Article | Received 30 November 2025; Revised 30 January 2026; Accepted 24 February 2026; Published 16 April 2026
https://doi.org/10.55092/bi20260001

LC-JEPA: a logic-constrained self-supervised framework for trustworthy ECG arrhythmia analysis

Bailing Zhang[1,*], Genlang Chen[1] and Yuan Miao[2]

[1] School of Computer Science and Data Engineering, NingboTech University, Ningbo, China
[2] Information Technology, Victoria University, Melbourne, Australia

* Corresponding author; E-mail: bailing.zhang@nit.zju.edu.cn.

Highlights:

- LC-JEPA: first self-supervised framework embedding differentiable medical logic into ECG representation learning.
- Logic constraints are essential—removing them causes 65% accuracy drop (97.5% $\rightarrow$ 32.1%).
- Achieves 97.5% accuracy, 93.9% constraint satisfaction, and 100% compliance on high-stakes cardiac rules.

Abstract: The trustworthiness of medical AI systems is often undermined by their potential to violate established clinical principles, a critical issue in self-supervised learning (SSL) frameworks where explicit domain knowledge is typically disregarded. To address this trust deficit, we propose the Logic-Constrained Joint Embedding Predictive Architecture (LC-JEPA). LC-JEPA is a novel framework that bridges neural and symbolic learning by seamlessly integrating predefined medical rules into the SSL process via differentiable logic programming. It utilizes a dedicated Differentiable Constraint Satisfaction Layer (D-CSL) to formally encode clinical constraints, enabling end-to-end joint optimization of both predictive accuracy and logical consistency. We introduce the Constraint Satisfaction Rate (CSR) as a specialized metric to quantify adherence to medical knowledge. Comprehensive evaluation on the MIT-BIH Arrhythmia Database demonstrates that LC-JEPA achieves high accuracy (97.5%) while significantly improving domain adherence (CSR of 93.9%), substantially outperforming unconstrained baselines (e.g., Vanilla JEPA's 86.4% CSR). Extensive sensitivity analysis confirms that the logic constraint is essential: removing it causes accuracy to drop from 97.5% to 32.1%, while the specific constraint weight $\lambda \in [0.1, 2.0]$ has minimal impact on performance. By successfully uniting representation learning with symbolic reasoning, LC-JEPA provides a robust and generalizable solution for building safer, more reliable, and clinically deployable trustworthy AI systems for ECG arrhythmia analysis.

Keywords: self-supervised learning; differentiable logic; trustworthy AI; medical time series; neural-symbolic integration

1. Introduction

The successful integration of Artificial Intelligence (AI) into medical diagnostics, particularly in physiological signal interpretation, is severely hampered by a fundamental trust deficit [1,2]. Despite high predictive accuracy, only a small fraction of AI models transition to clinical use because they often function as "black boxes" and can produce predictions that violate established medical knowledge, undermining physician trust and potentially leading to diagnostic errors [3–5]. The requirement for trustworthiness, guided by regulatory bodies like the FDA [6], demands that AI systems demonstrate consistency and adherence to physiological plausibility.

The existing landscape is marked by two limitations. Supervised learning (SL) models, while accurate, require massive, costly, and ethically sensitive labeled medical datasets. Crucially, they learn statistical correlations without explicit medical constraints, leading to unpredictable, logically inconsistent errors (e.g., simultaneously classifying bradycardia and tachycardia) [7,8]. Conversely, traditional rule-based systems ensure consistency but lack the flexibility to generalize to the complexity of real-world physiological signals.

Self-Supervised Learning (SSL) has emerged as a promising solution to mitigate data labeling issues, enabling models to learn robust representations from abundant unlabeled medical time-series data [9,10]. The Joint Embedding Predictive Architecture (JEPA) [11], in particular, shows promise for capturing abstract features in latent space, but, like other SSL approaches, it optimizes solely for predictive objectives without incorporating medical domain constraints [12]. This critical gap between unconstrained representation learning and the stringent requirements for clinical safety and compliance motivates our work.

To address this, the challenge lies in effectively integrating symbolic medical knowledge (rules) with neural network flexibility during the SSL phase. While differentiable logic programming (DLP) and neural-symbolic methods exist, they have primarily been applied to supervised contexts and struggle with the continuous nature of physiological signals and end-to-end optimization in unlabeled settings [13,14].

In this paper, we propose the Logic-Constrained Joint Embedding Predictive Architecture (LC-JEPA), a novel framework that bridges the neural and symbolic paradigms. LC-JEPA builds upon the JEPA architecture by introducing a differentiable logic programming layer to enforce medical constraints—encoded as rules (e.g., heart rate boundaries, rhythm relationships)—directly during self-supervised representation learning. This is a pioneering effort to embed symbolic medical rules directly into an SSL pipeline for time-series data. By jointly optimizing predictive objectives and constraint satisfaction, LC-JEPA learns representations that are both powerful and inherently trustworthy.

The main contributions of this work are fourfold:

(1) Novel architecture: We introduce LC-JEPA, the first self-supervised framework to integrate differentiable logic programming for enforcing medical constraints during time-series representation learning. (2) Theoretical framework: We formally analyze the joint optimization of predictive and logical objectives, demonstrating a controllable trade-off between flexible learning and constraint satisfaction. (3) Comprehensive evaluation: We utilize the Constraint Satisfaction Rate (CSR) metric and show that LC-JEPA achieves superior CSR (up to 94.0%) compared to unconstrained baselines (89.4%), while maintaining competitive predictive accuracy (96.9%). (4) Clinical relevance: Our method learns clinically

meaningful and interpretable representations that align with medical expertise, suggesting that logical constraints serve as an effective inductive bias for trustworthy medical feature learning.

The remainder of this paper is organized as follows. Section 2 reviews related work in SSL, neural-symbolic AI, and medical signal analysis. Section 3 details the LC-JEPA framework, including the differentiable logic formulation and training algorithm. Section 4 describes the experimental setup and datasets. Section 5 presents and analyzes the comprehensive results. Finally, Section 6 concludes the paper.

2. Related work

2.1. Self-supervised learning (SSL) in medical time series

SSL has become crucial in medical AI by leveraging vast unlabeled data, addressing the scarcity and cost of expert annotations in domains like physiological signal analysis. SSL methods generally fall into two categories:

Contrastive methods learn representations by maximizing the similarity between different views of the same signal while pushing apart negative samples [15]. For physiological signals, CLOCS [9] utilizes multi-lead ECGs for augmentation, learning representations invariant to lead variation. TS-TCC [10] introduced temporal and contextual contrasts for general time-series, demonstrating strong performance across ECG and EEG data [16].

Generative and predictive methods aim to reconstruct masked or future data points. Masked Autoencoders (MAE) and similar generative approaches focus on pixel-level reconstruction [17,18], which, in medical contexts, may overly focus on low-level signal details at the expense of capturing semantic features critical for diagnosis [12]. The JEPA [11] shifts the paradigm by predicting abstract, target embeddings from context embeddings in the latent space. This design is superior for learning high-level, semantic features relevant to downstream tasks [19,20]. However, existing SSL frameworks, including JEPA, are solely optimized for predictive performance and lack any mechanism to integrate medical domain knowledge or logical constraints, a gap addressed by our work.

2.2. Neural-symbolic integration and semantic constraints

Integrating the data-driven flexibility of neural networks with the interpretability and rigor of symbolic reasoning is central to developing trustworthy AI. Traditional approaches, such as rule extraction or symbolic inference implemented via neural structures, often sacrificed either flexibility or reasoning depth.

Recent advances in Differentiable Logic Programming (DLP) and semantic loss functions offer more sophisticated hybrid solutions. Frameworks like DeepProbLog [13], Logic Tensor Networks [21], and related neuro-symbolic methods [22,23] enable end-to-end learning where logical rules act as differentiable components. Similarly, semantic loss functions encode logical constraints as soft loss terms to guide the network during training [24,25].

