Enhancing Explainability of Deep Learning-Based ECG Diagnosis Using Large Language Models

Shi Wu
Data Science Institute
University of Technology Sydney
Sydney, NSW, Australia
Wu.Shi@uts.edu.au

Jianlong Zhou
Data Science Institute
University of Technology Sydney
Sydney, NSW, Australia
jianlong.zhou@uts.edu.au

Yifei Dong
Data Science Institute
University of Technology Sydney
Sydney, NSW, Australia
yifei.dong@uts.edu.au

Fang Chen
Data Science Institute
University of Technology Sydney
Sydney, NSW, Australia
fang.chen@uts.edu.au

Abstract

The electrocardiogram (ECG) is an essential diagnostic tool for monitoring heart health. Traditional manual methods for ECG interpretation are increasingly challenged by the complexity of heart diseases and the volume of ECG data. Recent advancements in artificial intelligence (AI), particularly deep learning, have improved the efficiency and accuracy of ECG diagnostics. However, the "black-box" nature of AI models poses trust and verification challenges. To address this, we propose a novel approach integrating Explainable AI (XAI) techniques with multimodal large language models to enhance ECG diagnostic interpretability. Our methodology employs a 2D convolutional neural network (CNN) to classify ECG signals, followed by Grad-CAM to generate heatmaps highlighting critical areas influencing AI decisions. These enhanced ECG images and their classifications are then input into a multimodal large language model to produce comprehensive explanatory outputs. This approach combines the visual processing power of CNNs with the contextual understanding of multimodal models, offering clearer insights into AI reasoning. Our research demonstrates improved diagnostic accuracy and interpretability, setting a new standard for AI integration in clinical practices. Specifically, our method achieved an F1-score of 0.67 for ECG classification using Inception and a BERT-Score of 0.818 for text generation using Gemini_GradCAM.

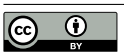
CCS Concepts

• **Computing methodologies** → *Natural language generation.*

Keywords

Electrocardiogram (ECG), Explainable AI (XAI), Multimodal Large Language Models (MLLMs), Grad-CAM

ACM Reference Format:
Shi Wu, Jianlong Zhou, Yifei Dong, and Fang Chen. 2024. Enhancing Explainability of Deep Learning-Based ECG Diagnosis Using Large Language Models. In *2024 The 8th International Conference on Advances in Artificial Intelligence (ICAAI 2024), October 17–19, 2024, London, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3704137.3704146



ICAAI 2024, London, United Kingdom

ACM ISBN 979-8-4007-1801-4/24/10
https://doi.org/10.1145/3704137.3704146

1 Introduction

The electrocardiogram (ECG) is a vital tool for monitoring heart health, aiding in the detection of arrhythmias, heart enlargement, ischemia, and inflammation [11]. Traditionally, interpreting ECGs has been a manual, time-consuming process requiring significant expertise. With the increasing complexity of heart diseases and the volume of ECG data, this manual approach struggles with scalability and speed. Recent advancements in artificial intelligence (AI), particularly deep learning, have significantly enhanced the efficiency and accuracy of ECG diagnostics [22]. AI models, especially deep convolutional neural networks (CNN), can autonomously analyze ECG data, matching the diagnostic accuracy of experienced cardiologists [12]. However, the adoption of such AI systems in clinical practice is hindered by their opaque decision-making processes, often referred to as the “black-box" problem, which makes it difficult for practitioners to trust and verify the AI’s decisions [1]. To address this, Explainable AI (XAI) techniques are being integrated, making AI decision processes transparent and understandable for healthcare professionals. Despite these advancements, the effectiveness of visual explanations in specialized fields like ECG diagnostics depends heavily on domain-specific knowledge. Non-specialists may struggle to grasp the significance of AI-highlighted ECG areas. Additionally, using multimodal large language models (LLMs) for image classification and diagnosis presents challenges, as these models excel in natural language processing but may falter with direct image classification [19].

