CardioRAG: A Retrieval-Augmented Generation Framework for Multimodal Chagas Disease Detection

Zhengyang Shen[1], Xuehao Zhai[2], Hua Tu[1], Mayue Shi[1,3]

[1] Department of Electrical and Electronic Engineering, Imperial College London, London, UK
[2] Department of Civil and Environmental Engineering, Imperial College London, London, UK
[3] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK

Abstract

Chagas disease affects nearly 6 million people worldwide, with Chagas cardiomyopathy representing its most severe complication. AI-based ECG screening is promising in settings where serological testing capacity is limited, but current machine learning approaches face challenges including limited accuracy, poor interpretability, and dependence on large labeled datasets. Moreover, they remain difficult to integrate with evidence-based clinical diagnostic indicators. We propose CardioRAG, a novel retrieval-augmented generation (RAG) framework that integrates large language model (LLM) with interpretable ECG clinical features. Variational autoencoder-learned representations were used for semantic case retrieval, providing similar ECG cases to guide LLM-based clinical reasoning. On an independent validation set of 100 cases, CardioRAG achieved 58.59% accuracy with 87.76% recall and F1 score of 0.68, effectively identified positive cases for prioritized serological testing. The framework provides a clinical evidence-based approach to Chagas disease screening that combines clinical knowledge with AI reasoning, also demonstrates a promising pathway for embedding clinical indicators into broader AI systems for diagnosis. This study was conducted as part of the PhysioNet Challenge 2025 (Team: ECGenius). The proposed model was unable to be scored on the hidden test set.

1. Introduction

Chagas disease is a neglected tropical disease affecting approximately 6 million people worldwide, with fewer than 10% aware of their infection status [1]. The disease can progress to Chagas cardiomyopathy (ChCM), where electrocardiographic abnormalities often precede overt structural heart disease [2]. ECG provides a pragmatic, low-cost tool for early risk stratification in resource-limited settings, enabling prioritized serological testing and more efficient resource allocation [3, 4].

Recently, modern data-driven approaches have enabled new paradigms for disease detection from physiological signals. Advanced machine-learning methods can model non-linear relationships between disease status and multivariate time-series signals, such as ECG [5] and movement [6]. However, current methods exhibit persistent limitations: (i) performance instability across domains due to population shift and limited calibration [7], (ii) limited clinical interpretability hindering trust and adoption, and (iii) dependence on large, well-curated labeled datasets that are scarce for neglected diseases.

To address these challenges, we introduce CardioRAG, a novel multimodal retrieval-augmented generation (RAG) framework integrating interpretable ECG clinical features with LLM-based diagnostic reasoning. It targets the critical screening scenario where high recall is essential for identifying potential Chagas cases for prioritized serological testing. This work makes three key contributions: (1) a clinically-grounded RAG pipeline combining established ECG biomarkers (RBBB, LAFB) with heart rate variability metrics, achieving consistent high recall performance ($>$85%) across different model configurations; (2) a VAE-based representation learning system coupled with demographic screening, enabling effective similar case retrieval; (3) empirical demonstration that prompt simplification and balanced case retrieval optimize RAG performance.

2. Methodology

We propose a comprehensive framework for automated Chagas disease detection that integrates deep ECG representation learning with retrieval-augmented generation (RAG) [8] for enhanced diagnostic reasoning. The system (Figure 1) processes 12-lead ECG with patient demographic data (age, sex) through three main stages: extraction of clinical features from ECG signals; VAE-based representation learning for semantic similarity [9]; and RAG-