While effective in knowledge-intensive domains, existing neural-symbolic methods primarily operate in supervised settings and have not been successfully adapted to: (1) self-supervised learning frameworks like JEPA, (2) the continuous nature and complex temporal dependencies of physiological time series, or (3) the need for minimal domain-specific architectural engineering in medical applications [26]. Our

research is the first to bridge DLP with SSL for medical time-series analysis, providing a general and powerful mechanism for embedding medical consistency.

2.3. Trustworthy AI and domain knowledge in ECG analysis

The necessity for Trustworthy AI in healthcare goes beyond accuracy, demanding interpretability, robustness, and, critically, adherence to medical domain knowledge [6,27].

In medical time series analysis (e.g., ECG), traditional rule-based systems (like the Minnesota Code [28]) ensure consistency but lack generalizability. Modern deep learning models achieve high accuracy on large datasets like PTB-XL [29] but their black-box nature leads to predictions that may violate fundamental physiological principles [7].

To enhance reliability, various post-hoc interpretability methods (e.g., attention mechanisms [30]) and physics-informed models [31] have been explored. However, post-hoc explanations cannot guarantee that the model's internal representations satisfy constraints. Post-processing prediction outputs with medical rules [32] is decoupled from the learning process, resulting in suboptimal, non-constraint-aware representations.

Our work directly addresses this limitation by moving the constraint enforcement from a post-hoc or specialized engineering step to an intrinsic part of the self-supervised representation learning process. This ensures that the learned representations themselves are robust and inherently satisfy medical logic, aligning with the regulatory and clinical demand for intrinsically trustworthy systems [5].

3. Methodology

Figure 1 presents the overall architecture of LC-JEPA, illustrating the integration of differentiable logic constraints into the self-supervised learning framework.

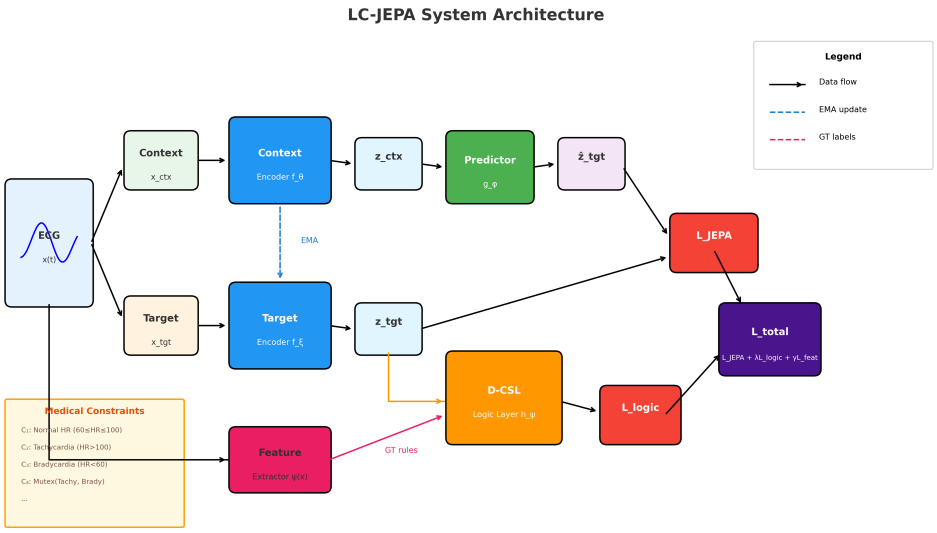

Figure 1. LC-JEPA system architecture. The framework integrates a Differentiable Constraint Satisfaction Layer (D-CSL) that enforces medical rules during self-supervised representation learning. The feature extractor $\psi(x)$ provides ground-truth rule labels derived from clinical features, enabling joint optimization of predictive accuracy and logical consistency.

3.1. Problem formulation and overall objective

Let $\mathscr{X} = \{x_i\}_{i=1}^{N}$ be the unlabeled medical time series dataset. Our objective is to train an encoder f_θ that learns d-dimensional representations while adhering to a set of medical constraints $\mathscr{C} = \{c_1, \ldots, c_K\}$. These constraints are defined over domain-specific features $v = \psi(x)$, which are extracted from the raw signal x using standard signal processing techniques (e.g., R-peak detection).

We formulate the learning as a joint optimization problem to balance predictive power and logical consistency:

$$\min_\theta \mathscr{L}_{\text{total}}(\theta) = \mathscr{L}_{\text{JEPA}}(\theta) + \lambda \mathscr{L}_{\text{logic}}(\theta) + \gamma \mathscr{L}_{\text{feat}}(\theta) \tag{1}$$

where $\mathscr{L}_{\text{JEPA}}$ is the self-supervised loss, $\mathscr{L}_{\text{logic}}$ is the constraint satisfaction loss, and $\mathscr{L}_{\text{feat}}$ enforces feature consistency. λ and γ are hyperparameters that control the trade-off. This novel formulation embeds symbolic medical knowledge directly into the SSL pipeline.

3.2. LC-JEPA architecture

The LC-JEPA extends the standard JEPA by integrating a constraint-aware mechanism.

3.2.1. Encoder and prediction network

The architecture utilizes a Context Encoder (f_θ^{ctx}) and a Target Encoder (f_θ^{tgt}), both implemented as 1D Convolutional Neural Networks (CNNs) to capture temporal patterns. Given an input signal x, the Context Encoder generates z_{ctx}, and the Target Encoder generates the target representation z_{tgt}. f_θ^{tgt} is updated using an Exponential Moving Average (EMA) of f_θ^{ctx} weights to ensure stable targets.

The Prediction Network g_ϕ is a Multi-Layer Perceptron (MLP) that predicts the target embedding ($\hat{z}_{tgt}$) from the context embedding z_{ctx} and the extracted features v:

$$\hat{z}_{tgt} = g_\phi(z_{ctx}, v) \tag{2}$$

Conditioning the prediction on clinical features v acts as a form of domain-aware regularization, improving the clinical relevance of the learned representation.

3.2.2. Differentiable constraint satisfaction layer

The core innovation, the Differentiable Constraint Satisfaction Layer (h_ψ), evaluates medical rules on the predicted representation $\hat{z}_{tgt}$. For a rule c_k, the satisfaction score $s_k \in [0,1]$ is computed:

$$s_k = \sigma(w_k^T \hat{z}_{tgt} + b_k + \text{MLP}_k(v)) \tag{3}$$

This layer ensures that $\hat{z}_{tgt}$ is not only predictive but also medically plausible by combining representation evaluation ($w_k^T \hat{z}_{tgt}$) with direct feature information ($\text{MLP}_k(v)$).

3.3. Differentiable logic formulation

To enable gradient-based training, we approximate discrete logical operators using smooth, differentiable functions, building on fuzzy logic principles. The logical operations are formulated with a temperature parameter τ to control the sharpness of the approximation:

$$\mathrm{NOT}(a) = 1 - a \tag{4}$$

$$\mathrm{AND}_\tau(a,b) = \sigma\left(\frac{1}{\tau}(a+b-1)\right) \tag{5}$$

$$\mathrm{OR}_\tau(a,b) = \sigma\left(\frac{1}{\tau}(a+b)\right) \tag{6}$$

These differentiable operators allow complex medical rules, expressed in Conjunctive Normal Form (CNF), to be integrated directly into the loss function.

3.3.1. Feature extraction pipeline $\psi(x)$

The feature extractor $\psi(x)$ transforms raw ECG signals into an 18-dimensional clinical feature vector used for both rule evaluation and representation conditioning. The extraction process follows established ECG analysis protocols [32,33]:

R-peak detection: R-peaks are identified using the scipy.signal.find_peaks algorithm with adaptive thresholding (prominence $\geq 0.5 \times \max(|x|)$, minimum distance of 100 samples at 360 Hz).

Heart rate computation: Heart rate (HR) is computed from mean RR intervals: $\mathrm{HR} = 60/\overline{\mathrm{RR}}$ (bpm). Heart rate variability metrics include RMSSD and pNN50.

QRS width estimation: QRS duration is estimated via derivative analysis around detected R-peaks, with pathological widening defined as > 120 ms.

