Transfering Clinical Knowledge into ECGs Representation: A Self-Supervised Approach for Interpretable, Unimodal-at-Inference Diagnosis

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep learning models have shown high accuracy in classifying electrocardiograms (ECGs), but their black box nature hinders clinical adoption due to a lack of trust and interpretability. To address this, we propose a novel three-stage training paradigm that transfers knowledge from multimodal clinical data (laboratory exams, vitals, biometrics) into a powerful, yet unimodal, ECG encoder. We employ a self-supervised, joint-embedding pre-training stage to create an ECG representation that is enriched with contextual clinical information, while only requiring the ECG signal at inference time. Furthermore, we leverage this enriched representation to provide clinically relevant explanations by training the model to predict associated laboratory abnormalities directly from the ECG embedding. Evaluated on the MIMIC-IV-ECG dataset, our model outperforms a standard signal-only baseline in multi-label diagnosis classification and successfully bridges a substantial portion of the performance gap to a fully multimodal model that requires all data at inference. Our work demonstrates a practical and effective method for creating more accurate and trustworthy ECG classification models. By converting abstract predictions into physiologically grounded explanations, our approach offers a promising path toward the safer integration of AI into clinical workflows.

1 Introduction

Electrocardiogram (ECG) exams are a fundamental diagnostic tool in medical practice, recording the heart's electrical activity to detect a wide range of cardiovascular conditions [Braunwald et al., 2015]. Their importance is substantial, especially considering that cardiovascular diseases remain the leading cause of global mortality [Roth et al., 2020]. With the advancement of artificial intelligence, the automatic classification of ECG exams using deep learning techniques has shown remarkable potential [Liu et al., 2021, Petmezas et al., 2022], offering high precision in identifying abnormalities and extracting complex information that can aid in diagnosis.

Despite the promising performance and significant accuracy achieved by deep learning models in healthcare, their adoption and trust among healthcare professionals remain a challenge [Reyes et al., 2020, Adeniran et al., 2024]. The main barrier to integrating these models into clinical practice lies in their black box nature—the lack of explainability and transparency about how decisions are made [Rosenbacke et al., 2024]. Physicians, accustomed to clinical reasoning based on evidence and causality, hesitate to trust systems whose internal processes are opaque, leading to mistrust and uncertainty regarding the safety and justification of the predictions [Koçak et al., 2025].

Frequently, explanations are presented as saliency maps, which often fail to bridge the trust gap due to a lack of robustness and misalignment with clinical concepts [Borys et al., 2023, Zhang et al.,

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2023]. To address this, we propose a novel multimodal training architecture that enriches an ECG model with knowledge from associated tabular clinical data. Instead of requiring all data modalities at inference, our approach uses a self-supervised, joint-embedding objective to transfer the rich context from laboratory values and vitals into a powerful, unimodal ECG encoder. This enriched encoder is not only more accurate but also inherently more interpretable. By training it to perform a secondary task of predicting lab abnormalities from the ECG signal alone, we create a system that can explain its diagnostic reasoning in terms of concrete, clinically relevant concepts.

In summary, we make the following contribution:

- We propose a joint-embedding pre-training framework to transfer knowledge from multimodal tabular data into a unimodal ECG encoder for the task of diagnosis classification;
- We introduce the prediction of laboratory abnormalities from the ECG embedding as a novel, clinically-grounded method for explaining the model's diagnostic outputs.

Related Work Our work integrates insights from three key areas. While deep learning for ECG classification is well-established, models often lack the trust of clinicians due to their "black box" nature [Reyes et al., 2020]. Current eXplainable AI (XAI) methods often rely on saliency maps, which can be misaligned with clinical reasoning [Borys et al., 2023]. Concurrently, self-supervised learning (SSL) has enabled the creation of powerful representations without labeled data [Zbontar et al., 2021], and multimodal learning seeks to create holistic models by fusing data sources [Kline et al., 2022]. However, most multimodal approaches require all data at inference time, hindering practicality [Alcaraz et al., 2025]. We are the first to unify these areas, using a multimodal SSL objective to distill knowledge into a practical, unimodal-at-inference model with a novel, clinically-grounded explanation mechanism. A detailed literature review is provided in the Appendix.

