# World Model Augmentation for Imbalanced Multi-Label ECG Classification

**Luiz Facury** [1] **Jose Geraldo Fernandes** [1] **Pedro Robles** [1] **Wagner Meira Jr.** [1]

## Abstract

Automated multi-label ECG classification struggles with severe class imbalance: rare co-occurring cardiac conditions are systematically underrepresented in clinical datasets. We address this with an action-conditioned latent transition model, pretraining a Joint-Embedding Predictive Architecture (LeJEPA) on over 700K longitudinal ECG pairs from MIMIC-IV-ECG, training it to predict how a patient's latent ECG representation changes between visits given the shift in their ICD label set. After pretraining, we repurpose the frozen dynamics model as a data augmentor — given a normal embedding and a target condition combination, it synthesises the corresponding abnormal embedding entirely in representation space, without generating a single additional waveform. Training a lightweight MLP probe on the resulting 2.7M-embedding dataset (721K real + 2M synthetic) achieves a macro-averaged AUROC of 0.743 across 76 ICD-coded conditions, recovering 55% of the gap between a real-data-only linear probe (0.687) and a fully fine-tuned encoder (0.789), with no encoder updates.

## 1. Introduction

Cardiac disease is the leading cause of death globally, making electrocardiography one of the most widely deployed diagnostic tools in medicine, with billions of recordings acquired annually. The availability of large annotated datasets such as MIMIC-IV-ECG (Gow et al., 2023) and PTB-XL (Wagner et al., 2020) has fuelled rapid progress in automated multi-label ECG classification (Hannun et al., 2019; Ribeiro et al., 2020), with deep networks now matching or exceeding cardiologist-level performance on common conditions. However, clinical ECG data is inherently long-tailed: while benign findings and frequent arrhythmias are abundantly represented, rare or co-occurring pathologies appear only a handful of times even in datasets of hundreds of thousands of recordings. This imbalance directly limits the performance of standard supervised models on the conditions that matter most clinically.

A natural approach to address unbalanced data is data augmentation, but existing approaches have a fundamental problem. Signal-space methods, such as GANs (Delaney et al., 2019) and diffusion models (Alcaraz & Strodthoff, 2023), can generate realistic minority-class waveforms, but are computationally expensive and offer limited control over multi-label condition combinations. Embedding-space interpolation such as SMOTE (Chawla et al., 2002) is cheap but ignores the structured geometry of deep representations and cannot model how specific pathological changes manifest in latent space.

We approach the imbalance problem from a different angle, asking: *can a model that understands how ECG representations evolve over time be repurposed to synthesise the representations of rare conditions?* To this end, we pretrain an action-conditioned latent transition model on longitudinal ECG pairs, conditioning latent-state predictions on the change in ICD diagnosis between consecutive recordings of the same patient. The trained dynamics model captures how pathological changes deform the representation space, and can be queried at will to synthesise embeddings for arbitrary target condition combinations — without ever generating a waveform.

Concretely, our contributions are:

- We adapt **LeJEPA** (Balestriero & LeCun, 2025) to longitudinal ECG pairs by introducing a label-delta action encoder, producing a transition model that predicts how a patient's ECG embedding changes with their clinical status.
- We propose **dynamics-based embedding augmentation**: the frozen predictor is used to synthesise 2M abnormal embeddings covering 300 multi-label condition combinations, directly addressing long-tail imbalance in representation space.
- On 76 ICD-coded conditions from MIMIC-IV-ECG, our augmented linear probe reaches AUROC 0.743, recovering 55% of the gap to a fully fine-tuned encoder

[*]Equal contribution [1]Department of Computer Science, Universidade Federal de Minas Gerais, Brazil. Correspondence to: Luiz Facury <luizfysouza@gmail.com>.

*LatinX in AI Workshop @ ICML 2026*, Seoul, South Korea. 2026.

while adding no parameters and requiring no additional waveforms.

## 2. Related Work

**ECG Deep Learning and Benchmarks.** Large-scale ECG classification has greatly advanced, with CNNs and transformers reaching cardiologist-level performance tasks (Hannun et al., 2019; Ribeiro et al., 2020). PTB-XL (Wagner et al., 2020) and its xResNet1d-50 benchmark (Strodthoff et al., 2020) remain the standard transfer evaluation, while MIMIC-IV-ECG (Gow et al., 2023) and its ICD-coded extension (Strodthoff et al., 2023) define the longitudinal, multi-label setting we target. Recent foundation models such as ECG-FM (McKeen et al., 2025), trained on over 1.5M recordings with hybrid SSL objectives, represent the current upper bound for representation-based approaches but do not address the longitudinal structure of repeated patient ECGs.

**Self-Supervised Learning for Biosignals.** Contrastive and predictive SSL have achieved competitive ECG representations without manual labels. CLOCS (Kiyasseh et al., 2021) uses positives at the patient and segment-level; 3KG (Gopal et al., 2021) designs physiologically grounded spatial augmentations; and Mehari & Strodthoff (Mehari & Strodthoff, 2022) provide a systematic comparison of SimCLR (Chen et al., 2020) , BYOL (Grill et al., 2020), and CPC (Oord et al., 2018) with an xResNet1d-50 encoder — the direct architectural predecessor of our encoder design. Masked-modelling approaches such as ST-MEM (Na et al., 2024) make use of the spatio-temporal structure of multi-lead recordings. However, all these methods employ a pretraining that treats each recording as an independent sample, leaving the rich temporal dynamics present in longitudinal datasets entirely unexploited.

