Clinically Interpretable Zero-Shot ECG Classification via Multimodal Learning and Expert-Aligned Descriptors

Luiz Facury de Souza, José Geraldo Fernandes, Pedro Robles Dutenhefner, Turi Andrade Rezende, Gisele L. Pappa, Gabriela Miana Paixão, Antonio Luiz Pinho Ribeiro, Wagner Meira Jr.

Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Abstract

The widespread adoption of artificial intelligence in clinical cardiology is hampered by a critical factor: the lack of transparency in automated electrocardiogram (ECG) interpretation systems. While deep learning models can accurately classify cardiac abnormalities, their "black-box" nature prevents clinicians from verifying the diagnostic reasoning, undermining clinical trust. To address this, we developed a multimodal diagnostic framework that emulates clinical reasoning by directly linking ECG signal features to their corresponding textual descriptions from clinical reports. Our system learns to recognize abnormalities like Left Bundle Branch Block (LBBB) not merely as a classification label, but by identifying and associating it with established diagnostic criteria, such as a 'widened QRS complex ($>$ 120 ms)'. Trained on a large dataset of paired ECG signals and narrative reports, our model achieves diagnostic accuracies comparable to conventional supervised models without requiring explicit training on classification labels. By grounding its predictions in clinically relevant subfeatures, the system provides transparent, verifiable evidence for its conclusions. This approach represents a paradigm shift toward AI systems that augment clinical decision-making with intelligible, evidence-based insights, fostering greater trust and facilitating integration into the diagnostic workflow.

1. Introduction

The 12-lead electrocardiogram (ECG) is a cornerstone of cardiovascular medicine, offering a non-invasive, cost-effective window into the heart's electrical function. With cardiovascular diseases remaining the leading cause of global mortality [1], the immense volume of ECGs performed annually strains healthcare systems and expert interpreters, leading to potential delays in diagnosis and adverse patient outcomes.

To meet this challenge, automated ECG analysis has evolved significantly. Supervised learning approaches have achieved cardiologist-level accuracy for a range of conditions [2–6]. More recently, self-supervised learning methods have demonstrated the ability to learn robust representations from vast quantities of unlabeled ECG data [7, 8].

Despite these triumphs, a significant barrier to clinical adoption persists: the "black-box" problem. The opaque nature of deep learning models poses tangible risks: without a clear rationale, clinicians cannot easily distinguish a correct, nuanced diagnosis from a coincidental correlation or an error caused by an out-of-distribution artifact. This fundamentally limits the role of AI to that of a preliminary screening tool rather than a trusted diagnostic partner. Standard explainability techniques, such as attention maps or Grad-CAM [9, 10], often highlight regions of the ECG signal but fail to provide explanations in a clinically meaningful lexicon.

A promising direction to bridge this trust gap is multimodal learning, integrating signals with clinical text [11–13]. This work builds on that foundation by proposing a framework that classifies ECGs in a zero-shot manner by grounding its predictions in verifiable, text-based criteria, which we term "subfeatures." Our central hypothesis is that by forcing a model to learn these direct signal-to-text correlations, it will develop a more robust and generalizable understanding of cardiac electrophysiology than a model trained on abstract labels alone.

The standard ECG signal is composed of three primary waveforms: the P-wave (atrial depolarization), the QRS complex (ventricular depolarization), and the T-wave (ventricular repolarization). These waveforms encode critical diagnostic features for the abnormalities studied here. For example, 1st-degree AV block manifests as a prolonged PR interval between the P-wave and QRS complex, while RBBB and LBBB critically alter QRS morphology and duration. Atrial fibrillation disrupts the regularity of P-waves and the rhythm of QRS complexes, while sinus tachycardia and bradycardia are defined by the rate of successive QRS complexes (the RR interval). Our framework is designed to explicitly learn these fundamental associations.

2. Methods

Our methodology correlates signal morphology with clinical terminology by leveraging a large, multimodally annotated dataset and a vision-language training paradigm.

2.1. Dataset

We used the CODE-15% dataset, a 15% subsample of the CODE study [2], containing 345,779 standard 10-second, 12-lead ECGs sampled at 400 Hz from 233,770 unique patients. We focused on six prevalent diagnoses: 1st-degree atrioventricular block (1dAVb), right bundle branch block (RBBB), left bundle branch block (LBBB), sinus bradycardia (SB), atrial fibrillation (AFIB), and sinus tachycardia (ST). All data were de-identified to protect patient privacy in accordance with ethical guidelines.

The dataset's key feature is the inclusion of detailed medical reports generated by the validated University of Glasgow ECG analysis program [14, 15]. An example is shown in Table 1. We trained our primary model using only the narrative signal descriptions, while a supervised baseline was trained on the diagnostic labels for comparison.

Table 1. Example of a clinical narrative report from the CODE-15 dataset, translated from Portuguese.