2 Method

Our approach is a three-stage training paradigm designed to create a powerful and interpretable ECG representation. We first pre-train a waveform encoder using a self-supervised, cross-modal objective to transfer knowledge from tabular clinical data. Subsequently, we fine-tune this encoder for two downstream relevant clinical tasks: a primary task of multi-label diagnosis classification Strodthoff et al. [2024]; and, a secondary task of laboratory values abnormality prediction Alcaraz and Strodthoff [2024].

2.1 Joint-embedding Pre-training

The primary goal of our pre-training stage is to learn a signal encoder, Φ_x, whose representations are enriched with the contextual information present in associated tabular clinical data. To achieve this, we frame the task as a self-supervised, joint-embedding problem where the model learns to align representations from these two distinct modalities.

For a given patient encounter, we have a raw ECG signal segment $x \in \mathbb{R}^{C \times L}$, where C is the number of leads and L is the sequence length, and a corresponding vector of tabular clinical data $m \in \mathbb{R}^{D}$, which includes demographics, vitals, and biometrics. The signal and tabular data are processed by their respective encoders, a powerful S4-based sequence model Φ_x and a MLP Φ_m. These backbones produce feature representations $h_x = \Phi_x(x)$ and $h_m = \Phi_m(m)$. Following the standard practice in self-supervised learning, these feature representations are then mapped into another embedding space using dedicated projector networks, Θ_x and Θ_m, $z_x = \Theta_x(h_x)$ and $z_m = \Theta_m(h_m)$, $z_x, z_m \in \mathbb{R}^{E}$, where E is the embedding dimension.

To align these embeddings without risking representational collapse, we employ the Barlow Twins loss function, $\mathcal{L}_{BT}$. This objective function does not rely on negative sampling and instead encourages the cross-correlation matrix computed between the embeddings Z_x and Z_m over a batch of N samples to be close to the identity matrix. The loss is composed of two terms: an invariance term that pulls the embeddings from the same patient together; and, a redundancy reduction term that decorrelates the different dimensions of the embedding vectors. The total pre-training objective is thus defined as in Equation 1.

$$\mathcal{L}_{je} = \mathcal{L}_{BT}(Z_x, Z_m) = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2 \tag{1}$$

where C is the cross-correlation matrix computed between the batch-normalized embeddings Z_x and Z_m, and λ is a hyperparameter balancing the two terms. By minimizing this loss, the signal encoder Φ_x is trained to produce representations h_x that are not only descriptive of the ECG signal itself but are also highly predictive of the rich, contextual information contained in the clinical data m. This process effectively transfers the clinical knowledge into the parameters of the waveform encoder.

2.2 Diagnosis Classification

Following the self-supervised pre-training stage, we leverage the learned, context-aware signal encoder Φ_x for the primary downstream task of clinical diagnosis prediction. To accomplish this, we adopt a standard fine-tuning protocol.

The pre-trained encoder Φ_x is partially frozen to prevent catastrophic forgetting of the transferred clinical knowledge. A new, task-specific classification head, Ψ_y, is then initialized and attached to the encoder. This head is a lightweight MLP that takes the feature representation $h_x = \Phi_x(x)$ and maps it to a logit for each of the K possible diagnoses.

The model is then trained on the subset of the data containing ground-truth diagnosis labels Y. The optimization objective is the standard Binary Cross-Entropy (BCE) loss, summed over all possible labels, as is common for multi-label classification problems. The classification loss is defined as in Equation 2.

$$\mathcal{L}_c = \mathcal{L}_{BCE}(Y, \hat{Y}) = -K^{-1} \sum_k^K [Y_k \log(\hat{Y}_k) + (1 - Y_k) \log(1 - \hat{Y}_k)] \tag{2}$$

where $Y \in \{0, 1\}^K$ is the one-hot encoded vector of true labels and $\hat{Y}$ is the vector of predicted logits from the model. This fine-tuning stage adapts the powerful, general-purpose representation learned during pre-training to the specific patterns required for the diagnostic task.