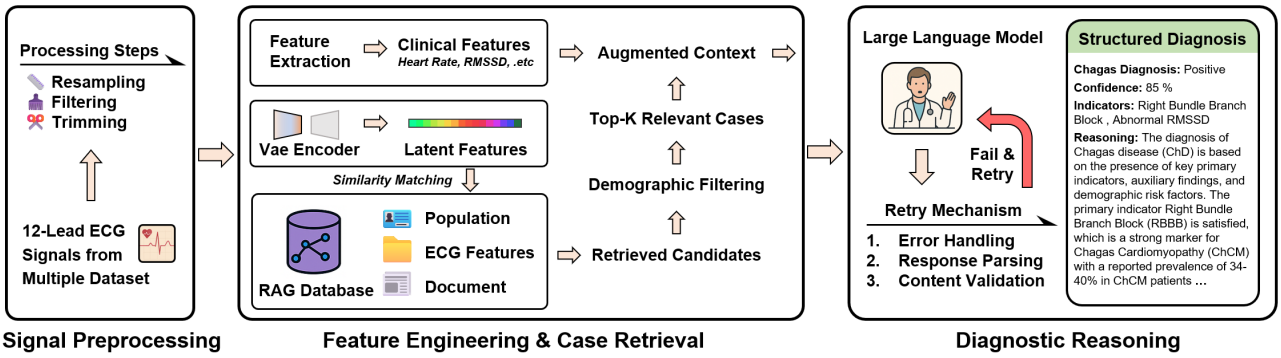

Figure 1: The CardioRAG Framework for Chagas disease diagnosis from 12-lead ECG signals. The system preprocesses raw ECG data, extracts clinical and latent features via VAE, retrieves relevant cases from a RAG database, and generates structured diagnoses with confidence scores using a large language model.

enhanced diagnostic decision making with large language models (LLMs).

2.1. Datasets and preprocessing

This study utilized three publicly available ECG datasets from the PhysioNet Challenge [3, 4, 10]: SaMi-Trop dataset, PTB-XL dataset, and CODE-15% dataset. All ECG signals underwent standardized pre-processing: (1) resampling recordings to 400 Hz using linear interpolation, (2) standardizing signal durations to 7 seconds through cropping or padding, and (3) filtering using the NeuroKit2 toolbox for noise removal and baseline correction.

2.2. Chagas-specific feature engineering

Chagas disease and ChCM manifest as specific ECG abnormalities, particularly conduction and rhythm disorders [2]. For conduction disorders, we implemented automated detection of right bundle branch block (RBBB) and left anterior fascicular block (LAFB) using Minnesota Code criteria [11]. RBBB and LAFB represent key ChCM manifestations, with prevalence rates of 34-40% and 23-39% respectively in ChCM patients [2]. Table 1 outlines the specific ECG parameters required for automated detection.

Table 1: ECG Parameters used for diagnosis

Feature	Target Leads	Required ECG Parameters
RBBB	I, II, III, aVL aVF, V1, V2	QRS duration, R wave duration, R peak duration, R wave amplitude, R' wave amplitude, S wave duration, S wave amplitude, net QRS deflection
LAFB	I, II, III, aVL aVF	QRS duration, Q wave duration, Q wave amplitude, QRS axis angle

For rhythm assessment, RR-derived metrics were extracted from lead V5, including ventricular rate and RMSSD (root mean square of successive differences). RMSSD serves as a short-term heart rate variability index, with reduced values significantly associated with Chagas disease [12]. These features, combined with demographic information (age and sex), form the comprehensive multimodal input to the RAG diagnostic system.

2.3. CardioRAG diagnostic architecture

The RAG framework addresses the fundamental challenge of labeled data scarcity in Chagas disease detection by enabling case-based reasoning via retrieval of similar historical cases. This diagnostic approach aligns with clinical practice, in which physicians rely on prior cases to guide complex diagnostic decisions. [8, 13].

Variational autoencoder for signal embedding. We employ a variational autoencoder (VAE) [9] to learn compact ECG representations that support effective similarity search. The encoder consists of four residual blocks with progressively increasing channels (32, 64, 128, 256). Each residual block contains two 1D convolutions with Batch Normalization, ReLU and a skip connection. The encoder outputs (μ) and log-variance ($\log \sigma^2$) parameters of a 256-dimensional latent distribution. Training employs the standard VAE objective:

$$L = L_{\text{recon}} + \beta \cdot L_{\text{KL}} \tag{1}$$

where $L_{\text{recon}} = E_{q_\phi(z|x)}[\log p_\theta(x|z)]$ is the reconstruction loss, $L_{\text{KL}} = D_{\text{KL}}(q_\phi(z|x)||p(z))$ is the KL divergence regularization term, and β is set to 0.1 based on validation performance.

Case retrieval mechanism. The retrieval process implements a two-stage search strategy combining VAE-based similarity with demographic filtering. Similarity search begins in the VAE latent space using cosine similarity to identify the k most similar cases (with k tuned on validation data). The secondary filtering computes a com-

posite similarity score:

$$S_{\text{composite}} = S_{\text{VAE}} + w_{\text{age}} \cdot S_{\text{age}} \quad (2)$$

where S_{VAE} is normalized VAE similarity, S_{age} reflects age similarity using a Gaussian kernel with $\sigma = 10$ years, and w_{age} is the weighting coefficient. Retrieved cases are formatted into structured context for the LLM, including patient demographics, detected clinical features, HRV metrics, and diagnostic labels, with length control to avoid prompt overflow.