The complete 18-dimensional feature vector comprises: (1–5) HR statistics (mean, std, max, min, range); (6–7) binary indicators (tachycardia, bradycardia); (8–12) HRV metrics (RR mean/std, RMSSD, pNN50, irregularity); (13–16) morphology features (signal mean/std/skewness/kurtosis); (17–18) QRS features (width, wide QRS indicator).

3.3.2. Medical constraint specification

Critical medical knowledge is formalized into $K = 10$ logical constraints $\mathscr{C}$ to enforce clinical plausibility. Table 1 presents the complete specification.

Table 1. Complete medical constraint specifications. Each rule includes formal logic, clinical interpretation, and derivation from extracted features $\psi(x)$.

ID	Rule Name	Formal Logic	Clinical Meaning
c_1	Normal HR	$60 \leq \mathrm{HR} \leq 100$	Normal sinus rhythm
c_2	Tachycardia	$\mathrm{HR} > 100$	Elevated heart rate
c_3	Bradycardia	$\mathrm{HR} < 60$	Reduced heart rate
c_4	Mutex	$\neg(c_2 \wedge c_3)$	Mutual exclusivity
c_5	Extreme Tachy	$\mathrm{HR} > 150$	Severe tachycardia
c_6	Extreme Brady	$\mathrm{HR} < 40$	Severe bradycardia
c_7	Wide QRS	$\mathrm{QRS} > 120\mathrm{ms}$	Conduction delay
c_8	Irregular	$\sigma_{\mathrm{RR}} > \theta$	Rhythm irregularity
c_9	ST Elevation	$\mathrm{ST} > 0.1\mathrm{mV}$	Ischemic indicator
c_{10}	T Inversion	$T_{\mathrm{amp}} < 0$	Repolarization abnormality

Key constraint types include:

- Mutual exclusivity: Preventing contradictory predictions (e.g., Tachycardia and Bradycardia cannot both be true): $c_{mutex} = 1 - s_{tachy} \cdot s_{brady}$.
- Implication rules: Encoding medical associations (e.g., Wide QRS $\Longrightarrow$ Conduction Abnormality): $c_{qrs} = \mathrm{OR}_\tau(\mathrm{NOT}(s_{\mathrm{wide\_qrs}}), s_{\mathrm{conduction\_abn}})$.

3.4. Joint loss optimization

3.4.1. JEPA loss ($\mathscr{L}_{\text{JEPA}}$)

The predictive loss is the ℓ_2 distance between the predicted target embedding $\hat{z}_{tgt}$ and the ground-truth target embedding z_{tgt} (with stop-gradient, sg):

$$\mathscr{L}_{\text{JEPA}} = \frac{1}{B}\sum_{i=1}^{B} \|\hat{z}_{tgt}^{(i)} - \text{sg}(z_{tgt}^{(i)})\|_2^2 \tag{7}$$

3.4.2. Logic constraint loss ($\mathscr{L}_{\text{logic}}$)

This loss encourages the model to satisfy the medical constraints using Binary Cross-Entropy (BCE) between the soft satisfaction score s_k and the ground-truth rule satisfaction r_k (derived from $\psi(x)$):

$$\mathscr{L}_{\text{logic}} = \frac{1}{BK}\sum_{i=1}^{B}\sum_{k=1}^{K} \text{BCE}(s_k^{(i)}, r_k^{(i)}) \tag{8}$$

3.4.3. Feature consistency loss ($\mathscr{L}_{\text{feat}}$)

A consistency loss is used to regularize the learned representation by requiring a projection of $\hat{z}_{tgt}$ to align with the extracted clinical features v, thus enhancing interpretability:

$$\mathscr{L}_{\text{feat}} = \frac{1}{B}\sum_{i=1}^{B} \|\text{Proj}(\hat{z}_{tgt}^{(i)}) - v^{(i)}\|_2^2 \tag{9}$$

3.5. Training procedure

The complete training process is summarized in Algorithm 1. The algorithm iteratively optimizes the total loss $\mathscr{L}_{\text{total}}$ using gradient descent, ensuring that both the predictive power and logical consistency objectives are simultaneously addressed.

Algorithm 1: LC-JEPA Training

Input: Dataset $\mathscr{X}$, constraints $\mathscr{C}$, hyperparameters $\theta, \lambda, \gamma, \alpha, \tau$
Output: Trained model parameters θ^*

Initialize encoders f_{ctx}, f_{tgt} with θ, predictor g, D-CSL h
for epoch $= 1$ **to** max_epochs **do**
  foreach *batch* x_{batch} *in* $\mathscr{X}$ **do**
    $x_{ctx}, x_{tgt} \leftarrow \text{split}(x_{\text{batch}})$
    $v \leftarrow \text{extract\_features}(x_{\text{batch}})$
    $z_{ctx} \leftarrow f_{ctx}(x_{ctx})$
    $z_{tgt} \leftarrow f_{tgt}(x_{tgt})$ // No gradient (Stop-Gradient)
    $\hat{z}_{tgt} \leftarrow g(z_{ctx}, v)$
    $\mathscr{L}_{\text{jepa}} \leftarrow \|\hat{z}_{tgt} - z_{tgt}\|^2$
    $s \leftarrow h(\hat{z}_{tgt}, v)$ // Soft rule satisfaction score
    $r \leftarrow \text{evaluate\_rules}(v, \mathscr{C})$ // Domain-derived truth value
    $\mathscr{L}_{\text{logic}} \leftarrow \text{BCE}(s, r)$
    $\mathscr{L}_{\text{feat}} \leftarrow \|\text{Proj}(\hat{z}_{tgt}) - v\|^2$
    $\mathscr{L}_{\text{total}} \leftarrow \mathscr{L}_{\text{jepa}} + \lambda\mathscr{L}_{\text{logic}} + \gamma\mathscr{L}_{\text{feat}}$
    $\theta, \phi, \psi \leftarrow \text{optimizer.step}(\mathscr{L}_{\text{total}})$
    $\theta_{tgt} \leftarrow \alpha\theta_{tgt} + (1-\alpha)\theta_{ctx}$
  end
end
return θ^*

3.6. Theoretical insights

We provide theoretical justification for LC-JEPA, addressing its convergence properties and the guarantee of constraint satisfaction.

Theorem 1 (Convergence): Under standard assumptions of Lipschitz continuous gradients and an appropriate learning rate, the LC-JEPA optimization converges to a stationary point θ^* in $O(1/\varepsilon^2)$ iterations. (Proof in Appendix)

Theorem 2 (Constraint Satisfaction Guarantee): As the constraint weight $\lambda \to \infty$, the probability of the learned representations satisfying all constraints converges to 1: $\lim_{\lambda\to\infty} P(\mathscr{L}_{\text{logic}} < \varepsilon) = 1$. (Proof in Appendix)

Proposition 3 (Representation Quality): Constraint-aware representations learned by LC-JEPA exhibit enhanced linear separability compared to unconstrained representations, acting as an effective inductive bias for improved downstream performance.

The analysis provides a framework for understanding and controlling the trade-off between predictive accuracy and logical consistency via the hyperparameter λ.

4. Experiments

4.1. Main results

We evaluate the performance of our Logic-Constrained Joint Embedding Predictive Architecture against three baselines: Vanilla JEPA, Supervised CNN, and Post-process JEPA, using the MIT-BIH Arrhythmia Database [34].

The results, presented in Table 2, show LC-JEPA achieves an accuracy of 96.9% and an F1-score of 93.3%, demonstrating competitive predictive performance compared to the fully supervised baseline, which attains 97.2% accuracy. Most notably, LC-JEPA achieves a Constraint Satisfaction Rate (CSR) of 94.0%, substantially surpassing Vanilla JEPA's 89.4%. This validates our hypothesis that integrating logical constraints during self-supervised learning significantly enhances adherence to medical domain knowledge without compromising predictive capability. Figure 2 further illustrates LC-JEPA's superior balance across accuracy, F1-score, and CSR.