2.3 Reconstruction Finetuning

A primary motivation for our three-stage approach is to produce a waveform representation that is not only predictive but also interpretable. We train the signal encoder Φ_x to perform a reconstruction task of laboratory abnormalities that are related to the context seem in the pre-training.

For this purpose, we introduce a second, independent head Ψ_m, which is attached to the same pre-trained encoder Φ_x. This head is trained to predict a multi-label vector $M^* \in \{0, 1\}^P$, where P is the number of distinct laboratory abnormalities defined in our dataset (e.g., "Hemoglobin_high", "Urea Nitrogen_low"). Each element $m_p^* \in M^*$ is a binary indicator of a specific lab abnormality.

Similar to the classification phase, the reconstruction head Ψ_m is a lightweight MLP that takes the signal representation h_x and outputs a vector of logits $\hat{M}^* = \Psi_m(\Phi_x(x))$, one for each of the P lab abnormalities. The model is trained by minimizing the BCE loss, as defined in Equation 2, now applied to the lab abnormality targets m^* as in Equation 3.

$$\mathcal{L}_r = \mathcal{L}_{BCE}(M^*, \hat{M}^*) \tag{3}$$

The ability of the model to successfully perform this pseudo-reconstruction task, using only the ECG signal as input, serves to provide a mechanism for model explainability. For any given diagnosis prediction, we can now also output the concurrent lab abnormalities predicted from the same ECG embedding, offering clinicians a direct, data-driven insight into the potential physiological drivers behind the model's primary prediction. Essentially we offer something like: "The model predicts diagnosis Y, and it's doing so as it also is predicting a associated lab abnormality M_p^* (e.g., high creatinine)."

Table 1: Main results comparing our joint-embedding (JE) and reconstruction approach to established baselines. Filled circles (•) indicate a data requirement, while empty circles (○) indicate it is not required.

	Evaluation		Inference Requirement			Training Requirement		
Model	**Diagnoses**	**Lab**	**Routine**	**Lab**	**Diagnoses**	**Routine**	**Lab**	**Diagnoses**
Supervised Signal-Only	0.768	-	○	○	○	○	○	•
Multimodal Lab Prediction	-	**0.762**	•	○	○	•	•	○
Multimodal Classification	**0.826**	-	•	•	○	•	•	•
JE + Reconstruction	0.795	0.701	○	○	○	•	•	•

3 Experimental Setup and Results

We evaluate our paradigm on the MIMIC-IV-ECG dataset [Johnson et al., 2023], focusing on the emergency department subset per the protocol in Alcaraz et al. [2025], which also defines our data splits and backbone architecture, we detail the experiment setup in the Appendix. We report macro-averaged AUROC and compare against three key baselines: a **Supervised Signal-Only** model [Strodthoff et al., 2024]; a **Multimodal Lab Prediction** model serving as an upper bound for the reconstruction task [Alcaraz and Strodthoff, 2024]; and a fully **Multimodal Classification** model serving as a practical upper bound for diagnosis [Alcaraz et al., 2025].

As shown in Table 1, the results suggest the effectiveness of our approach. Our final model, which operates using only the ECG signal at inference time, achieves a classification performance improvement over the signal-only baseline indicating the presence of tabular data knowledge transfer. While it does not reach the performance of the multimodal upper bound which requires full access to all data at test time, our method successfully bridges a significant portion of this performance gap without sacrificing the practicality of unimodal deployment.

Furthermore, we also assess the explanations by evaluating its performance on the laboratory abnormality prediction task. Our approach achieves similar performance to its correspondence baseline, again we treat them as an upper bound because of the data requirement at inference.

