CLIC: CONTEXTUAL LANGUAGE-INFORMED CARDIAC PATHOLOGY CLASSIFICATION

Giovani D. Lucafó, Rafael da Costa Silva, João Lucas Luz Lima Sarcinelli,
Andre Guarnier De Mitri & Diego Furtado Silva
Institute of Mathematical and Computer Sciences
Universidade de São Paulo
São Carlos, São Paulo, Brazil
{giovanilucafo, rafael.csilva, andremitri,
joao.luz,diegofsilva}@usp.br

ABSTRACT

The electrocardiogram (ECG) is the gold standard for non-invasive diagnosis of cardiac pathologies and is a fundamental pillar of cardiovascular medicine. Recent progress in deep learning has led to the development of robust automated classifiers that achieve high performance by processing raw physiological signals. However, in clinical practice, diagnosis is rarely based solely on the signal. Cardiologists commonly support their interpretation with the patient's characteristics and the specific data-acquisition context. Despite this, most current algorithms remain restricted to signal-only analysis, failing to integrate technical metadata and demographic variables. This paper proposes **C**ontextual **L**anguage-**I**nformed **C**ardiac pathology classification (CLIC), a multimodal framework that significantly enhances diagnostic precision by encoding these variables through natural language. We demonstrate that translating patient-level contextual data into descriptive text provides an informative anchor that helps the model disambiguate complex physiological patterns. We further investigate the use of Large Language Models to synthesize richer clinical descriptions and observe that, while these generated texts remain competitive, controlled template-based contextual clinical text leads to consistent improvements in downstream classification performance.

Track: Research

1 INTRODUCTION

Modern machine learning-based applications increasingly rely on large, heterogeneous datasets that collect information from multiple sources and modalities. In this context, multimodal learning has emerged as an approach to building richer, more complex representations (Li & Tang, 2025). However, compared to domains such as vision and language, the integration of different modalities with health time series remains less explored (Trirat et al., 2024). Few studies address the fusion of signals, clinical notes, and structured records in a unified framework (Liu et al., 2026; Li et al., 2023).

In healthcare, where time series are widely used to monitor patients, the limitations of unimodal approaches are evident. Medical decision-making relies on multiple sources of information, including physiological signals, demographic data, and textual reports, which provide complementary views of patient health. Relying on a single modality can lead to incomplete or suboptimal diagnosis, whereas richer representations can improve classification performance. Despite the success of deep learning for physiological time series, most existing architectures focus only on signal data and do not leverage the multimodal nature of medical data (Liu et al., 2024; Wang et al., 2023).

In this work, we address this gap in the context of cardiology by proposing **C**ontextual **L**anguage-**I**nformed **C**ardiac pathology classification (CLIC), a multimodal supervised learning framework that integrates electrocardiogram (ECG) time series signals with textual information for cardiac pathology classification. As depicted in Figure 1, rather than treating text as an auxiliary input, CLIC investigates how contextual language representations can enrich ECG-based feature learning, leading to more informative representations for downstream classification tasks. Specifically, we

explore two multimodal configurations of CLIC, which will be detailed further in the next section. These configurations allow us to analyze the impact of increasingly expressive textual context on multimodal representation learning.