> **Signal Description:** Rightward and superior QRS axis deviation. P wave: normal amplitude and duration. PR interval: normal duration. QRS: normal axis and amplitudes. Prolonged duration with RSR pattern in V1 and a wide S wave in lateral leads. ST segment and T wave: secondary changes due to RBBB. QTc: abnormal.
> **Conclusion:** Sinus tachycardia; right bundle branch block.
> **Labels:** 1dAVb=0, RBBB=1, LBBB=0, SB=0, AFIB=0, ST=1.

2.2. Model and Training

The architecture comprises a 1D Vision Transformer (ViT) [16] for the ECG signal and a BioBERTpt model [17, 18] for the clinical text. The ViT architecture was selected for its proven ability to model long-range dependencies, crucial for analyzing ECGs where diagnostic clues can span significant time intervals. The raw signal is segmented into non-overlapping patches, which are then linearly projected to form the input sequence for the Transformer encoder. Our training strategy, inspired by CLIP [19], aligns the representations from these two modalities.

This training paradigm is conceptually twofold. The first objective, a contrastive signal loss, builds a robust ECG encoder by training the model to recognize two augmented versions of the same ECG as being similar, while pushing them apart from other ECGs in a batch. This ensures the learned features are invariant to minor noise. The second, and more critical, objective is the signal-text alignment loss. Here, the model learns to minimize the distance in the shared embedding space between an ECG signal's representation and that of its paired narrative report. This process, performed end-to-end using the Adam optimizer, explicitly forces the model to learn the visual manifestation of clinical descriptions, effectively creating a cross-modal dictionary between ECG patterns and medical language.

2.3. Evaluation Protocol

Evaluation was performed on the CODE Test subset of 827 high-quality ECGs. The zero-shot model was evaluated via a clinically intuitive process: for a new ECG, its signal embedding is compared to a library of pre-computed text embeddings of "subfeatures." These subfeatures were crafted by consulting clinical guidelines to ensure they represent canonical diagnostic criteria. A diagnosis is made if the cosine similarity score for a subfeature surpasses a threshold optimized on a validation set to balance sensitivity and specificity, providing a direct, textual justification for each finding. For the supervised baseline, the same ViT architecture was used, but its output was fed into a classification head trained with a standard cross-entropy loss against the diagnostic labels.

3. Results

Our experiments show that the zero-shot multimodal approach achieves diagnostic performance on par with a fully supervised model, while offering full interpretability.

3.1. Diagnostic Accuracy

As detailed in Table 2 and Table 3, our interpretable model achieved a mean accuracy of 0.968 and a mean AUC of 0.861. This performance is highly competitive with the supervised model's mean accuracy of 0.973 and mean AUC of 0.871. One possible interpretation for the interpretable model's stronger performance on LBBB and AFIB is that the narrative descriptions for these conditions contain rich morphological descriptors (e.g., 'notched QRS', 'absent P-waves') that provide a more powerful training signal than a simple binary label.

Table 2. Accuracy comparison of the interpretable (zero-shot) and supervised models across all six conditions.

Abnormality	**Interpretable**	**Supervised**
1dAVb	0.961	0.970
RBBB	0.966	0.959
LBBB	0.959	0.981
SB	0.984	0.986
AFIB	0.978	0.983
ST	0.960	0.959
Average	0.968	0.973

Table 3. AUC comparison of the interpretable (zero-shot) and supervised models, showing competitive performance.

Abnormality	**Interpretable**	**Supervised**
1dAVb	0.670	0.696
RBBB	0.843	0.982
LBBB	0.987	0.992
SB	0.861	0.896
AFIB	0.862	0.708
ST	0.944	0.954
Average	0.861	0.871

3.2. Clinical Interpretability

The framework's primary contribution is its delivery of transparent, evidence-based diagnoses. For a diagnosis of RBBB, the model must find high similarity with the sub-feature: *"QRS complex duration >120 ms with characteristic RBBB morphology."* As shown in Figure 1, the system correctly identified an RBBB case exhibiting these pathognomonic features. This granular, criteria-based explanation moves beyond abstract visualizations, offering a verifiable audit trail for each diagnosis. This is critical not only for clinical trust but also for educational purposes, allowing trainees to see a clear link between textbook criteria and their real-world presentation on an ECG.

4. Conclusion

We have presented a multimodal framework for ECG analysis that achieves high diagnostic accuracy without sacrificing interpretability. Our primary contribution is demonstrating that high performance can be achieved without sacrificing transparency, a critical step toward clinically integrated AI.

The clinical relevance of this approach is significant. The workflow implication is a potential shift from simple AI-driven alerts to interactive diagnostic sessions where a physician can review a finding and see the specific supporting evidence. In resource-limited settings, it could function as a dependable screening and triaging tool. The core innovation is the shift from opaque predictions to a collaborative, evidence-based dialogue between the clinician and the AI.