4 Conclusion

In this work, we introduced a novel training paradigm that creates a powerful, practical, and interpretable ECG-based diagnostic model. Our primary contribution is demonstrating that a self-supervised, joint-embedding objective can effectively transfer knowledge from multimodal clinical data into a unimodal encoder, retaining the practical advantage of requiring only the ECG at inference time. Furthermore, we established a new mechanism for model interpretability by predicting associated laboratory abnormalities from the same latent representation, offering a more intuitive alternative to input-space heatmaps. A key limitation is the risk of catastrophic forgetting in our sequential training, which we mitigated by freezing early layers; future work could explore advanced continual learning techniques, a detailed discussion is provided in the Appendix. We also plan to expand our framework to include unstructured text. By converting abstract predictions into physiologically grounded explanations, our approach offers a promising path toward the safer integration of AI into clinical workflows.

References

Adewale Abayomi Adeniran, Amaka Peace Onebunne, and Paul William. Explainable ai (xai) in healthcare: Enhancing trust and transparency in critical decision-making. *World J. Adv. Res. Rev*, 23:2647–2658, 2024.

Juan Miguel Lopez Alcaraz and Nils Strodthoff. Cardiolab: Laboratory values estimation and monitoring from electrocardiogram signals–a multimodal deep learning approach. *arXiv preprint arXiv:2411.14886*, 2024.

Juan Miguel Lopez Alcaraz, Hjalmar Bouma, and Nils Strodthoff. Enhancing clinical decision support with physiological waveforms—a multimodal benchmark in emergency care. *Computers in Biology and Medicine*, 192:110196, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Katarzyna Borys, Yasmin Alyssa Schmitt, Meike Nauta, Christin Seifert, Nicole Krämer, Christoph M Friedrich, and Felix Nensa. Explainable ai in medical imaging: An overview for clinical practitioners–beyond saliency-based xai approaches. *European journal of radiology*, 162:110786, 2023.

Eugene Braunwald, Douglas L Mann, Douglas P Zipes, P Libby, and RO Bonow. Braunwald's heart disease: a textbook of cardiovascular medicine. In *Braunwald's heart disease: A textbook of cardiovascular medicine*, pages 1028–1028. 2015.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. Survey of explainable ai techniques in healthcare. *Sensors*, 23(2):634, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training 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)*, pages 4171–4186, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Jyoti Gupta and KR Seeja. A comparative study and systematic analysis of xai models and their applications in healthcare. *Archives of Computational Methods in Engineering*, 31(7):3977–4002, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Tim Hulsen. Explainable artificial intelligence (xai): concepts and challenges in healthcare. *AI*, 4(3): 652–666, 2023.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171, 2022.

Burak Koçak, Andrea Ponsiglione, Arnaldo Stanzione, Christian Bluethgen, João Santinha, Lorenzo Ugga, Merel Huisman, Michail E Klontzas, Roberto Cannella, and Renato Cuocolo. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and interventional radiology*, 31(2):75, 2025.

Qika Lin, Yifan Zhu, Xin Mei, Ling Huang, Jingying Ma, Kai He, Zhen Peng, Erik Cambria, and Mengling Feng. Has multimodal learning delivered universal intelligence in healthcare? a comprehensive survey. *Information Fusion*, page 102795, 2024.

Xinwen Liu, Huan Wang, Zongjin Li, and Lang Qin. Deep learning in ecg diagnosis: A review. *Knowledge-Based Systems*, 227:107187, 2021.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

Georgios Petmezas, Leandros Stefanopoulos, Vassilis Kilintzis, Andreas Tzavelis, John A Rogers, Aggelos K Katsaggelos, and Nicos Maglaveras. State-of-the-art deep learning methods on electrocardiogram data: systematic review. *JMIR medical informatics*, 10(8):e38454, 2022.

Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2(3): e190043, 2020.