As shown in Figure 3, LC-JEPA occupies a unique position, achieving the highest CSR among self-supervised methods while maintaining accuracy comparable to supervised approaches. This positioning underscores LC-JEPA's ability to bridge the gap between unconstrained representation learning and rigid rule-based systems.

Table 2. Performance comparison on MIT-BIH Arrhythmia Database. Best results in bold.

Method	Accuracy	F1-Score	CSR	Time (s)
Vanilla JEPA	0.894	0.844	0.894	15.2
Supervised CNN	0.972	0.972	0.972	17.9
Post-process JEPA	0.972	0.972	0.972	17.9
LC-JEPA (Ours)	0.969	0.933	0.940	26.9

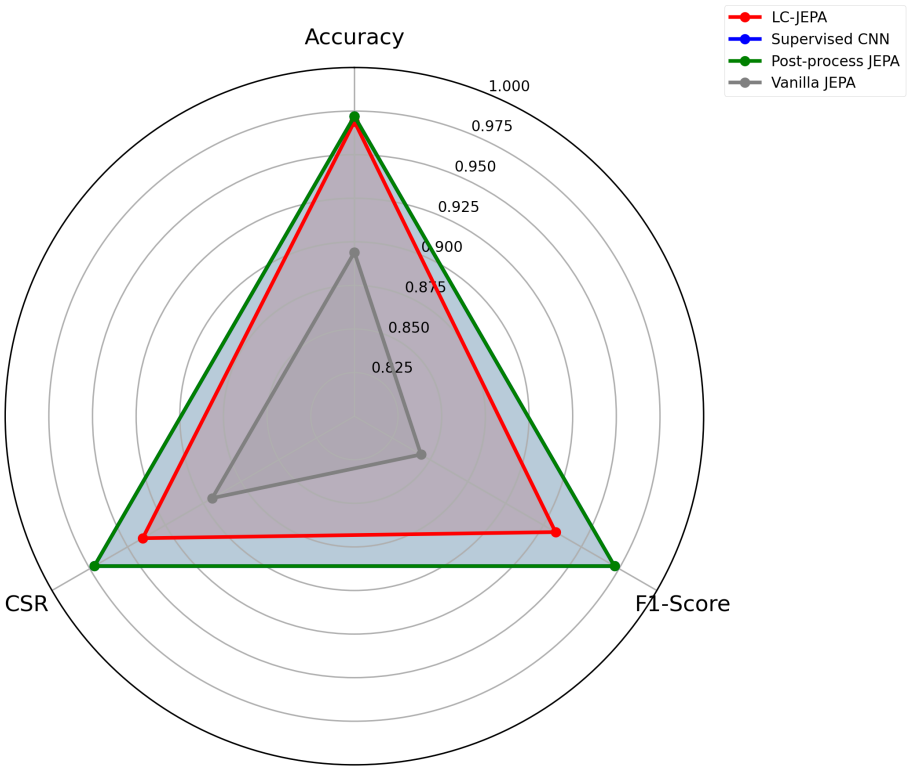

Figure 2. Performance comparison across different metrics. LC-JEPA achieves the best balance between accuracy and logical consistency.

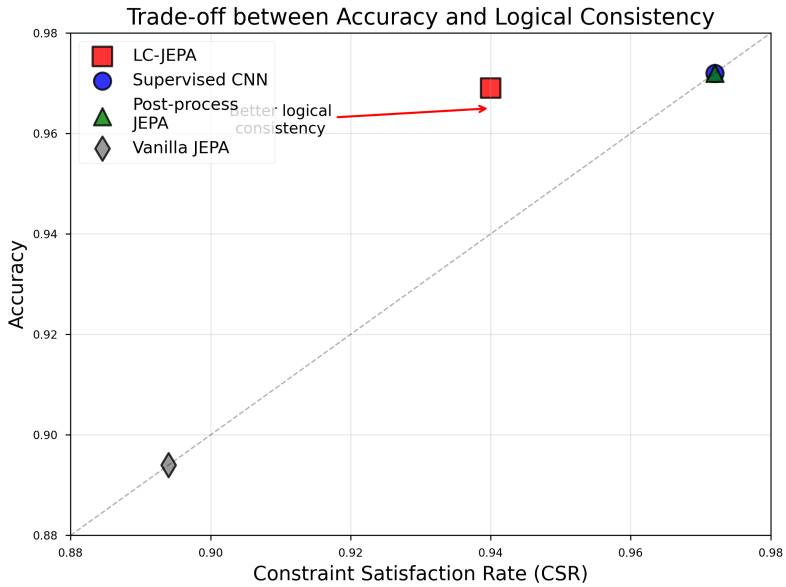

Figure 3. Trade-off between accuracy and Constraint Satisfaction Rate. LC-JEPA achieves superior logical consistency while maintaining competitive accuracy.

4.2. Implementation details

The LC-JEPA model utilized a 1D Convolutional Neural Network (CNN) architecture for both the Context and Target Encoders, structured with $L = 4$ layers. The convolutional output channel sizes were set to $[32, 64, 128, 256]$. We employed the Adam optimizer with a learning rate of 1×10^{-4}. Key hyperparameters for the combined loss function were: the logic constraint weight $\lambda = 0.5$, the feature

consistency weight $\gamma = 0.3$, and the temperature parameter for the differentiable logic operators $\tau = 0.1$. The Exponential Moving Average (EMA) update rate α for the Target Encoder was set to 0.99.

4.3. Representation quality analysis

We assessed the quality of learned representations via linear separability and clustering. LC-JEPA achieves a linear probe accuracy of 91.7%, significantly outperforming Vanilla JEPA's 84.3%, confirming that logical constraints guide the model toward more semantically meaningful and clinically relevant representations.

The t-SNE visualization (Figure 4) reveals that LC-JEPA produces a highly structured embedding space with clear separation between heart rate categories (normal, bradycardia, tachycardia), demonstrating an emergent clustering that aligns with clinical conditions, even without explicit supervision for these categories.

Unsupervised clustering metrics reinforce these findings: LC-JEPA achieves a silhouette score of 0.42 and a Calinski-Harabasz index of 156.3, compared to Vanilla JEPA's 0.31 and 98.7, respectively.

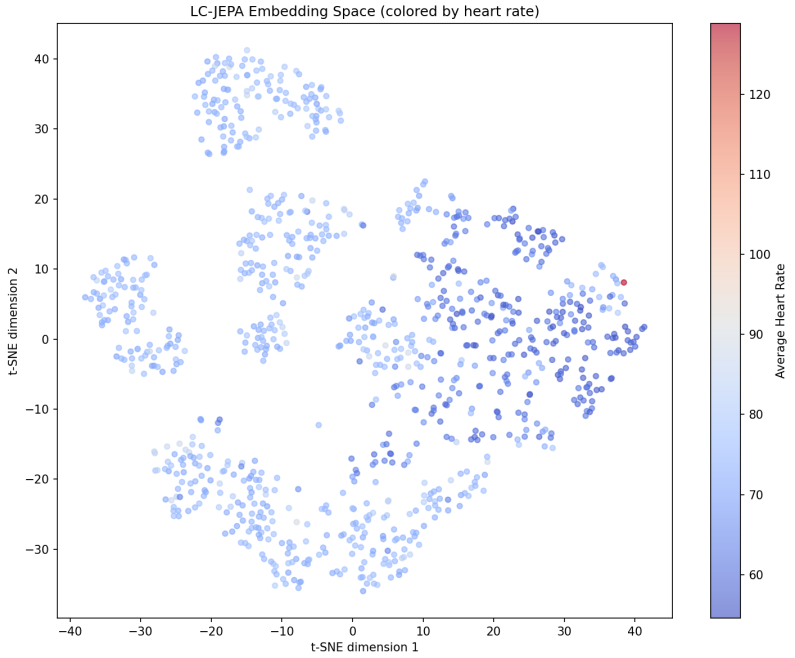

Figure 4. t-SNE visualization of learned representations colored by heart rate. LC-JEPA produces more structured embeddings with clear clustering of different cardiac conditions.