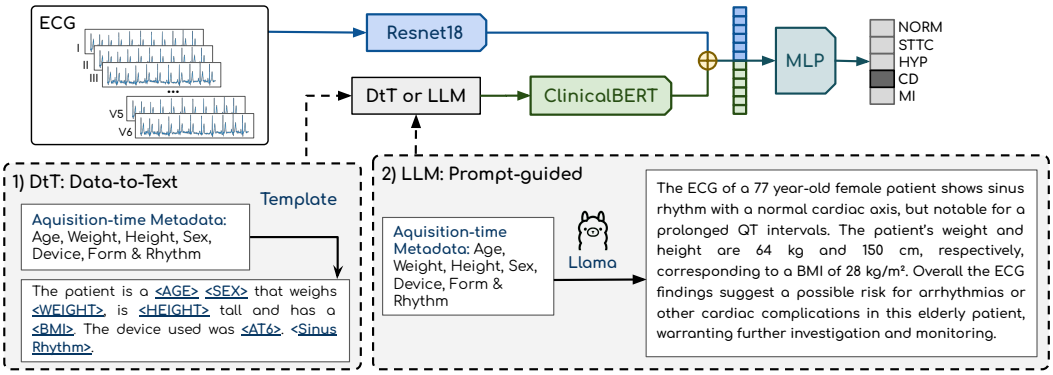

Figure 1: Illustration of the **C**ontextual **L**anguage-**I**nformed **C**ardiac pathology classification (CLIC) framework. The framework consists of two input data workflows: a Resnet18 that receives a 12-lead ECG as input, and a ClinicalBERT that receives a contextual clinical text generated via a template-based strategy called **Data-to-text** (1) and a **Prompt-guided** strategy that uses Llama (2).

Unlike many multimodal approaches in physiological signal analysis that rely on additional clinical sources such as electronic health records (EHRs) (Kim et al., 2025; Lalam et al., 2023; Yang & Wu, 2021; Lyu et al., 2023), the proposed CLIC framework operates in a self-contained setting, using only information directly available at acquisition time.

Our results show that incorporating contextual variables improves classification performance and representation quality compared to ECG-only baselines. However, the integration strategy is important. Direct concatenation of numerical attributes provides a strong and stable baseline, yet transforming metadata into natural language yields the best overall results, especially in a deterministic Data-to-Text setting with the ClinicalBERT language encoder. In contrast, LLM-generated reports do not consistently improve the semantic quality of the embeddings beyond structured approaches. Although CLIC is designed to generalize across signal types and datasets, this study focuses on ECG-based diagnosis as a case study to evaluate different strategies for combining physiological signals with structured and text-based contextual information.

2 METHODOLOGY

We formulate the addressed problem as a supervised multiclass classification task for ECG-based pathology classification. We conduct our experiments using the PTB-XL dataset (Wagner et al., 2020), a large-scale publicly available ECG benchmark. This dataset meets our computational constraints and provides a rich set of structured metadata, including diagnostic statements and clinical annotations, which support the multimodal experimental setting considered in this work. The classification is performed over five categories: Normal ECG (NORM), Myocardial Infarction (MI), ST/T Change (STTC), Conduction Disturbance (CD), and Hypertrophy (HYP). We followed the standard split into training, validation, and test sets and observed a notable class imbalance, particularly for Conduction Disturbances and Hypertrophy classes, posing additional challenges for supervised learning. The signals in PTB-XL consist of 500 Hz 12-lead ECGs, accompanied by the patient's data, such as sex, age, and weight, and automatically annotated rhythm and morphology information. We followed the dataset's recommended stratification using the `strat_fold` column. Folds 1 through 8 were allocated for training, fold 9 for validation, and fold 10 for testing. The target value corresponds to a single diagnostic class.

As the base model for the ECG, we employ an unimodal ECG classifier based on a ResNet18 (He et al., 2015) architecture adapted for time series modality. The ECG encoder maps each input signal to a 512-dimensional embedding, which is subsequently fed into a multilayer perceptron (MLP)

classifier. The MLP consists of two hidden layers with 256, 64, and 5 units, in this order. We employ the same architecture as the backbone in the multimodal scenario.

To incorporate contextual patient information, we evaluate two strategies to construct textual representations from ECG metadata, as shown in Figure 1. These text representations are combined with the ECG signal embeddings. Both strategies use the same metadata variables: age, sex, height, weight, collection device, signal morphology, and rhythm. The height and weight attributes were used to calculate body mass index (BMI), a commonly used metric for weight-related health risks. This selected set of variables corresponds to information that can be collected at acquisition time by health professionals in a clinical environment. The multimodal strategies are:

(1) Data to text: This method converts structured metadata into short natural language descriptions using template-based substitution. It provides deterministic and consistent textual context while preserving clinical reasoning.