Rikard Rosenbacke, Åsa Melhus, Martin McKee, and David Stuckler. How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: Systematic review. *JMIR AI*, 3:e53207, 2024.

Gregory A Roth, George A Mensah, Catherine O Johnson, Giovanni Addolorato, Enrico Ammirati, Larry M Baddour, Noël C Barengo, Andrea Z Beaton, Emelia J Benjamin, Catherine P Benziger, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the gbd 2019 study. *Journal of the American college of cardiology*, 76(25):2982–3021, 2020.

Nils Strodthoff, Juan Miguel Lopez Alcaraz, and Wilhelm Haverkamp. Prospects for artificial intelligence-enhanced electrocardiogram as a unified screening tool for cardiac and non-cardiac conditions: an explorative study in emergency care. *European Heart Journal-Digital Health*, 5(4): 454–460, 2024.

Ran Xiao, Cheng Ding, Xiao Hu, Gari D Clifford, David W Wright, Amit J Shah, Salah Al-Zaiti, and Jessica K Zègre-Hemsey. Integrating multimodal information in machine learning for classifying acute myocardial infarction. *Physiological Measurement*, 44(4):044002, 2023.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

Jiajin Zhang, Hanqing Chao, Giridhar Dasegowda, Ge Wang, Mannudeep K Kalra, and Pingkun Yan. Revisiting the trustworthiness of saliency methods in radiology ai. *Radiology: Artificial Intelligence*, 6(1):e220221, 2023.

Appendix

A Related Work

Our work is primarily related to prior work on eXplainable Artificial Intelligence (XAI), self-supervised learning, and multimodal training in healthcare. To the best of our knowledge, our work is the first attempt to unify insights from these areas for automatic ECG classification.

eXplainable Artificial Intelligence in Healthcare The growing adoption of deep learning models in medical applications, such as the automatic classification of electrocardiograms (ECG), has demonstrated impressive performance in identifying various cardiac abnormalities [Liu et al., 2021, Petmezas et al., 2022]. However, the black box nature of many of these deep learning algorithms limits their widespread acceptance by healthcare professionals [Reyes et al., 2020, Adeniran et al., 2024, Rosenbacke et al., 2024, Koçak et al., 2025]. The lack of transparency in the decision-making processes of these models generates reluctance, especially in high-risk environments. Traditional XAI methods, such as activation maps in the input space, are often employed to visualize the regions of the input that most influence a model's prediction [Chaddad et al., 2023]. However, in the medical context, these visual explanations are often insufficient. Cardiologists interpret ECGs based on established clinical concepts, such as P-wave morphology or ST-segment elevation, not on raw signal intensity. Explanations that do not resonate with this established clinical lexicon are difficult to interpret and, therefore, less useful for diagnostic validation or learning [Borys et al., 2023, Zhang et al., 2023]. In response to these limitations, there is a growing demand for more intuitive and clinically relevant XAI approaches that can bridge the gap between abstract AI predictions and human understanding, thereby fostering greater trust and facilitating the integration of AI into clinical decision support systems [Hulsen, 2023, Gupta and Seeja, 2024].

Self-Supervised Learning Self-supervised learning (SSL) has emerged as a powerful paradigm for learning meaningful representations from unlabeled data. Methodologies can be broadly categorized into two families. The first, joint-embedding methods, learn by enforcing consistency between the embeddings of two or more augmented views of the same data point. These approaches, such as SimCLR [Chen et al., 2020], DINO [Caron et al., 2021], Barlow Twins [Zbontar et al., 2021], and VICReg [Bardes et al., 2021], primarily use a discriminative objective to pull representations of the same instance together while preventing representational collapse. The second family consists of reconstruction-based methods, which learn by reconstructing the original input from a corrupted or masked version. This generative approach includes classic models like Variational Autoencoders (VAEs) [Kingma and Welling, 2013] and Generative Adversarial Networks (GANs) [Goodfellow et al., 2020], as well as modern powerhouses like Denoising Diffusion models [Ho et al., 2020] and Masked Language Models like BERT [Devlin et al., 2019]. Our pre-training stage falls into the joint-embedding category, leveraging its strength in learning abstract, transferable features.