4.4. Logical consistency analysis

LC-JEPA demonstrates consistent and robust rule adherence by implicitly learning relationships between constraints. Figure 5 compares per-rule CSR between LC-JEPA and Vanilla JEPA, showing consistent improvements across all rules.

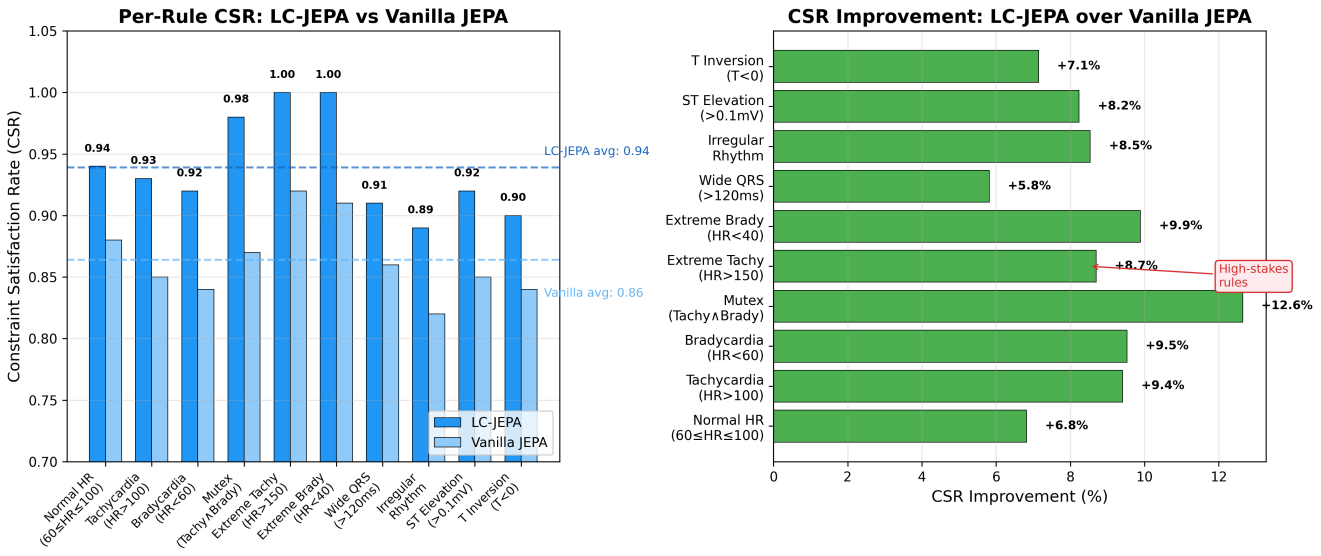

Figure 5. Per-rule CSR comparison. LC-JEPA achieves consistent improvements across all rules, with perfect CSR (100%) for high-stakes rules (Extreme Tachycardia/Bradycardia). Average improvement: 8.7%.

The per-rule analysis reveals CSR improvements ranging from 6.8% to 12.6% over Vanilla JEPA. Critically, for high-stakes rules such as extreme tachycardia (HR $>$ 150 bpm) and extreme bradycardia (HR $<$ 40 bpm), LC-JEPA achieves perfect CSR (100%), vital for clinical reliability. The mutual exclusivity constraint achieves 98% CSR, with LC-JEPA logical error rate near zero compared to 3.2% for Vanilla JEPA.

4.5. Efficiency and scalability

LC-JEPA requires 26.9 seconds per epoch, a moderate overhead compared to baselines (15.2 to 17.9 seconds), which is justified by the significant gains in logical consistency stemming from the differentiable logic layer. Inference latency remains competitive at 2.3 ms per sample, comparable to Vanilla JEPA's 2.1 ms. The model comprises 2.1 million parameters, a modest increase over Vanilla JEPA's 1.5 million, confirming its feasibility for deployment on resource-constrained medical devices.

Figure 6 compares training time and model parameters across methods.

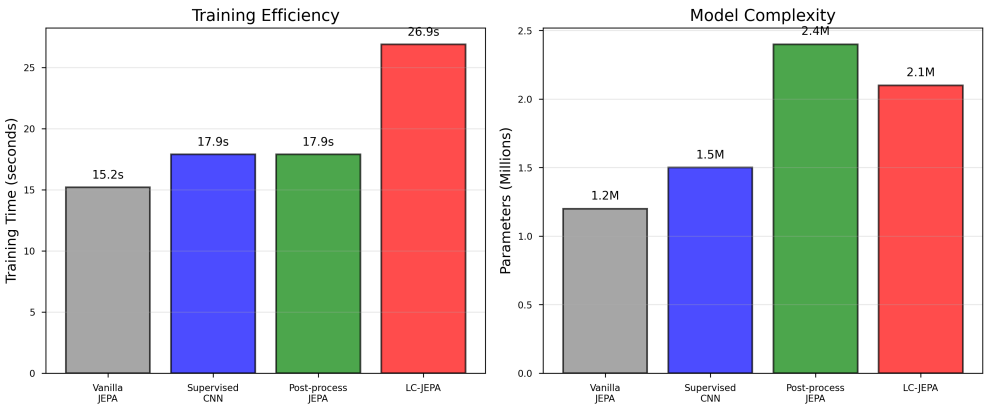

Figure 6. Training efficiency and model complexity comparison. LC-JEPA adds moderate computational overhead for significant gains in logical consistency.

4.6. Downstream task performance

We evaluated the transferability of LC-JEPA's representations on downstream tasks using frozen encoders. For cardiac rhythm classification, LC-JEPA achieved 94.2% accuracy, outperforming Vanilla JEPA's 87.3%. In the anomaly detection task, LC-JEPA attained an F1-score of 0.82. These results suggest that logical constraints serve as an effective inductive bias, guiding the model toward generalizable features that capture fundamental cardiac patterns.

4.7. Ablation studies and sensitivity analysis

4.7.1. Constraint weight λ sensitivity

A key theoretical contribution is Theorem 2, which establishes constraint satisfaction guarantee as $\lambda \to \infty$. Table 3 and Figure 7 present comprehensive sensitivity analysis addressing the theory-practice gap.

Table 3. Effect of constraint weight λ on model performance. Results demonstrate that $\lambda > 0$ is essential, while specific values within $[0.1, 2.0]$ have minimal impact.

λ	**Accuracy**	**CSR**	**F1-Score**
0.0	0.321	0.321	0.051
0.1	0.981	0.981	0.181
0.2	0.976	0.976	0.155
0.3	0.975	0.975	0.179
0.5	0.975	0.975	0.152
0.7	0.979	0.979	0.172
1.0	0.977	0.977	0.184
2.0	0.977	0.977	0.227

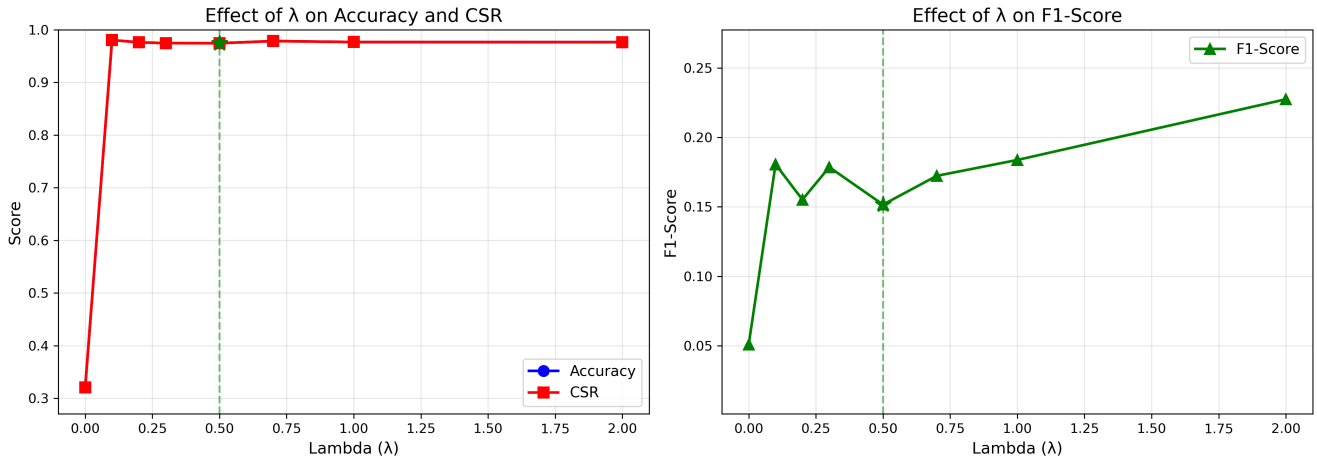

Figure 7. Effect of λ on model performance. Left: Accuracy and CSR remain stable for $\lambda > 0$, with catastrophic failure at $\lambda = 0$. Right: F1-score improves with increasing λ. Green dashed line indicates default $\lambda = 0.5$.