To address these challenges, we propose a new approach using multimodal large language models to interpret deep learning model outputs: A 2D CNN model classifies ECG signals, and Grad-CAM produces heatmaps to highlight critical ECG areas influencing the AI’s decision. These enhanced ECGs are input into a multimodal model, generating comprehensive explanations. Specifically, we make the following contributions:

- Unlike previous works that rely solely on CNNs or LLMs, our method combines the visual processing power of CNNs with the contextual understanding of multimodal models. A 2D

CNN model to classify preprocessed ECG images, ensuring diagnostic accuracy and avoiding the unreliability caused by the hallucinations of large models, while the multimodal LLMs generate detailed textual explanations, leveraging their understanding of medical terminology. This integration addresses the limitations of standalone models by providing both accurate diagnostics and comprehensive explanations.
- We uniquely apply Grad-CAM to ECG diagnostics and input the enhanced images into multimodal models to generate textual explanations. This dual-modality approach overcomes the limitations of standalone Grad-CAM, which needs domain knowledge and lacks textual explanations, by offering clear, comprehensive interpretations.
- By using ECG images instead of raw signals, our methodology aligns more closely with real-world clinical practices where ECGs are often stored as images. This practical alignment enhances the applicability and adoption potential of our framework in clinical settings.

2 Related Work

Deep learning's high efficiency in complex tasks such as ECG diagnosis classification stems from their ability to autonomously extract and utilize high-level abstract features from data. However, this automatic extraction of high-level features also complicates the models' decision-making explanation. For example, Gradient-weighted Class Activation Mapping (Grad-CAM) visualizes feature maps in the last convolutional layers, highlighting signal parts focused on for classification decisions [14]. Attention mechanisms integrate into models, spotlighting influential heartbeats or waveform segments [20]. Local Interpretable Model-agnostic Explanations (LIME) create interpretable local models around complex decision boundaries, identifying decisive input features [13]. SHapley Additive exPlanations (SHAP) quantify input features' impact on predictions, based on cooperative game theory [8]. These methods adapt well to deep learning's complexity, offering intuitive, flexible explanations across various architectures, significantly aiding in ECG analysis and enhancing trust and accuracy in medical diagnostics.

Furthermore, recent research tries to convert complex medical signals into medical reports with LLMs. For instance, BioSignal Copilot [7], an innovative signal-to-text (Sig2Txt) engine designed to convert medical signals into technical or clinical reports. Yu et al. [23] propose a zero-shot, retrieval-augmented approach utilizing advanced language models like LLaMA2 and GPT-3.5 to generate diagnostic results for medical conditions such as arrhythmias and sleep apnea without the need for training samples. However, the aforementioned method heavily relies on domain knowledge, requiring extensive manual feature extraction and knowledge base construction. Multimodal large models can directly input images, allowing the model to automatically capture important features and generate text in an end-to-end manner [19]. But this kind of approach often leads to hallucinations and is not suitable for direct use in the medical field due to the risk of inaccurate interpretations and diagnoses.

In this paper, we aim to integrate machine learning explanation techniques into LLMs for achieving both the interpretability and accuracy of the diagnostic reports.

3 Methodology

In this study, we explored the use of the raw ECG signals of 12 ECG leads and the ECG image of 12 ECG leads separately along with prompt sentences to classify ECG data with a multimodal large language model. It was found that the use of ECG image data resulted in better classification performance. Therefore, this study uses a separate CNN vision model for ECG classifications. Based on this, we compared two methods: inputting the original ECG 12 lead images and classification labels into the multimodal large language model, and using Grad-CAM to mark important features, followed by the multimodal large language model performing image captioning on the marked images. This comparison aimed to assess which approach better aligns image features with textual explanations, thereby enhancing the interpretability and accuracy of the diagnostic reports.

3.1 Data and Preprocessing

In this study we use the PTB-XL dataset [21] because it encompasses both ECG data in the form of 12-lead recordings and comprehensive diagnostic information. 12-lead ECG images instead of raw data are used because of two primary reasons: (1) in real clinical practice, ECGs are commonly recorded and stored as images rather than raw signal data. Therefore, employing images aligns the research methodology with practical clinical scenarios, ensuring the findings are directly applicable without additional data processing steps; (2) it has been demonstrated in previous research that training models using data from all 12 leads significantly enhances diagnostic performance. Each lead provides unique electrical activity information from different areas and angles of the heart, leading to a more comprehensive understanding of cardiac function. Integrating data from all leads improves the richness and diversity of features captured by the model, enhancing its ability to identify complex patterns and diagnose various cardiac conditions more accurately.

