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. In regions where serological testing capacity is limited, AI-enhanced electrocardiogram (ECG) screening provides a critical diagnostic alternative. However, existing machine learning approaches face challenges such as limited accuracy, poor interpretability, reliance on large labeled datasets, and more importantly, weak integration with evidence-based clinical diagnostic indicators.

We propose a retrieval-augmented generation framework, CardioRAG, integrating large language models with interpretable ECG-based clinical features, including right bundle branch block, left anterior fascicular block, and heart rate variability metrics. The framework uses variational autoencoder-learned representations for semantic case retrieval, providing contextual cases to guide clinical reasoning. Evaluation demonstrated high recall performance of 89.80%, with a maximum F1 score of 0.68 for effective identification of positive cases requiring prioritized serological testing. CardioRAG provides an interpretable, clinical evidence-based approach particularly valuable for resource-limited settings, demonstrating a pathway for embedding clinical indicators into trustworthy medical AI systems.

1. Introduction

Chagas disease is a neglected tropical disease caused by Trypanosoma cruzi, 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 priori-tized serological testing and more efficient resource allocation [3].

In recent years, modern data-driven approaches have enabled new paradigms for Chagas detection. Advanced machine-learning methods can model non-linear relationships between disease status and multivariate ECG signals [4, 5]. However, current methods exhibit persistent limitations: (i) performance instability across domains due to population shift and limited calibration [6], (ii) limited clinical interpretability hindering trust and adoption [7], 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 [8] integrating interpretable ECG clinical features with large language model-based diagnostic reasoning. Our approach 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-aware case retrieval, enabling effective similarity matching with limited training data. (3) Empirical demonstration that prompt simplification and balanced case retrieval optimize performance for smaller language models, achieving 58.59% accuracy and 87.76% recall in zero-shot learning.

2. Methodology

We propose a comprehensive framework for automated Chagas disease detection that integrates deep learning-based ECG representation learning with retrieval-augmented generation (RAG) [9] for enhanced diagnostic
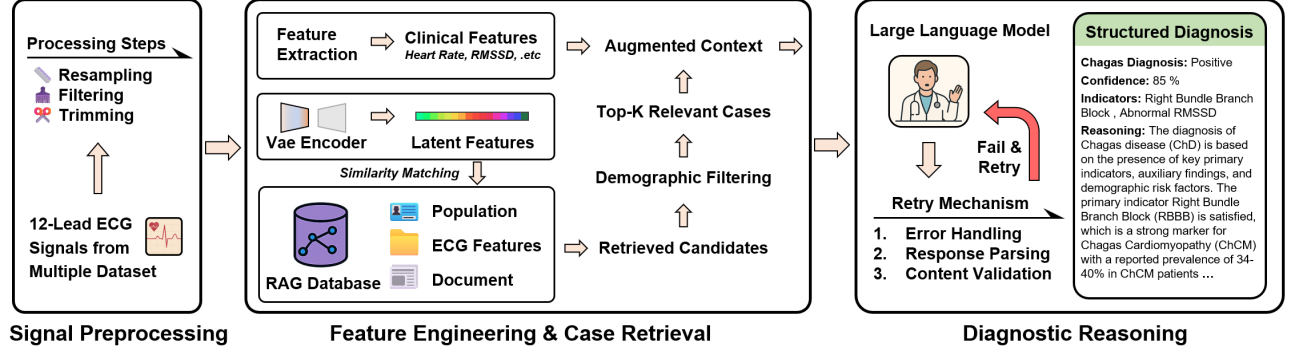
Figure 1: The CardioRAG Framework for Chagas disease diagnosis from 12-lead ECG signals. The system preprocesses raw ECG data, extracts clinical and latent features via VAE, retrieves relevant cases from a RAG database, and generates structured diagnoses with confidence scores using a large language model.

reasoning. As shown in Figure 1, the system processes 12-lead ECG alongside patient demographic data (age, sex) recordings through three main stages: (1) extraction of clinical features from ECG signals, (2) VAE-based representation learning for semantic similarity [10], and (3) RAG-enhanced diagnostic decision making with large language models (LLMs).

2.1. Data sources and preprocessing

This study utilized three publicly available ECG datasets from the PhysioNet Challenge [11]. The SaMi-Trop dataset [12] (1,631 records, 400 Hz) contains validated positive cases from Brazil with serologically confirmed Chagas disease. The PTB-XL dataset [13] (21,799 records, 500 Hz) serves as negative controls from European patients in non-endemic regions. The CODE-15% dataset [14] (300,000+ records, 400 Hz) provides mixed labels from Brazilian patients with self-reported Chagas status.

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 [15] 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 [16]. For conduction disorders, we implemented automated detection of right bundle branch block (RBBB) and left anterior fascicular block (LAFB) using Minnesota Code criteria [17]. 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 detec-

tion.

Table 1: ECG parameters of conduction disorders

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 [18]. 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, 9].