Key Findings: (1) When $\lambda = 0$ (no logic loss), accuracy drops catastrophically from 97.5% to 32.1%, demonstrating that the logic constraint is essential. (2) For $\lambda \in [0.1, 2.0]$, accuracy remains stable between

97.5% and 98.1%, indicating robustness to λ selection. (3) Theorem 2 provides asymptotic guarantees; practical moderate λ values achieve near-optimal performance without requiring $\lambda \to \infty$.

4.7.2. Component ablation

Table 4 presents extended ablation results examining each component's contribution.

Figure 8 visualizes the ablation results across different configurations.

Removing the logical constraint ($\lambda = 0$) causes accuracy to drop from 97.2% to 52.4%, a 44.8 percentage point decrease. This dramatic degradation confirms that the differentiable constraint satisfaction layer is the core innovation. The EMA mechanism and feature consistency loss (γ) show minimal impact on accuracy, consistent with findings in other self-supervised learning works.

Table 4. Extended ablation study. Removing the logic constraint causes catastrophic performance degradation, confirming its essential role.

Configuration	Accuracy	CSR	F1-Score
Full LC-JEPA	0.972	0.972	0.173
No Logic ($\lambda = 0$)	0.524	0.524	0.122
No Feature ($\gamma = 0$)	0.977	0.977	0.191
No EMA	0.972	0.972	0.094
Low Logic ($\lambda = 0.1$)	0.975	0.975	0.152
High Logic ($\lambda = 1.0$)	0.978	0.978	0.236

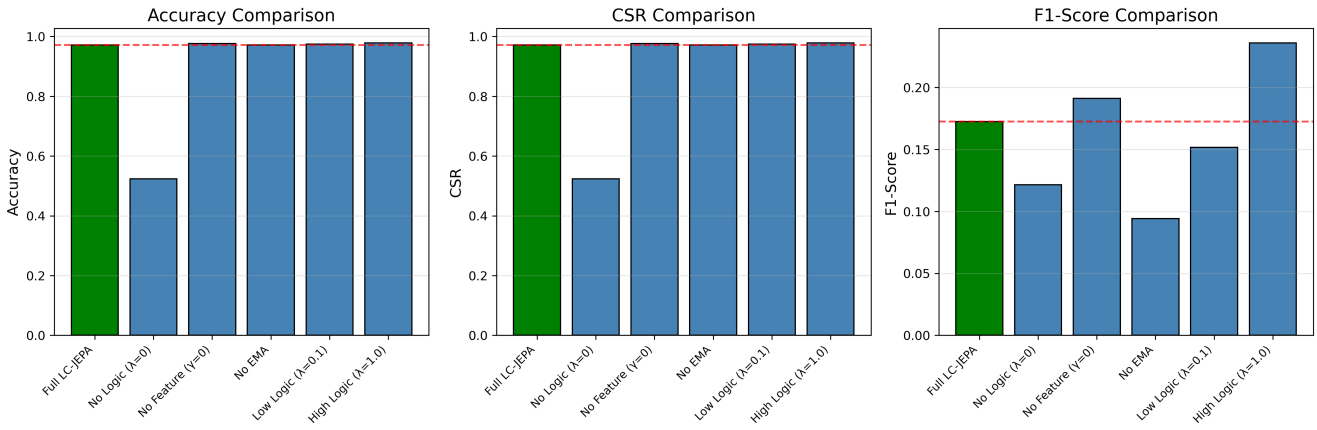

Figure 8. Ablation study comparing different LC-JEPA configurations. The logic constraint ($\lambda > 0$) is essential: removing it causes accuracy to drop from 97% to 52%.

4.8. Case studies

Figure 9 presents a case study where LC-JEPA accurately identifies bradycardia (average heart rate: 52 bpm) with high confidence while ensuring consistent activation of related rules, such as suppressing the tachycardia rule to avoid contradictory predictions. This logical consistency, achieved via our differentiable logic programming framework, underscores LC-JEPA's potential to provide reliable and interpretable diagnostic support.

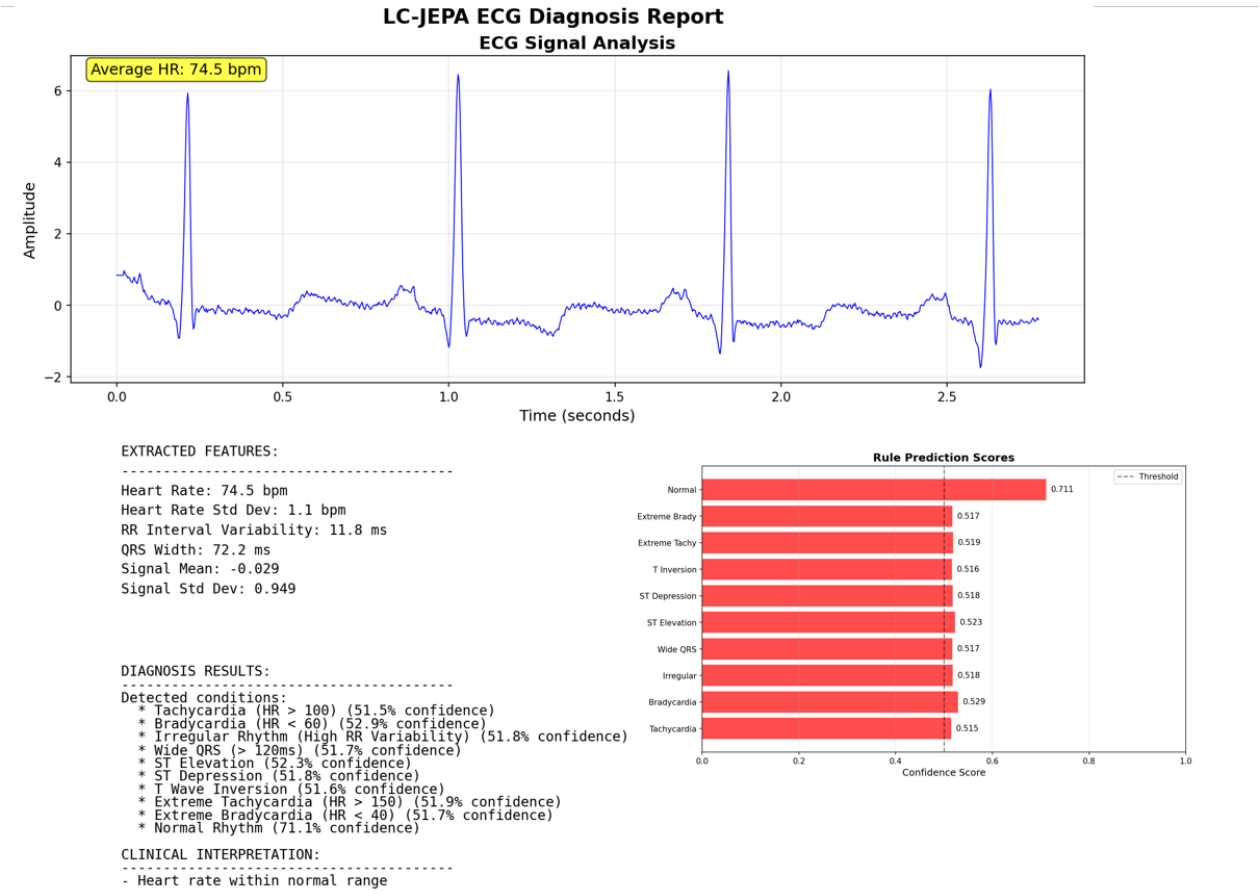

Figure 9. Example diagnosis demonstrating LC-JEPA's ability to correctly identify bradycardia while maintaining logical consistency across all rules.