LLM powered diagnostic reasoning. The LLM receives structured prompts containing patient features and retrieved similar cases, generating diagnostic predictions with confidence scores and clinical reasoning. The LLM output follows a structured JSON format containing: (1) binary diagnosis (POSITIVE/NEGATIVE), (2) confidence percentage, (3) detailed clinical reasoning, (4) identified diagnostic indicators, (5) relevant risk factors, and (6) other cardiac findings. Note while confidence scores are generated, we found them unreliable for smaller LLMs and thus focus evaluation on binary diagnostic performance.

Example (LLM-generated diagnostic rationale). *"The patient presents with RBBB_satisfaction indicating a right bundle branch block, which is consistent with Chagas. The low RMSSD in Lead V6 (7.8 ms) strongly suggests a heart rhythm abnormality indicative of Chagas. No other significant ECG findings are noted, and the data supports a clear positive diagnosis."* ***Diagnosis:*** *Chagas positive.*

3. Results and analysis

We evaluated the CardioRAG framework using the DeepSeek-R1:1.5b language model on a test set of 100 patients, consisting of 50 consecutive positive cases from the SaMi-Trop dataset and 50 consecutive negative controls from the PTB-XL dataset. Our experiments focused on two critical aspects: the impact of prompt engineering and RAG retrieval strategies on diagnostic performance.

Note on official Challenge submission. Our CardioRAG framework was unable to be scored on the official hidden test set due to technical constraints. Instead, we submitted a supervised deep neural network baseline, D-Net [6], achieved an official Challenge score of 0.190 (ranked 29) under our team name (ECGenius). Therefore, the official leaderboard score and ranking do not reflect the performance of the CardioRAG model in this work.

3.1. Prompt engineering

Figure 2 presents the performance comparison across four prompt configurations. Counterintuitively, the "P2 Simplified Clinical" configuration achieved the best performance with 58.59% accuracy, 87.76% recall, and 67.72% F1 score, representing significant improvements

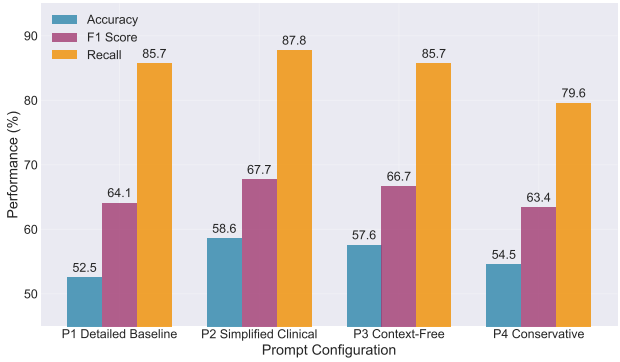

Figure 2: Impact of prompt engineering (top-k retrieved case, k=8). Configurations: P1 Detailed prompt (baseline, full ECG criteria and clinical instructions), P2 Simplified Clinical (without detailed ECG criteria for RBBB/LAFB), P3 Context-Free (without diagnostic background), P4 Conservative (includes cautionary guidance for positive diagnoses).

over the "P1 Detailed Baseline" (52.53% accuracy, 85.71% recall, 64.12% F1). This 6.06 percentage point accuracy improvement suggests that for smaller language models, concise prompts focusing on key decision factors outperform exhaustive clinical descriptions with detailed RBBB/LAFB detection criteria.

Notably, adding cautionary instructions ("P4 Conservative") decreased performance to 54.55% accuracy, indicating that overly conservative prompting may bias the model toward indecision. The optimal configuration maintained essential clinical context while avoiding information overload. In the annalysis, one case could not produce a valid structured output from the language model and was therefore excluded from the subsequent evaluation.

3.2. Retrieval strategies

Table 2 demonstrates the impact of retrieval augmentation on diagnostic performance. The relationship between the number of retrieved cases (k) and accuracy follows an inverted U-shape, with optimal performance at k=8 (58.59% accuracy). The baseline prompt (P1) without RAG achieved a low recall of 48.98%, which is significantly lower than all configurations with RAG. This demonstrated that RAG effectively enhanced the LLM's diagnostic performance.