(2) Prompt guided: This method generates more complex clinical descriptions using a large language model. In our experiments, we use the 8 billion parameter Llama 3.1 (Touvron et al., 2023) model to produce a concise clinical ECG report conditioned on available metadata.

We designed a prompt to guide Llama in generating the ECG reports. The model was instructed to act as a cardiology specialist and produce a single paragraph that simulates a clinical report when given the data described above. To ensure factual consistency, we constrained the model to use only the provided information and to avoid generating false information (e.g., heart rate values not present in the input).

The generated text from both methods were encoded using a pretrained ClinicalBERT (Huang et al., 2020) model, resulting in a 768-dimensional embedding. The ClinicalBERT parameters are frozen during the model training. Furthermore, we evaluate an additional baseline in which the same metadata used in the multimodal configurations are encoded as numerical features and directly concatenated with the ECG embeddings before classification via the MLP. This baseline was designed to assess whether a language encoder such as BERT can capture extrinsic information beyond what is available through direct numerical feature concatenation. We refer to this model as **ECG+Attr**.

The models were trained for up to 1000 epochs, with an early stopping patience of 50 epochs. We used a batch size of 16 across all sets, with data shuffling applied exclusively to the training dataloader. The network was optimized using the Adam optimizer with a learning rate of 10^{-3}, and the loss was computed using Binary Cross-Entropy with unnormalized logits. Finally, to ensure statistical reliability, all experiments were executed across 5 independent runs, and we report the average performance metrics along with their standard deviations.

Both unimodal settings use a two-layer MLP for classification. In multimodal settings, we extend this architecture by adding an MLP layer to aggregate the additional embeddings introduced during the fusion stage. The source code is available at the repository [1].

3 RESULTS

Table 1 summarizes the performance of both unimodal and multimodal architectures across all independent runs. Overall, the results demonstrate that CLIC consistently improves ECG classification when meaningful textual context is incorporated.

The unimodal ECG baselines achieve strong performance for the majority class (NORM) but struggle with minority and clinically complex conditions, particularly Conduction Disturbance (CD), reflecting the limitations of ECG-only representations. The LLM-generated (CLIC-LLM) multimodal variant shows modest gains for STTC. However, the performance does not consistently surpass the ECG+Attr unimodal baseline across all classes.

Alternatively, CLIC-DtT achieves the best overall performance, outperforming almost all baseline metrics, including general accuracy and macro-average, while maintaining low variance across runs. Particularly for minority classes, improvements are more pronounced, including CD and STTC, suggesting that structured template-based semantic text provides a more stable integration of demographic and signal metadata than both raw attribute concatenation and LLM-generated descriptions.

[1] https://github.com/giovanidl/CLIC

One possible explanation is that LLM-generated reports introduce lexical variability and implicit reformulations that may dilute information. As a result, the textual encoder may extract less stable representations than those from the direct and standardized template sentences.

For Conduction Disturbance (CD), which is known to be more frequent in older male patients (Haimovich et al., 2024; Monin et al., 2016), the inclusion of contextual metadata appears to provide clinically relevant prior information beyond the ECG signal alone. When encoded with a domain-specific language model, the textual representation may incorporate extrinsic medical knowledge about age- and sex-related comorbidities. This additional semantic structure can help the model capture associations that are not explicitly represented in the signal features.

Table 1: Performance comparison (mean ± std) across models using class-wise F1-score, Recall, General Accuracy and Macro-F1. **Bold** and underline values indicate the model with the best and second-best metric for that class, respectively.