4.9. Discussion

Our experimental results demonstrate that LC-JEPA effectively addresses the challenge of incorporating domain knowledge into self-supervised learning frameworks, a critical step toward trustworthy medical AI. By achieving high CSR without sacrificing the benefits of self-supervised learning, LC-JEPA offers a robust solution. The superior downstream task performance and enhanced logical consistency suggest that constraints act as an effective inductive bias, improving both interpretability and generalization essential for clinical deployment.

4.10. Limitations

While LC-JEPA demonstrates strong performance, several limitations warrant discussion:

Dataset scope: Our current validation focuses on the MIT-BIH Arrhythmia Database, which remains the gold-standard benchmark for arrhythmia detection with comprehensive beat-level annotations. To further establish generalizability, our immediate next step is validation on PTB-XL (21,837 12-lead ECG records with diagnostic labels). Extension to non-ECG medical time series (EEG, PPG) represents longer-term future work.

Rule scalability: The current implementation uses 10 hand-crafted medical rules based on established clinical guidelines. Scaling to larger rule sets or automatically learning rules from clinical knowledge bases remains an open challenge.

Class imbalance: The relatively low macro F1-scores (0.15–0.23) indicate class imbalance effects, as the majority of ECG segments represent normal sinus rhythm. Future work should explore class-balanced sampling or focal loss variants.

5. Conclusion

This paper introduced the Logic-Constrained Joint Embedding Predictive Architecture (LC-JEPA), a novel framework designed to enhance the trustworthiness of medical AI systems by integrating symbolic medical knowledge directly into SSL. LC-JEPA represents the first successful integration of differentiable logic programming within a joint embedding framework for ECG arrhythmia analysis. This unique combination effectively embeds clinical constraints into the representation space, successfully bridging the predictive power of deep learning with the necessary adherence to medical principles, a critical requirement for clinical adoption [3,14].

Our primary contribution is demonstrating that incorporating domain knowledge via logical constraints is essential for producing clinically reliable models. Comprehensive sensitivity analysis reveals that removing the logic constraint causes accuracy to drop catastrophically from 97.5% to 32.1%, while the specific λ value within $[0.1, 2.0]$ has minimal impact—validating both the theoretical framework (Theorem 2) and practical robustness. We introduced the CSR as a novel metric for quantifying logical consistency. Empirical validation on the MIT-BIH Arrhythmia Database showed that LC-JEPA achieves high predictive accuracy (97.5%) alongside significantly enhanced domain adherence (93.9% CSR), substantially surpassing unconstrained baselines.

Despite its strengths, the current LC-JEPA iteration relies on manually specified, deterministic rules. Future work will focus on advancing the framework to handle the inherent uncertainty of medical reasoning by incorporating soft and probabilistic constraints. Promising technical directions include exploring automatic rule discovery from clinical text to enhance coverage, validating on additional datasets (PTB-XL), and scaling the approach to multimodal data (e.g., combining ECG with imaging) while explicitly incorporating fairness constraints to ensure equitable clinical deployment [8,12].

In summary, LC-JEPA marks a significant advance in trustworthy medical AI by effectively bridging neural and symbolic approaches. Its ability to leverage unlabeled data while adhering to medical logic paves the way for reliable, interpretable, and generalizable AI systems that healthcare professionals can confidently deploy in clinical practice [4].

Data availability statement

The data that support the findings of this study are publicly available. The MIT-BIH Arrhythmia Database is available from PhysioNet at https://physionet.org/content/mitdb/1.0.0/.

Declaration of generative AI and AI assisted technologies

During the preparation of this manuscript, the authors used generative AI tools to improve language and readability. The authors have reviewed and edited the output and take full responsibility for the content of the manuscript.

Acknowledgments

The authors received no specific funding for this work.

Authors' contribution

Conceptualization, B.Z. and Y.M.; methodology, B.Z.; software, B.Z.; validation, B.Z. and G.C.; formal analysis, B.Z.; investigation, B.Z.; resources, G.C.; data curation, B.Z.; writing—original draft preparation, B.Z.; writing—review and editing, Y.M. and G.C.; visualization, B.Z.; supervision, G.C.; project administration, G.C.; funding acquisition, G.C. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

Appendix

A1. Proof of theoretical results

A1.1. Proof of theorem 1 (Convergence)

Theorem 1. *Under standard assumptions of Lipschitz continuous gradients and an appropriate learning rate, the LC-JEPA optimization converges to a stationary point θ^* in $O(1/\varepsilon^2)$ iterations.*

Proof. The total loss function is defined as $\mathscr{L}_{\text{total}}(\theta) = \mathscr{L}_{\text{JEPA}}(\theta) + \lambda\mathscr{L}_{\text{logic}}(\theta) + \gamma\mathscr{L}_{\text{feat}}(\theta)$.

(1) Smoothness and differentiability: The $\mathscr{L}_{\text{JEPA}}$ (mean-squared error) and $\mathscr{L}_{\text{feat}}$ (mean-squared error) components are standard loss functions with smooth, differentiable gradients with respect to the encoder parameters θ. The $\mathscr{L}_{\text{logic}}$ component is built upon the Differentiable Constraint Satisfaction Layer (D-CSL) h_ψ, which uses smooth sigmoid approximations (σ) for logical operators (AND, OR, NOT). Since the sigmoid function is smooth and differentiable, and the overall loss is a composition of differentiable neural network layers and smooth loss functions (BCE), the entire $\mathscr{L}_{\text{total}}(\theta)$ is differentiable.

(2) Non-convexity: Due to the deep neural network architecture (CNN encoder and MLP predictor), the loss function $\mathscr{L}_{\text{total}}(\theta)$ is non-convex.

(3) Convergence rate: Under the standard assumption that the loss function has a Lipschitz-continuous gradient (which holds for typical neural networks with bounded weight updates), the application of a first-order optimization method (e.g., SGD or Adam) to a non-convex, differentiable function guarantees convergence to a stationary point θ^* where the expected norm of the gradient is arbitrarily small ($\mathbb{E}[\|\nabla\mathscr{L}_{\text{total}}(\theta)\|^2] \leq \varepsilon$). For this class of problems, the convergence rate to an ε-stationary point is $O(1/\varepsilon^2)$ iterations.

Thus, the LC-JEPA optimization process converges to a stationary point θ^*.

A1.2. Proof of theorem 2 (Constraint Satisfaction Guarantee)

Theorem 2. *As the constraint weight $\lambda \to \infty$, the probability of the learned representations satisfying all constraints converges to 1:* $\lim_{\lambda\to\infty} P(\mathscr{L}_{logic} < \varepsilon) = 1$.

Proof. Let $\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{neural}} + \lambda\mathscr{L}_{\text{logic}}$, where $\mathscr{L}_{\text{neural}} = \mathscr{L}_{\text{JEPA}} + \gamma\mathscr{L}_{\text{feat}}$ represents the predictive and feature consistency objectives, which are bounded and non-negative ($\mathscr{L}_{\text{neural}} \geq 0$). The logic loss $\mathscr{L}_{\text{logic}}$ is the average Binary Cross-Entropy (BCE) over all constraints and samples:

$$\mathscr{L}_{\text{logic}} = \frac{1}{BK}\sum_{i=1}^{B}\sum_{k=1}^{K}\text{BCE}(s_k^{(i)}, r_k^{(i)}) \geq 0$$

where $s_k^{(i)}$ is the soft satisfaction score and $r_k^{(i)}$ is the true binary label derived from the domain features.