Before building the model on the PTB-XL dataset, we preprocess the data to ensure quality and consistency. For ECG signal data, the steps include smoothing filtering, 50Hz notch filtering, R-wave segmentation extraction, and baseline drift removal. After preprocessing, the dataset is split into training and test sets. For the diagnosis text report data associated with each ECG signal, we use Google Translate to convert German and Swedish texts into English. The diagnostic text is then tokenized, removing stop words and special characters. Finally, we create vocabulary and vector representations of the diagnostic text.

3.2 Framework of Diagnosis Report Generations

This paper proposes a novel framework for the ECG diagnosis report generation as shown in Figure 1. It is divided into several key steps, each detailing how to extract and interpret data from ECG images to provide meaningful medical insights.

3.2.1 CNN Classifier. : Popular 2D CNN models include ResNet-50 [3], VGG-16[15], DenseNet[5], and InceptionNet[16]. Each model offers unique advantages in the classification of electrocardiogram diagnoses: ResNet-50 with its deeper structure and residual connections, VGG-16 for its simplicity, DenseNet for efficient feature

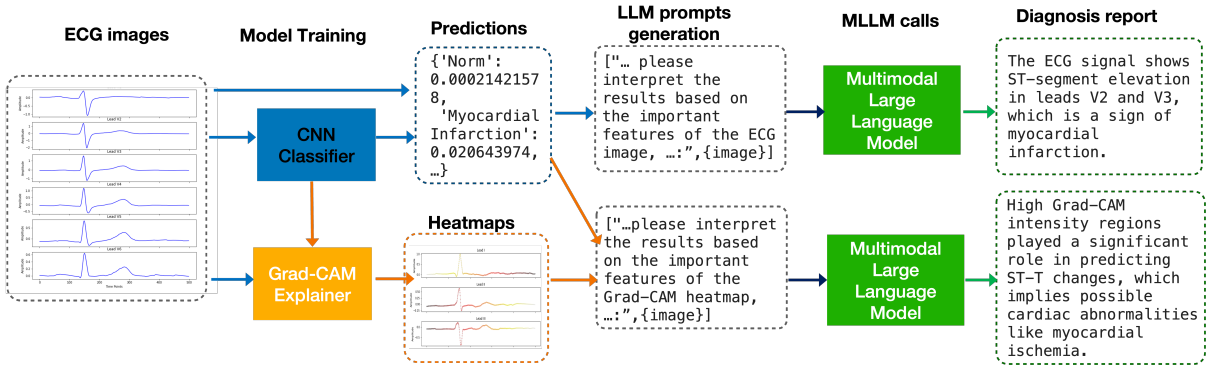


Figure 1: Overview of the proposed framework

reuse, and InceptionNet for capturing multi-scale features [18]. We deployed these models on our data to determine the most suitable model for ECG classification, considering both performance and interpretability.

3.2.2 Grad-CAM Explianer. : Grad-CAM stands out among various XAI methods due to its significant interpretability benefits, particularly within the context of CNN [1]. It excels in providing class-specific explanations, showcasing how the model's focus varies across different predicted classes. This is especially useful in multi-label classification tasks, offering a distinct advantage over methods like SHAP and LIME, which typically focus on the average contribution of features without providing class-specific insights [2]. Additionally, Grad-CAM's utility in image-related deep learning applications, such as ECG diagnosis, is particularly valuable; the heatmaps it generates can be directly superimposed on medical images, offering intuitive and actionable insights for medical professionals [4]. This capability for clear visual explanations, along with its high compatibility with pre-trained CNNs without requiring architectural modifications, and its computational efficiency makes Grad-CAM an invaluable tool in fields requiring transparent decision-making processes [6].