Model	NORM (1004)		MI (544)		STTC (283)		CD (122)		HYP (245)	
	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall
ECG Only	0.831±0.009	0.862±0.037	0.705±0.009	0.679±0.037	0.570±0.023	0.597±0.025	0.360±0.013	0.308±0.033	0.553±0.012	0.522±0.026
ECG+Attr	0.872±0.005	0.897±0.023	0.766±0.008	0.791±0.013	0.684±0.021	0.701±0.044	0.424±0.017	0.334±0.036	0.594±0.016	**0.528±0.038**
CLIC-DtT	**0.887±0.010**	**0.936±0.018**	**0.781±0.013**	**0.792±0.055**	**0.767±0.005**	**0.761±0.050**	**0.490±0.051**	**0.421±0.069**	**0.597±0.030**	0.491±0.048
CLIC-LLM	0.871±0.003	0.909±0.011	0.757±0.009	0.750±0.025	0.695±0.009	0.712±0.063	0.401±0.060	0.341±0.085	0.587±0.029	0.526±0.082

Overall Performance

Metric	ECG Only	ECG+Attr	CLIC-DtT	CLIC-LLM
Macro-F1	0.604±0.019	0.668±0.006	**0.704±0.015**	0.662±0.007
General Accuracy	0.714±0.017	0.773±0.005	**0.800±0.008**	0.770±0.003

To further examine the impact of multimodality on representation learning, we analyzed the embedding space using UMAP (McInnes et al., 2020), as shown in Figure 2. Analyzing the latent space structure, the ECG-only representations are sparse and unstructured, whereas the ECG+Attr embeddings form dense but highly mixed clusters. The CLIC-DtT model demonstrates superior organization, forming a continuous manifold with more evident separation between normal and pathological classes. This suggests that CLIC-DtT successfully integrates multimodal signals without allowing static attributes to dominate the embedding space.

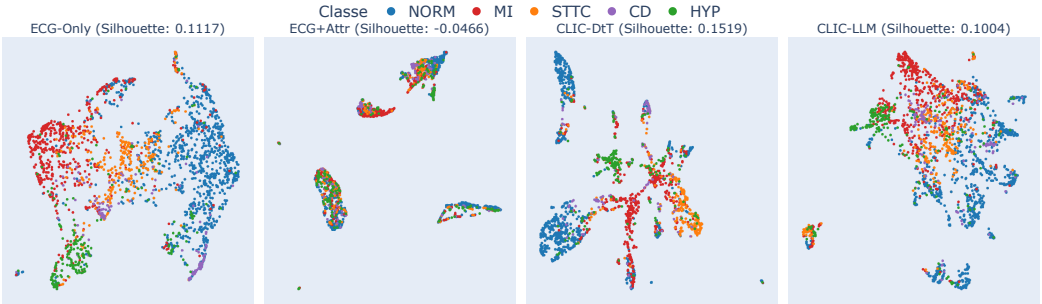

Figure 2: UMAP visualization of embedding distributions across models.

4 Conclusion

Our results show that incorporating contextual patient information improves performance compared to the ECG-only baseline. Even being a simple addition, the direct concatenation of structured attributes ECG+Attr consistently overcomes the ECG-only scenario. Among the multimodal strategies, the Data-to-Text approach achieves greater stability and better overall performance, while the LLM-generated variant yields similar results on some metrics but does not consistently surpass the structured attribute baseline. These findings indicate that deterministic textual generation can provide more stable supervision. In practice, the LLM text did not sufficiently enhance the semantic content of the embeddings to outperform the template-based text in the assessed protocol, suggesting that the additional linguistic complexity did not translate into more discriminative multimodal representations, potentially requiring task-specific alignment or fine-tuning to improve performance. We

also observed that contextual metadata is particularly beneficial for the CD class, which is strongly associated with demographic factors. These observations suggest that semantic information captured by the ClinicalBERT encoder boosted the representations' discriminability for classes that are more susceptible to demographic factors.