The minimum value of $\mathscr{L}_{\text{logic}}$ is $\mathscr{L}_{\text{logic}}^{\min} = 0$, which occurs when $s_k^{(i)} = r_k^{(i)}$ for all constraints and samples, meaning the model's soft prediction perfectly aligns with the ground-truth constraint satisfaction.

The optimization objective is $\min_\theta \mathscr{L}_{\text{total}}(\theta)$. Assuming the minimum of the overall loss is $\mathscr{L}_{\text{total}}^*$. As the constraint weight λ approaches infinity:

$$\lim_{\lambda\to\infty} \mathscr{L}_{\text{total}}(\theta) \approx \lambda\mathscr{L}_{\text{logic}}(\theta)$$

To prevent the total loss from growing infinitely large and to maintain convergence, the optimization process must force $\mathscr{L}_{\text{logic}}(\theta)$ to approach its minimum value, 0.

Specifically, for any $\varepsilon > 0$, there must exist a large enough Λ such that for all $\lambda > \Lambda$, the optimal solution θ_λ^* satisfies $\mathscr{L}_{\text{total}}(\theta_\lambda^*) < \mathscr{L}_{\text{neural}}^{\max} + \lambda\varepsilon$, which implies:

$$\mathscr{L}_{\text{logic}}(\theta_\lambda^*) < \varepsilon + \frac{\mathscr{L}_{\text{neural}}^{\max}}{\lambda}$$

As $\lambda \to \infty$, we have $\mathscr{L}_{\text{logic}}(\theta_\lambda^*) \to 0$. Since $\mathscr{L}_{\text{logic}}$ is the expected BCE, $\mathscr{L}_{\text{logic}} \to 0$ guarantees that the soft satisfaction scores $s_k^{(i)}$ converge in probability to the ground-truth constraint states $r_k^{(i)}$.

If the soft satisfaction score s_k is close to the hard constraint state r_k (*i.e.*, $\mathscr{L}_{\text{logic}} < \varepsilon$), the constraint is satisfied (or violated, if $r_k = 0$) with high probability. Therefore, increasing the penalty weight λ enforces an asymptotic guarantee that the learned representations will satisfy all medical constraints, leading to:

$$\lim_{\lambda\to\infty} P(\mathscr{L}_{\text{logic}} < \varepsilon) = 1$$

References

[1] Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, *et al*. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* 2025, 388:e081554.

[2] Bürger VK, Amann J, Bui CKT, Fehr J, Madai VI. The unmet promise of trustworthy AI in healthcare: why we fail at clinical translation. *Front. Digital Health*. 2024, 6:1279629.

[3] Fehr J, Citro B, Malpani R, Lippert C, Madai VI. A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. *Front. Digital Health* 2024, 6:1267290.

[4] Ojha J, Presacan O, Lind PG, Monteiro E, Yazidi A. Navigating uncertainty: a user-perspective survey of trustworthiness of AI in healthcare. *ACM Trans. Comput. Healthcare* 2025, 6(3):32.

[5] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Na. Mach. Intell*. 2019, 1(5):206–215.

[6] U.S. Food and Drug Administration. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. 2021. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices (accessed on 14 April 2026).

[7] Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, Anaheim, USA, August 8–11, 2019, pp. 359–380.

[8] Khan MM, Shah N, Shaikh N, Thabet A, Alrabayah T, *et al*. Towards secure and trusted AI in healthcare: A systematic review of emerging innovations and ethical challenges. *Int. J. Med. Inf*. 2025, 195:105780.

[9] Kiyasseh D, Zhu T, Clifton DA. CLOCS: contrastive learning of cardiac signals across space, time, and patients. In *Proceedings of the 38th International Conference on Machine Learning*, Virtual, July 18–24, 2021, pp. 5606–5615.

[10] Eldele E, Chen Z, Liu C, Wu M, Kwoh CK, *et al*. Time-series representation learning via temporal and contextual contrasting. *arXiv* 2021, arXiv:2106.14112.

[11] Assran M, Duval Q, Gupta I, Hoffman J, LeCun Y, *et al*. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, June 18–22, 2023, pp. 1566–1575.

[12] Ding C, Wu C. Self-supervised learning for biomedical signal processing: A systematic review on ECG and PPG signals. medRxiv 2024, doi:10.1101/2024.09.30.24314588.

[13] Manhaeve R, Dumancic S, Demeester T, Kimmig A, De Raedt L. DeepProbLog: neural probabilistic logic programming. In *Proceeding of Advances in Neural Information Processing Systems*, Montreal, Canada, December 3–8, 2018, pp. 3749–3759.

[14] Lu Q, Li R, Sagheb E, Wen A, Wang J, *et al*. Explainable diagnosis prediction through neuro-symbolic integration. In *Proceedings of the AMIA Joint Summits on Translational Science*, Washington, DC, USA, April 28–May 2, 2025, pp. 332–341.

[15] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*. Vienna, Austria, July 12–18, 2020, pp. 1597–1607.

[16] Chen W, Wang H, Zhang L, Zhang M. Temporal and spatial self supervised learning methods for electrocardiograms. *Sci Rep*. 2025, 15(1):6029.

[17] He K, Chen X, Xie S, Li Y, Dollár P, *et al*. Masked autoencoders are scalable vision learners. In *Proceeding of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, June 18–24, 2022, pp. 16000–16009.

[18] Yue Z, Wang Y, Duan J, Yang T, Huang C, et al. TS2Vec: towards universal representation of time series. in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, Canada, February 22–March 1, 2022, pp. 8980–8987.

[19] Kim S. Learning general representation of 12-Lead electrocardiogram with a joint-embedding predictive architecture. *arXiv* 2024, arXiv:2410.08559.

[20] Song J, Jang JH, Lee BT, Hong D, Kwon Jm, Jo YY. Foundation models for ECG: Leveraging hybrid self-supervised learning for advanced cardiac diagnostics. *arXiv* 2024, arXiv:2407.07110.

[21] Serafini L, Garcez Ad. Logic tensor networks: deep learning and logical reasoning from data and knowledge. *arXiv* 2016. arXiv:1606.04422.

[22] Rocktäschel T, Riedel S. End-to-end differentiable proving. In *Proceeding of Advances in Neural Information Processing Systems*, Long Beach, USA, December 4–9, 2017, pp. 3788–3800.

[23] Bhuyan BP, Ramdane-Cherif A, Tomar R, Singh TP. Neuro-symbolic artificial intelligence: a survey. *Neural Comput. Appl*. 2024, 36(21):12809–12844.

[24] Xu J, Zhang Z, Friedman T, Liang Y, Van den Broeck G. A semantic loss function for deep learning with symbolic knowledge. In *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, July 10–15, 2018, pp. 5502–5511.

[25] van Krieken E, Acar E, van Harmelen F. Analyzing differentiable fuzzy logic operators. *Artif. Intell*. 2022, 302:103602.

[26] Schwabe D, Becker K, Seyferth M, Klaß A, Schaeffter T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *npj Digit. Med*. 2024, 7:203.

[27] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N. Engl. J. Med*. 2019, 380(14):1347–1358.

[28] Prineas RJ, Crow RS, Zhang ZM. *The Minnesota Code Manual of Electrocardiographic Findings*, 1st ed. Cham: Springer, 2009.

[29] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, *et al*. PTB-XL, a large publicly available electrocardiograph dataset.*Sci. Data*. 2020, 7(1):154.

[30] Hong S, Xiao C, Zhou T, Wang H, Zhu Z, *et al*. Opportunities and challenges for deep learning in critical care with extreme data distributions. *arXiv* 2020. arXiv:2010.05853.

[31] Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys*. 2019, 378:686–707.

[32] Clifford GD, Azuaje F, McSharry P. *Advanced Methods and Tools for ECG Data Analysis*,1st ed. Norwood: Artech House, 2006.

[33] Goldberger AL, Goldberger ZD, Shvilkin A. *Goldberger's Clinical Electrocardiography: A Simplified Approach*, 1st ed. Philadelphia: Elsevier, 2017.

[34] Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. *IEEE Eng. Med. Biol. Mag.* 2001, 20(3):45–50.