3.2.3 MLLM Prompts Generation. : Once we obtain the disease classification labels, they can then be used as inputs into a LLM. The simplest form of this would be using the classification labels as prompts or starting points for generating text. Additionally, the original ECG images or heatmaps marked with important features using Grad-CAM can be incorporated into a multimodal large language model. For example,

Prompt: ["the prediction of this ECG signal with 12 lead is following: {label}, please interpret the results based on the important features of the ECG image, answer it in two sentences or less in English:",{image}]

3.2.4 Diagnosis report. : This step involves leveraging multimodal large languages models to interpret ECG classification results by integrating their image-text understanding abilities with existing medical knowledge bases. This enables the system to generate detailed explanations, drawing on recognized patterns in ECG images and cross-referencing them with authoritative medical literature, enhancing diagnostic insights and user trust. The current two most powerful multimodal large language models are GPT-4V [9] and Gemini [17]. GPT-4V is distinguished by its ability to deliver precise and concise responses, effectively managing complex scenarios and excelling in human interactions and emotional intelligence aspects. While Gemini stands out for its detailed and expansive responses, which include relevant images and links, excelling in providing a rich narrative and visual detail, which is crucial for deep comprehension and thorough analysis [10].

Finally, an interactive interface displays the classification results, explanatory images, and explanatory texts to users, allowing medical professionals to understand the diagnostic information of each ECG case.

3.3 Performance Evaluation Metrics

In our framework, we employ two distinct models tailored to different aspects of the task, which makes the evaluation method more comprehensive and specialized.

For the visual model that classifies images, we are using standard classification metrics such as precision, recall, and F1 score. These metrics are ideal for this purpose as they quantify how well the model identifies and classifies the correct categories within the images. Precision measures the accuracy of the positive predictions, recall assesses how well the model captures all relevant instances, and the F1 score provides a balance between precision and recall, important for evaluating the overall performance of the model especially when the class distribution is uneven.

To assess the quality of text report generated by the multimodal models, we have chosen to use BERT-Score ([24]) as the evaluation metric. BERT-Score leverages a pre-trained BERT model to measure the semantic similarity between candidate text and reference text. This metric was chosen because, in medical image description, ground truth is valuable but often limited to specific phrases. Multimodal large models generate text with greater novelty and creativity, offering richer perspectives than standard benchmarks. While expert's manual evaluation is ideal for accuracy and relevance, it is expensive, time-consuming, and hard to scale. BERT-Score serves as a consistent, scalable, and cost-effective alternative to assess semantic accuracy and depth, which can effectively evaluate the quality of the model's textual output and serve as a useful preliminary filter to identify high-quality descriptions.

4 Experiments and Results

4.1 Experiment Settings

The PTB-XL dataset contains 21,837 clinical 12-lead ECG records from 18,885 patients, annotated with 71 SCP-ECG statements. These statements include 44 diagnostic, 19 form (4 overlapping with diagnostic), and 12 rhythm statements. The dataset's diagnostic labels are organized into 5 superclasses and 24 subclasses, offering granularity in analysis. It includes 12-lead ECG signals sampled at 500 Hz, along with demographic information, and diagnostic annotations.

The preprocessed data was saved in two formats: images and arrays. The image format is used for prompts in MLLMs, while the array format is used as input for image classification algorithms.We applied several 2D CNN algorithms, including ResNet-50, VGG-16, DenseNet, and InceptionNet, on the preprocessed PTB-XL dataset to determine the most suitable model for ECG classification based on performance metrics like precision, recall, F1-score, and accuracy.

In the implementation of Grad-CAM, we targeted the last convolutional, as it contains rich high-level features that reflect specific spatial information. The gradients of this layer with respect to the class of interest were computed, and global average pooling was applied to determine the importance of each feature map. These weighted feature maps were then combined and passed through a ReLU function to emphasize areas contributing most to the class prediction.

For the multimodal large language models, we utilized the APIs of GPT-4 Turbo and Gemini Pro Vision 1.0. These models were used to generate detailed diagnostic reports based on the enhanced ECG images and their classification labels. The generated reports were then compared to the dataset's diagnostic reports, used as ground truth, and evaluated using BERT-Score to measure the semantic similarity and quality of the text descriptions.

When using large language models for inference, it's crucial to configure certain parameters to ensure reproducibility and minimize randomness. The main configuration parameters include several key settings: a low temperature (0.1), a top k of 1, a top p of 1, and a max outputs tokens of 2048 words.