Multimodal Training in Healthcare Clinical reality is inherently multimodal; a patient's status is defined by a combination of physiological signals, lab values, imaging data, and clinical notes. Multimodal machine learning aims to integrate these heterogeneous data sources to build more robust and accurate models for a holistic understanding of patient health [Kline et al., 2022, Lin et al., 2024]. A common strategy is feature fusion, where representations from different modalities are combined at an early, intermediate, or late stage to make a final prediction [Xiao et al., 2023]. For instance, a late-fusion model might combine the outputs of separate encoders for ECGs and tabular data, as is done in our multimodal baseline [Alcaraz et al., 2025]. While powerful, these fusion-based approaches typically require all data modalities to be present at inference time, which can be a significant practical barrier in clinical workflows. Our work presents an alternative: using multimodal data during training via a knowledge transfer objective to enrich a unimodal encoder, thereby gaining the benefits of multimodal context without the constraint of multimodal input during deployment.

Automatic ECG Classification Deep learning has become the state-of-the-art approach for the automatic interpretation of ECGs, capable of identifying a wide range of cardiac and even non-cardiac conditions with high accuracy. Our work builds directly upon recent advancements in this area. We situate our contribution relative to three distinct but related lines of research: (i) supervised, unimodal models that classify diagnoses from the ECG signal alone [Strodthoff et al., 2024]; (ii) multimodal

models designed to predict laboratory abnormalities from ECG and routine clinical data [Alcaraz and Strodthoff, 2024]; and (iii) state-of-the-art multimodal fusion models that achieve high performance but require all data sources at inference time, serving as a practical upper bound for the classification task [Alcaraz et al., 2025].

B Three-stage algorithm

We organised the approach into a fluxogram, as despicted in Figure 1.

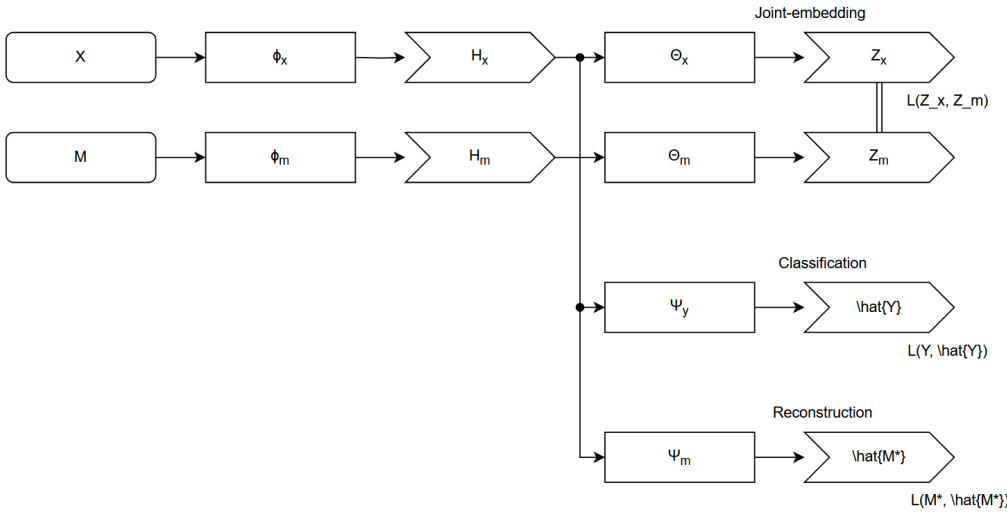

Figure 1: Schematic of the proposed multimodal training architecture. Our method begins with a (i) joint-embedding pre-training stage, where an ECG encoder Φ_x learns to produce representations H_x that are aligned with embeddings from tabular clinical data M. Subsequently, this single, enriched encoder is finetuned for two downstream tasks: (ii) a primary multi-label diagnosis classification task $\mathcal{L}_c$; and, (iii) a secondary laboratory abnormality reconstruction-like task $\mathcal{L}_r$, which provides the mechanism for model interpretability. Crucially, only the ECG signal X is required at inference time.