While this work focuses on a single downstream classification task, it opens directions for future research. These include exploring alternative input signals (e.g., EEG and PPG), novel multimodal fusion strategies, and language model fine-tuning. Overall, CLIC demonstrates that integrating structured and text-based contextual information is a feasible complement to ECG representation learning and it also indicates that deterministic texts may be as effective as fully generative approaches.

ACKNOWLEDGMENTS

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brazil. Process Number #2022/03176-1, #2023/05041-9, and #2025/13341-8.

REFERENCES

Julian S. Haimovich, Paolo Di Achille, Victor Nauffal, Pulkit Singh, Christopher Reeder, Xin Wang, Gopal Sarma, Jelena Kornej, Emelia J. Benjamin, Anthony Philippakis, Puneet Batra, Patrick T. Ellinor, Steven A. Lubitz, and Shaan Khurshid. Frequency of electrocardiogram-defined cardiac conduction disorders in a multi-institutional primary care cohort. *JACC: Advances*, 3(7):101004, 2024. doi: 10.1016/j.jacadv.2024.101004.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2020. URL https://arxiv.org/abs/1904.05342.

Junmo Kim, Young-Kwan Kim, and Kwangsoo Kim. Multimodal electronic health record foundation models with electrocardiogram for cardiovascular disease prediction. *medRxiv*, Nov 2025. doi: 10.1101/2025.11.10.25339886. Preprint.

Sravan Kumar Lalam, Hari Krishna Kunderu, Shayan Ghosh, Kumar A. Harish, Ashim Prasad, Francisco Lopez-Jimenez, Samir Awasthi, Zachi I. Attia, Samuel J. Asirvatham, Paul Andrew Friedman, Rakesh Barve, and Melwin Babu. Ecg representation learning with multi-modal ehr data. *Transactions on Machine Learning Research*, 2023, November 2023. ISSN 2835-8856. Publisher Copyright: © 2023, Transactions on Machine Learning Research. All rights reserved.

Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ecg zero-shot learning, 2023. URL https://arxiv.org/abs/2303.12311.

Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey, 2025. URL https://arxiv.org/abs/2411.17040.

Xingyue Liu, Feizhong Zhou, Hanguang Xiao, Zhipeng Li, Shuai Liu, and Lingling Qian. A survey on large language models for medical time series. *Expert Systems with Applications*, 311:131364, 2026. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2026.131364. URL https://www.sciencedirect.com/science/article/pii/S0957417426002770.

Yuansan Liu, Sudanthi Wijewickrema, Christofer Bester, Stephen O'Leary, and James Bailey. Time series representation learning with supervised contrastive temporal transformer, 2024. URL https://arxiv.org/abs/2403.10787.

Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction, 2023. URL https://arxiv.org/abs/2208.10240.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

Jonathan Monin, Sébastien Bisconte, Aurore Nicaise, Anne-Pia Hornez, Olivier Manen, and Eric Perrier. Prevalence of intraventricular conduction disturbances in a large french population. *Annals of Noninvasive Electrocardiology*, 21(5):479–485, 2016. doi: 10.1111/anec.12331.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Patara Trirat, Yooju Shin, Junhyeok Kang, Youngeun Nam, Jihye Na, Minyoung Bae, Joeun Kim, Byunghyun Kim, and Jae-Gil Lee. Universal time-series representation learning: A survey, 2024. URL https://arxiv.org/abs/2401.03717.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset (version 1.0.1). PhysioNet, 2020. URL https://physionet.org/content/ptb-xl/1.0.1/. RRID:SCR_007345.

Yihe Wang, Yu Han, Haishuai Wang, and Xiang Zhang. Contrast everything: A hierarchical contrastive framework for medical time-series, 2023. URL https://arxiv.org/abs/2310.14017.

Bo Yang and Lijun Wu. How to leverage multimodal ehr data for better medical predictions?, 2021. URL https://arxiv.org/abs/2110.15763.