4.2 Results

4.2.1 ECG Classification. Table 1 shows the classification performance of all tested models. According to Table 1, Inception performs the best overall with precision, recall, and accuracy all at 0.70. DenseNet has slightly better F1-score (0.67) than ResNet-50 (0.66). VGG-16 has the lowest performance in all metrics. As a result, we chose Inception as the backbone of our framework.

Table 1: Classification performance of all tested models

Model	Precision	Recall	F1-score	Accuracy
VGG-16[15]	0.62	0.65	0.62	0.65
DenseNet[5]	0.69	0.69	0.67	0.69
Inception[16]	**0.70**	**0.70**	**0.67**	**0.70**
ResNet-50[3]	0.68	0.69	0.66	0.69

Table 2 illustrates the classification performance for ECG data with the use of Inception model. It shows that Inception performs well for the most common class (NORM) but struggles significantly with the less frequent classes (especially CD and HYP). This disparity suggests that the model might benefit from techniques such as class balancing, data augmentation for underrepresented classes, or using a more sophisticated model that can better capture the nuances of the less frequent classes.

Table 2: ECG class classification performance with Inception

Class	Precision	Recall	F1-score	Instances
NORM	0.70	0.94	0.80	844
MI	0.73	0.57	0.64	490
STTC	0.67	0.59	0.63	326
CD	0.61	0.31	0.41	167
HYP	0.80	0.11	0.19	74

4.2.2 ECG Diagnosis Report Generation. In this study, two types of inputs of ECG image and visual interpretation Grad-CAM are input to MLLMs of Gemini and GPT-4V respectively for ECG diagnosis report generations, achieving four groups of testing results as shown in Figure 2. Figure 2compares the four methods: Gemini_Image, Gemini_GradCAM, GPT-4V_Image, and GPT-4V_GradCAM across Precision, Recall, and F1-score. Gemini_GradCAM showed the best overall performance with the highest Precision (0.816), Recall (0.821), and F1-score (0.818). GPT-4V_GradCAM also performed well, closely following Gemini_GradCAM. Both Grad-CAM methods outperformed their respective image-only methods, indicating that Grad-CAM significantly enhances ECG diagnosis report generation performance in terms of precision, recall, and F1-score.

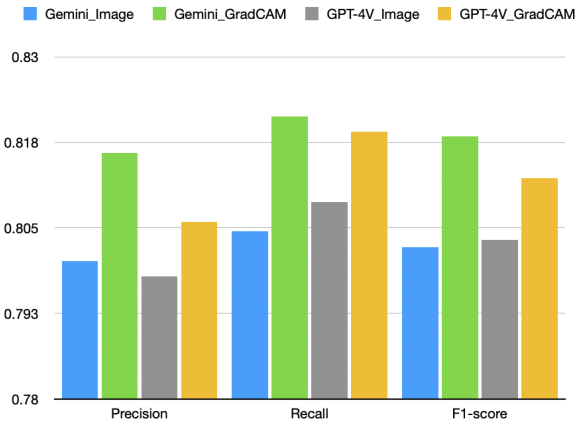


Figure 2: Performance of diagnosis report generations

5 Discussion and Limitations

The framework proposed in this paper is composed two key models: one focusing on classifying images into different ECG classes and the other on generating text descriptions (i.e. ECG report) for those images. This architecture allows us to explore how different types of input, raw images with classification labels and images enhanced by Grad-CAM with labels, affect the quality of the generated diagnosis report. Interestingly, the results indicate that current pre-trained

multimodal large models, even under zero-shot learning conditions, are capable of generating high-quality text in specialized fields like medical field. This underscores the models' potential to adapt and produce relevant output without prior specific training on similar tasks. The use of Grad-CAM images further result in more precise and insightful text descriptions, confirming that this method enhances the model's ability to focus on and interpret the most relevant features of an image.

However, there are three main drawbacks to this method: the PTB-XL dataset used for classifications has a limited number of samples, particularly in CD and HYP classes, which can impact classification performance. Furthermore, the diagnosis reports used as ground truth are not always complete or fluent, affecting the evaluation metrics for text generations. Although BERT-Score can evaluate the semantic similarity between the generated text and the reference text, it cannot fully replace expert evaluation, especially in cases requiring precise medical or technical descriptions. Expert assessments provide a deeper check on accuracy and relevance, which BERT-Score lacks. Third, the quality of the text generations heavily depends on the performance of the initial image classification model. If the accuracy of the classification model is not high, then the descriptions generated, whether from raw images or Grad-CAM processed images, may significantly deviate from actual conditions, thereby affecting the overall effectiveness and application value of the system.