Table 2: Comparison of retrieval configurations

Configuration	Accuracy%	Recall%	F1 Score
P1 No RAG	54.55	48.98	0.52
P1 RAG k=8	52.53	85.71	0.64
P2 RAG k=8	58.59	87.76	0.68
P2 RAG k=8 (bal)	58.59	89.80	0.68
P2 RAG k=16	52.53	77.55	0.62

The performance degradation observed at k=16 (52.53% accuracy) may be attributed to the introduction of excessive retrieved cases, which likely added noise rather than providing useful diagnostic context and potentially overwhelmed the LLM's reasoning capacity. In contrast, the balanced retrieval strategy at k=8 achieved the highest recall and F1 score, suggesting the influence of maintaining an appropriate proportion of representative positive and negative examples within the retrieval set.

These findings suggest alignment with our prompt engineering results, indicating that both prompt quality and RAG quantity may significantly influence LLM diagnostic performance. Neither maximal information provision nor extreme simplification yields optimal performance. Instead, balanced, focused contextual guidance appears to achieve the best diagnostic reasoning outcomes.

4. Discussion

Our CardioRAG framework represents a zero-shot learning paradigm based on LLMs, achieved substantially higher recall. It enables effective learning with limited data while explicitly incorporating established clinical indicators (RBBB, LAFB, HRV), providing interpretability advantages over black-box networks [14]. In contrast, recent ECG-based Chagas detection work has focused on supervised deep learning [5], achieving moderate performance but requiring extensive labeled data.

The observed 58–59% accuracy ceiling may reflect limitations associated with the relatively small model size (1.5B parameters) in our implementation. Larger models and further multi-modal information, such as temporal disease patterns, should be evaluated. The computational requirements (25-40 seconds per case) and dependence on proprietary LLMs require investigation of model compression and open-source alternatives for practical deployment in resource-constrained settings.

5. Conclusion

CardioRAG integrates LLMs with interpretable ECG clinical features for Chagas disease screening, achieving 87.76% recall. Our analysis revealed that simplified prompts outperformed detailed descriptions; moderate case retrieval (k=8) with balanced cases achieved optimal performance. The main limitation is limited accuracy, suggesting a need to explore larger language models.

The framework's high recall makes it valuable for initial screening and patient triaging especially in low-resource regions. Moreover, by integrating RAG with LLM-based diagnostic reasoning, the framework could inherently co-evolve with advances in large language models. Future work should explore the integration of larger LLMs, enhanced multi-modal retrieval, and clinical validation across diverse populations.

References

[1] World Health Organization. Chagas Disease, 2025. URL https://www.who.int/health-topics/chagas-disease.

[2] Acquatella H. Echocardiography in Chagas Heart Disease. Circulation 2007;115(9):1124–1131.

[3] Reyna MA, et al. Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025. In Computing in Cardiology 2025, volume 52. 2025; 1–4.

[4] Reyna MA, et al. Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025, 2025. URL https://arxiv.org/abs/2510.02202.

[5] Jidling C, et al. Screening for Chagas Disease from the Electrocardiogram Using a Deep Neural Network. PLoS Neglected Tropical Diseases 2023;17(7):e0011118.

[6] Shen Z, Gao B, Shi M. COBRA: Multimodal Sensing Deep Learning Framework for Remote Chronic Obesity Management via Wrist-Worn Activity Monitoring. In IUPESM World Congress on Medical Physics and Biomedical Engineering 2025. Adelaide, Australia, 2025; .

[7] Patrini G, et al. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017; 1944–1952.

[8] Lewis P, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Advances in Neural Information Processing Systems, volume 33. 2020; 9459–9474.

[9] Kingma DP, Welling M. Auto-Encoding Variational Bayes, 2022. URL https://arxiv.org/abs/1312.6114.

[10] Goldberger AL, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 2000;101(23):e215–e220.

[11] Prineas RJ, Crow RS, Zhang ZM. The Minnesota Code Manual of Electrocardiographic Findings. Springer, 2010.

[12] Ribeiro ALP, et al. Power-Law Behavior of Heart Rate Variability in Chagas' Disease. The American Journal of Cardiology 2002;89(4):414–418.

[13] Ng KKY, et al. RAG in Health Care: A Novel Framework for Improving Communication and Decision-Making by Addressing LLM Limitations. Nejm Ai 2025; 2(1):AIra2400380.

[14] Abbasian Ardakani A, et al. Interpretation of Artificial Intelligence Models in Healthcare: A Pictorial Guide for Clinicians. Journal of Ultrasound in Medicine 2024; 43(10):1789–1818.

Address for correspondence:

Mayue Shi

Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford OX3 7DQ, UK.

mayue.shi@eng.ox.ac.uk and m.shi16@imperial.ac.uk.