C Detailed Experiments

We evaluate our proposed joint-embedding and reconstruction learning paradigm on two distinct downstream tasks: multi-label diagnosis classification; and, laboratory abnormality prediction. To conduct this evaluation, our study utilizes the MIMIC-IV-ECG dataset, a large, publicly available resource uniquely suited for our multimodal research. It contains over 200,000 12-lead electrocardiograms from patients admitted to the Beth Israel Deaconess Medical Center, with a critical linkage to the rich clinical data within the main MIMIC-IV database [Johnson et al., 2023]. This linkage provides access to a comprehensive set of corresponding tabular data for each ECG, including: (i) laboratory test results; (ii) vital signs and biometrics; and, (iii) ICD-coded discharge diagnoses. The availability of synchronously recorded signals and comprehensive clinical context is fundamental to our approach, as it enables the self-supervised pre-training and provides the ground-truth labels for our downstream tasks. Following established prior work, our experiments focus on the subset of data originating from the emergency department to reflect a challenging and practical clinical screening scenario [Strodthoff et al., 2024]. Our experimental setup, including data splits, backbone architecture, and evaluation algorithms, rigorously follows the protocol established in Alcaraz et al. [2025] to ensure a fair comparison with all baselines. For the purpose of the lab prediction task, we discriminate the tabular data into two types: the valuable laboratory exam values; and the remaining tabular features, which we term routine clinical data (biometrics and vital signs). All results are reported as the macro-averaged Area Under the Receiver Operating Curve (AUROC) on the held-out test set.

D Limitations

Sequential Training Paradigm A primary limitation of our work lies in the three-stage sequential training paradigm. This approach introduces the risk of catastrophic forgetting [McCloskey and Cohen, 1989], where the model may lose some of the rich, transferred knowledge from the pre-training stage during subsequent finetuning. While a joint, multi-task learning approach was considered, we found the simultaneous optimization of the three distinct loss functions to be highly challenging and unstable. To mitigate this, we partially froze the early layers of the signal encoder, preserving the foundational representations learned during pre-training. Future work could explore more advanced techniques, such as using the transferred representations as a regularization target during the classification phase, to more explicitly retain the pre-trained knowledge. Or training a standard multimodal *all-modes required* model and later train another similar signal encoder to replicate the representation of the former, given that it encodes knowledge from the other modes.

Evaluation of Explanations Our evaluation of the model's interpretability is indirect. We use performance on the laboratory abnormality prediction task as a proxy for the quality of the explanation, demonstrating that the ECG embedding contains physiologically relevant information. However, this assessment does not establish a causal relationship between the predicted abnormalities and the final diagnoses, nor does it explicitly highlight which specific abnormalities are most salient for a given diagnostic prediction from the clinician's perspective. The model learns strong correlations, but further investigation, potentially involving causal inference methods or direct clinician feedback studies, would be required to untangle these into clinically actionable insights.

Expanding Multimodal Learning to Text Our choice of a joint-embedding framework was deliberate, with future extensions in mind—in particular, the generalization of a $n > 2$ multimodal training with the integration of unstructured text data, such as clinical notes, which are available in MIMIC-IV dataset. This rich textual information could further improve classification performance and enable the generation of more natural, text-based explanations for the model's predictions. We plan to generalize the joint-embedding objective to align representations from all three modalities (ECG, tabular, text). Crucially, we propose using the ECG's latent space as a central anchor, aligning the other modalities to it. This *anchor-based* approach, inspired by the teacher-student approach in some SSL methods, is hypothesized to scale more effectively than the naive alternative of optimizing all pairwise combinations of modalities.