-1 Accuracy	Top-5 Accuracy
Supraventricular Premature Contractions (80)	13.75%	22.50%
Second-degree Atrioventricular Block (Wenckebach) (1)	0.00%	100.00%
Bradycardia (12)	8.33%	16.67%
First-degree Atrioventricular Block (81)	48.15%	58.02%
Suspected Anteroseptal Infarction (46)	36.96%	58.70%
Severe Bradycardia (5)	40.00%	40.00%
Atrial Fibrillation (316)	57.59%	64.56%
Poor R Wave Progression (146)	32.19%	39.73%
Left Ventricular Hypertrophy (296)	13.18%	25.68%
Incomplete Right Bundle Branch Block (191)	35.60%	40.31%
Abnormal Q Wave (51)	3.92%	7.04%
Intraventricular Conduction Delay (75)	28.00%	30.67%
Ventricular Premature Contractions (156)	10.26%	21.79%
Suspected Left Anterior Fascicular Block (53)	69.81%	86.79%
Anteroseptal Infarction (63)	52.38%	53.97%
Sinus Bradycardia (52)	57.69%	61.54%
Complete Left Bundle Branch Block (59)	59.32%	83.05%
Mild Left Ventricular Hypertrophy with Left Atrial Enlargement (4)	25.00%	25.00%
Supraventricular Tachycardia (3)	0.00%	0.00%
RSR' Pattern (75)	24.00%	41.33%
Suspected Lateral Wall Infarction (28)	0.00%	3.57%
Suspected Anterior Wall Infarction (53)	7.55%	28.30%
Lateral Wall Infarction (79)	27.85%	27.85%
Tachycardia (54)	0.00%	5.56%
Suspected Mild ST-T Abnormality (40)	15.00%	30.00%
Left Anterior Fascicular Block (25)	88.00%	88.00%
Atrial Flutter (6)	16.67%	33.33%
Suspected High Posterior Wall Infarction (5)	0.00%	0.00%
Left Atrial Enlargement (174)	6.90%	17.82%
Suspected Acute Inferior Wall Infarction (5)	0.00%	0.00%
Possible Lateral Wall Infarction (40)	0.00%	0.00%
Anterior Wall Infarction (57)	24.56%	24.56%
Mild Left Axis Deviation (293)	42.66%	59.39%
High Voltage (Leads Corresponding to Left Ventricle) (126)	22.22%	44.44%
Frequent Supraventricular Premature Contractions (3)	33.33%	33.33%
Right Axis Deviation (209)	25.36%	33.01%
Left Axis Deviation (147)	18.37%	30.61%
Possible Anteroseptal Infarction (5)	0.00%	20.00%
Left Posterior Fascicular Block (2)	0.00%	0.00%
Supraventricular Trigeminy (2)	0.00%	0.00%
Biventricular Hypertrophy (17)	0.00%	0.00%
Prolonged QT Interval (214)	12.62%	18.22%
Mild Left Ventricular Hypertrophy (55)	21.82%	30.91%
Acute Anterior Wall Infarction (4)	0.00%	0.00%
Low Voltage (Limb Leads) (140)	53.57%	54.29%
Severe Right Axis Deviation (48)	18.75%	18.75%
Ventricular Couplet (18)	77.78%	77.78%
Subacute Anteroseptal Infarction (3)	0.00%	0.00%
Right Atrial Enlargement (73)	16.44%	26.03%
Mild Right Ventricular Hypertrophy (39)	0.00%	5.13%
Normal (639)	78.40%	90.45%
Clockwise Rotation (187)	9.63%	11.23%
Counterclockwise Rotation (203)	41.38%	48.77%
Right Ventricular Hypertrophy (10)	0.00%	10.00%
Ventricular Rhythm (1)	0.00%	0.00%

Continued on next page

Continued from previous page

Diagnosis (Data Counts)	Top-1 Accuracy	Top-5 Accuracy
T-wave Elevation (20)	45.00%	55.00%
S1, S2, S3 Pattern (25)	20.00%	24.00%
Mild ST Elevation (21)	9.52%	14.29%
Ventricular Bigeminy (13)	76.92%	84.62%
Possible Anterior Wall Infarction (12)	0.00%	0.00%
Ventricular Tachycardia (3)	0.00%	33.33%
Sinoatrial Block (1)	0.00%	0.00%
Indeterminate Arrhythmia (4)	0.00%	0.00%
Subacute Anterior Wall Infarction (3)	33.33%	33.33%
Subacute Lateral Wall Infarction (4)	75.00%	75.00%
Subacute Inferior Wall Infarction (8)	62.50%	62.50%
Mild Right Ventricular Hypertrophy with Left Atrial Enlargement (11)	36.36%	45.45%
Right Ventricular Hypertrophy with Right Atrial Enlargement (2)	100.00%	100.00%
Low Voltage (Chest Leads) (19)	31.58%	31.58%
Second-degree Atrioventricular Block (Mobitz) (1)	0.00%	0.00%
Mild Right Ventricular Hypertrophy with Right Atrial Enlargement (2)	100.00%	100.00%
Right Ventricular Hypertrophy with Left Atrial Enlargement (4)	75.00%	75.00%
Suspected Acute Lateral Wall Infarction	0.00%	0.00%
Ventricular Premature Contractions Couplets (2)	66.67%	66.67%
Ventricular Trigeminy (2)	0.00%	100.00%
Supraventricular Bigeminy (2)	0.00%	50.00%
Complete Atrioventricular Block (2)	0.00%	0.00%
Possible High Posterior Wall Infarction (2)	0.00%	0.00%
Acute Lateral Wall Infarction (1)	0.00%	0.00%
Suspected Acute Anterior Wall Infarction (2)	0.00%	0.00%

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This study showed the effectiveness of contrastive learning for ECG and text in both Japanese and many labels settings.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: It was shown that there are differences compared to previous studies in some ECG reports. It was also noted that training with different data is necessary for ECG reports such as hypertrophy.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [No]

Justification: Some reports lacked sufficient labels, which prevented adequate validation.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: Since the data used is from within a hospital, validation with open data remains necessary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The experiments were conducted using closed data from within the hospital.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The libraries and hyperparameters used in the experiments are described in the Methods section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This study is not focused on testing statistical significance but rather on evaluating Japanese language capability and many labels capabilities.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Information about the GPUs used in the experiments is provided in the Methods section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Detailed information about the models and training methods used in the experiments has been provided to facilitate the possibility of following experiments.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The conclusion includes implications for application to multimodal medical AI.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The paper uses the medical data of our hospital with considerations for privacy.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The source of the data has been added to the Methods section.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Details of the experiments have been included in the Methods section, and results from experiments with many labels have been documented in the Appendix.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This study did not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This study do not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.