6 Conclusion and Future Work

This study demonstrates the potential of integrating Explainable AI (XAI) techniques with multimodal large language models to enhance the interpretability and accuracy of ECG diagnosis. By employing a 2D CNN for ECG signal classification and utilizing Grad-CAM to highlight critical areas influencing AI decisions, we were able to generate comprehensive explanatory outputs through multimodal models. This approach combines the strengths of visual and textual data processing, offering clearer insights into AI reasoning and bridging the gap between AI capabilities and human interpretation. Despite some limitations, such as dependency on classification accuracy and the need for expert evaluation, our framework sets a new standard for AI integration in clinical practices. Future efforts will focus on improving the alignment of image classification and text generation to further streamline the diagnostic process and ensure robust, reliable, and interpretable AI-driven medical diagnostics.

References

[1] Theresa Bender and et al. 2023. Analysis of a Deep Learning Model for 12-Lead ECG Classification Reveals Learned Features Similar to Diagnostic Criteria. *IEEE Journal of Biomedical and Health Informatics* (2023).

[2] David Cian, Jan van Gemert, and Attila Lengyel. 2020. Evaluating the performance of the LIME and Grad-CAM explanation methods on a LEGO multi-label image classification task. *arXiv preprint arXiv:2008.01584* (2020).

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[4] Steven A Hicks and et al. 2021. Explaining deep neural networks for knowledge discovery in electrocardiogram analysis. *Scientific reports* 11, 1 (2021), 10949.

[5] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. Densely Connected Convolutional Networks. arXiv:1608.06993 [cs.CV]

[6] Vicneswary Jahmunah, Eddie YK Ng, Ru-San Tan, Shu Lih Oh, and U Rajendra Acharya. 2022. Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals. *Computers in Biology and Medicine* 146 (2022), 105550.

[7] Chunyu Liu, Yongpei Ma, Kavitha Kothur, Armin Nikpour, and Omid Kavehei. 2023. BioSignal Copilot: Leveraging the power of LLMs in drafting reports for biomedical signals. *medRxiv* (2023), 2023–06.

[8] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[9] OpenAI, Josh Achiam, and et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[10] Zhangyang Qi, Ye Fang, Mengchen Zhang, Zeyi Sun, Tong Wu, Ziwei Liu, Dahua Lin, Jiaqi Wang, and Hengshuang Zhao. 2023. Gemini vs GPT-4V: A Preliminary Comparison and Combination of Vision-Language Models Through Qualitative Cases. arXiv:2312.15011 [cs.CV]

[11] Nikita Rafie, Anthony H Kashou, and Peter A Noseworthy. 2021. ECG interpretation: clinical relevance, challenges, and advances. *Hearts* 2, 4 (2021), 505–513.

[12] Pranav Rajpurkar, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng. 2017. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836* (2017).

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of KDD 2016*. 1135–1144.

[14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[15] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]

[16] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going Deeper with Convolutions. arXiv:1409.4842 [cs.CV]

[17] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[18] Tabassum Islam Toma and Sunwoong Choi. 2022. A Comparative Analysis of 2D Deep CNN Models for Arrhythmia Detection Using STFT-Based Long Duration ECG Spectrogram. In *2022 Thirteenth International Conference on Ubiquitous and Future Networks (ICUFN)*. 483–488.

[19] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209* (2024).

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[21] P. Wagner, N. Strodthoff, R. D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter. 2020. PTB-XL, a large publicly available electrocardiography dataset. *Scientific data* 7, 1 (2020), 154.

[22] Joel Xue and Long Yu. 2021. Applications of machine learning in ambulatory ecg. *Hearts* 2, 4 (2021), 472–494.

[23] Han Yu, Peikun Guo, and Akane Sano. 2023. Zero-Shot ECG Diagnosis with Large Language Models and Retrieval-Augmented Generation. In *Machine Learning for Health (ML4H)*. PMLR, 650–663.

[24] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).