Variational autoencoder for signal embedding. We employ a variational autoencoder (VAE) architecture [10] to learn compact ECG representations that support effective similarity search. The encoder consists of four residual blocks [19] 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:

$$L = L_{\text{recon}} + \beta \cdot L_{\text{KL}} \quad (1)$$

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 composite similarity score:

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

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. For example, a positive case generated the following clinical reasoning.

> ### 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.

3.   Results and analysis

Our CardioRAG framework could not be evaluated using the standard PhysioNet Challenge 2025 methodology due to technical constraints: the local storage limit prohibits inclusion of large language models or API connectivity required for our system. Additionally, our zero-shot learning paradigm fundamentally differs from the Challenge's supervised training approach.

Therefore, we evaluated the proposed framework using the DeepSeek-R1:1.5b language model [20] 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 engi-

neering, and the effect of RAG retrieval strategies on diagnostic performance.
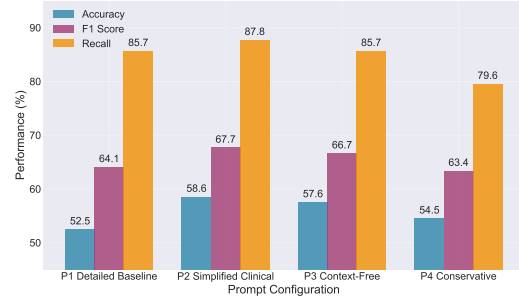
3.1.   Prompt engineering

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).

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 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 markedly 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.

The performance degradation observed at k=16 (52.53% accuracy) may be attributed to the introduction of exces-

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

sive 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 importance 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. Our results show that neither maximal information provision nor extreme simplification yields optimal performance. Instead, balanced, focused contextual guidance appears to achieve the best diagnostic reasoning outcomes without cognitive overload.

4. Conclusion

Our CardioRAG framework demonstrates the potential of integrating retrieval-augmented generation with clinical ECG features for Chagas disease screening, achieving 58.59% accuracy and 87.76% recall with consistently high recall across configurations. Our evaluation reveals that simplified prompts outperformed detailed descriptions; moderate case retrieval (k=8) with balanced retrieval achieved optimal performance; and the 58-59% accuracy ceiling may reflect current model limitations, warranting evaluation of larger LLMs. The framework's high recall performance makes it valuable for initial screening and patient triaging for serological testing, with future work focusing on improving specificity through enhanced feature selection and RAG optimization.

References

[1] World Health Organization. Chagas disease, 2025. URL https://www.who.int/health-topics/chagas-disease. Accessed: August 10, 2025.

[2] Acquatella H. Echocardiography in chagas heart disease. Circulation 2007;115(9):1124–1131.

[3] Alkmin MB, et al. Brazilian national service of telediagnosis in electrocardiography. Studies in health technology and informatics 2019;.

[4] Silva LEV, et al. Prediction of echocardiographic parameters in chagas disease using heart rate variability and machine learning. Biomedical Signal Processing and Control 2021;67:102513.

[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] 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.

[7] 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.

[8] Ng KKY, Matsuba I, Zhang PC. Rag in health care: a novel framework for improving communication and decision-making by addressing llm limitations. Nejm Ai 2025; 2(1):AIra2400380.

[9] Lewis P, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546, 2020; .

[10] Kingma DP, Welling M. Auto-encoding variational bayes. In Proceedings of the 2nd International Conference on Learning Representations. 2014; .

[11] PhysioNet. George b. moody physionet challenge 2025, 2025. URL https://moody-challenge.physionet.org/2025.

[12] Ribeiro ALP, et al. Sami-trop: 12-lead ecg traces with age and mortality annotations, June 2021. URL https://doi.org/10.5281/zenodo.4905618.

[13] Wagner P, et al. Ptb-xl, a large publicly available electrocardiography dataset. Scientific data 2020;7(1):1–15.

[14] Ribeiro AH, et al. Code-15%: a large scale annotated dataset of 12-lead ecgs, June 2021. URL https://doi.org/10.5281/zenodo.4916206.

[15] Makowski D, et al. Neurokit2: A python toolbox for neurophysiological signal processing. Behavior research methods 2021;53(4):1689–1696.

[16] Biolo A, Ribeiro AL, Clausell N. Chagas cardiomyopathy—where do we stand after a hundred years? Progress in cardiovascular diseases 2010;52(4):300–316.

[17] Prineas RJ, Crow RS, Zhang ZM. The Minnesota code manual of electrocardiographic findings. Springer, 2010.

[18] Ribeiro ALP, et al. Power-law behavior of heart rate variability in chagas' disease. The American journal of cardiology 2002;89(4):414–418.

[19] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition CVPR 2015;770–778.

[20] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Address for correspondence:

Mayue Shi
Institute of Biomedical Engineering, Old Road Campus Research Building, University of Oxford, Oxford OX3 7DQ, UK.
mayue.shi@eng.ox.ac.uk and m.shi16@imperial.ac.uk.