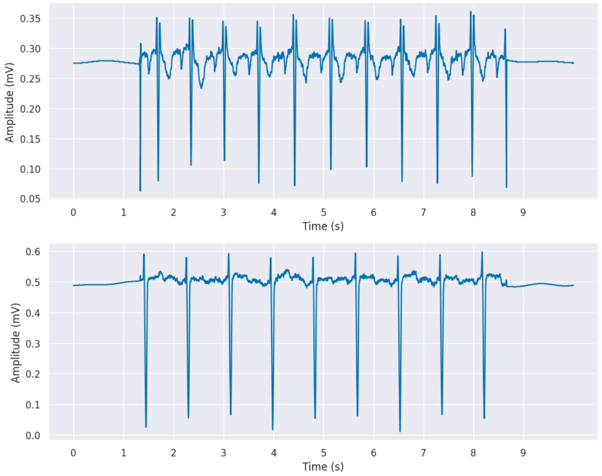

Figure 1. Comparison of lead V1 tracings. Top: An ECG with Right Bundle Branch Block (RBBB), showing the characteristic widened QRS complex (> 120 ms) and morphological pattern. Bottom: An ECG showing normal sinus rhythm for comparison.

This study has several limitations. First, its foundation on a single dataset means that generalizability must be confirmed on external datasets from diverse patient populations. Second, the reliance on machine-generated reports may not capture the variability of human-authored notes. Finally, our diagnostic scope was limited to six common conditions.

Future directions will focus on clinical translation. Key efforts will include expanding the diagnostic lexicon to include rarer arrhythmias; developing an interactive clinical interface where clinicians can query the model to highlight corresponding waveform evidence; and ultimately, conducting prospective clinical trials to rigorously assess the tool's real-world impact on diagnostic accuracy and clinician confidence. This work represents a step towards a new generation of medical AI designed to be collaborative partners in clinical reasoning.

Acknowledgments

This work is partially supported by CNPq, CAPES, Fapemig, Amazon AWS, as well as projects CIIA-Saúde and IAIA - INCT on AI.

References

[1] Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G, Ahmed M, Aksut B, Alam T, Alam K, et al.

Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. Journal of the American college of cardiology 2017;70(1):1–25.

[2] Ribeiro AH, Ribeiro MH, Paixão GM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MP, Andersson CR, Macfarlane PW, Meira Jr W, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. Nature communications 2020;11(1):1760.

[3] Pyakillya B, Kazachenko N, Mikhailovsky N. Deep learning for ecg classification. In Journal of physics: conference series, volume 913. IOP Publishing, 2017; 012004.

[4] Liu X, Wang H, Li Z, Qin L. Deep learning in ecg diagnosis: A review. Knowledge Based Systems 2021; 227:107187.

[5] Chen CY, Lin YT, Lee SJ, Tsai WC, Huang TC, Liu YH, Cheng MC, Dai CY. Automated ecg classification based on 1d deep learning network. Methods 2022;202:127–135.

[6] Singh PN, Mahapatra RP. A novel deep learning approach for arrhythmia prediction on ecg classification using recurrent cnn with gwo. International Journal of Information Technology 2024;16(1):577–585.

[7] Mehari T, Strodthoff N. Self-supervised representation learning from 12-lead ecg data. Computers in biology and medicine 2022;141:105114.

[8] Sarkar P, Etemad A. Self-supervised learning for ecg-based emotion recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020; 3217–3221.

[9] Anand A, Kadian T, Shetty MK, Gupta A. Explainable ai decision model for ecg data of cardiac disorders. Biomedical Signal Processing and Control 2022;75:103584.

[10] Ganeshkumar M, Ravi V, Sowmya V, Gopalakrishnan E, Soman K. Explainable deep learning-based approach for multilabel classification of electrocardiogram. IEEE Transactions on Engineering Management 2021;70(8):2787–2799.

[11] Liu C, Wan Z, Ouyang C, Shah A, Bai W, Arcucci R. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. arXiv preprint arXiv240306659 2024;.

[12] Phan T, Le D, Brijesh P, Adjeroh D, Wu J, Jensen MO, Le N. Multimodality multi-lead ecg arrhythmia classification using self-supervised learning. In 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, 2022; 01–04.

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

[14] Macfarlane P, Devine B, Latif S, McLaughlin S, Shoat D, Watts M. Methodology of ecg interpretation in the glasgow program. Methods of information in medicine 1990; 29(04):354–361.

[15] Macfarlane P, Devine B, Clark E. The university of glasgow (uni-g) ecg analysis program. In Computers in Cardiology, 2005. IEEE, 2005; 451–454.

[16] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv201011929 2020;.

[17] Schneider ETR, de Souza JVA, Knafou J, Oliveira LESe, Copara J, Gumiel YB, Oliveira LFAd, Paraiso EC, Teodoro D, Barra CMCM. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online: Association for Computational Linguistics, November 2020; 65–72. URL `https://www.aclweb.org/anthology/2020.clinicalnlp-1`

[18] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pretraining of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019; 4171–4186.

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

Address for correspondence:

Luiz Facury de Souza
Universidade Federal de Minas Gerais
Av. Pres. Antônio Carlos, 6627 - Pampulha
Belo Horizonte, MG, 31270-901, Brazil
luizfysouza@ufmg.br