**Joint-Embedding Predictive Architectures and World Models.** JEPA-style objectives (Assran et al., 2023; Bardes et al., 2024) learn by predicting target representations from context in latent space, avoiding the pitfalls of pixel-level reconstruction and yielding semantically richer features. Recent works, such as LeJEPA (Balestriero & LeCun, 2025), which we adopt as our backbone, strengthens this framework with SIGReg, a characteristic-function sketch that provably prevents representational collapse. The idea of conditioning latent predictions on actions is central to model-based reinforcement learning (Ha & Schmidhuber, 2018; Hafner et al., 2025), where world models learn to simulate future states given agent actions. We adapt this principle to the clinical setting by treating the change in the set of ICD labels of a patient between consecutive visits as a discrete "action", allowing the model to simulate the embedding of a future pathological state.

**Synthetic Augmentation for Class Imbalance.** Class imbalance is a persistent challenge in clinical ECG classification, where rare multi-condition combinations are severely underrepresented. SMOTE (Chawla et al., 2002) addresses this via interpolation in feature space, but was not designed for the high-dimensional, structured embeddings produced by deep encoders. Signal-space generative models, such as GAN-based synthesis (Delaney et al., 2019) and diffusion models conditioned on diagnostic labels (Alcaraz & Strodthoff, 2023), can produce realistic minority-class waveforms but are computationally expensive and operate on the raw signal rather than the learned representation. Our approach bypasses waveform generation entirely: by passing a normal embedding and a target label combination through the frozen dynamics model, we obtain synthetic embeddings that directly inhabit the encoder's representation space, providing multi-label control at negligible computational cost.

## 3. Method

### 3.1. Dataset and Task Setup

We use the MIMIC-IV-ECG dataset (Gow et al., 2023), a longitudinal collection of over 800,000 12-lead ECGs annotated with ICD codes mapped to $C = 76$ binary condition labels. The longitudinal structure allows us to create consecutive same-patient pairs $(x_t, x_{t+1})$. To prevent data leakage, all train, validation, and test folds are strictly patient-disjoint, ensuring no longitudinal pair crosses evaluation boundaries. We use folds 0–17 for training (721,002 recordings, 389,309 without any ICD-coded condition, denoted *normal*), fold 18 for validation, and fold 19 for testing. Performance is reported as macro-averaged AUROC.

### 3.2. Latent Transition Model Pretraining

We adopt **LeJEPA** (Balestriero & LeCun, 2025), a Joint-Embedding Predictive Architecture, as our world model backbone. The model encodes a waveform $x \in \mathbb{R}^{12\times T}$ through an xResNet1d-50 backbone $f_\theta$ and global average pooling to produce $h_t \in \mathbb{R}^{256}$, then projects it to $z_t = g_\phi(h_t) \in \mathbb{R}^{256}$ via a three-layer MLP projector with hidden dimensions $[2048, 2048, 256]$.

The transition between consecutive ECGs is encoded as the label delta $a_t = y_{t+1} - y_t \in \{-1, 0, 1\}^C$, capturing the full multi-label pathology change. We do not explicitly filter clinically implausible label transitions or noisy ICD codes, relying on the scale of the pretraining dataset to marginalize out infrequent labeling artifacts. An action encoder $p_\psi$ (two-layer MLP, $76 \to 512 \to 256$) maps $a_t$ to embedding space, and a predictor $r_\xi$ produces the future state estimate $\hat{z}_{t+1} = r_\xi([z_t;\, p_\psi(a_t)])$.

The pretraining loss is:

$$\mathcal{L} = (1-\lambda)\underbrace{\|\hat{z}_{t+1} - z_{t+1}\|_2^2}_{\mathcal{L}_{\text{pred}}} + \lambda\underbrace{\left[\mathcal{L}_{\text{SIG}}(z_t) + \mathcal{L}_{\text{SIG}}(z_{t+1})\right]}_{\mathcal{L}_{\text{SIG}}}, \tag{1}$$

with $\lambda = 0.1$. SIGReg prevents representational collapse by enforcing that real encoder outputs follow $\mathcal{N}(0, I)$ via a characteristic-function sketch over random projections. It is applied independently to both $z_t$ and $z_{t+1}$, but not to the predictor output $\hat{z}_{t+1}$.

### 3.3. Dynamics-Based Embedding Augmentation

After pretraining, we freeze the encoder and use the dynamics model to synthesize abnormal training embeddings. For each normal embedding $z_n$ and a target condition combination $\mathcal{C}$, we pass them through the frozen predictor to obtain a synthetic embedding:

$$z_{\text{aug}} = r_\xi([z_n;\, p_\psi(\mathcal{C})])\,. \tag{2}$$

The label assigned to $z_{\text{aug}}$ is the target combination $\mathcal{C}$ itself.

We enumerate all condition combinations with 1–4 co-occurring conditions appearing at least 50 times in the training set, selecting the top 300 by frequency. Synthetic samples are allocated inversely proportional to combination frequency, so that rare condition combinations receive more augmented samples:

$$n_{\mathcal{C}} \propto \frac{1}{\text{count}(\mathcal{C})}, \qquad \sum_{\mathcal{C}} n_{\mathcal{C}} = 2{,}000{,}000. \tag{3}$$

This yields a training set of 2.7M embeddings (721K real + 2M synthetic).

### 3.4. Downstream Probe

We train an MLP classifier on the frozen embeddings. The architecture consists of a linear projection to 512 dimensions, two pre-norm residual blocks, LayerNorm, and a linear output head for $C = 76$ conditions. We use BCE loss with class-specific positive weights (clamped at 5) and down-weight augmented samples by $w_{\text{aug}} = 0.3$. Training uses AdamW (lr $= 10^{-4}$, weight decay $10^{-2}$) with cosine annealing and early stopping (patience 50). We compare against: (i) a **real-only** probe trained on the 721K real embeddings only, and (ii) a fully **supervised** xResNet1d-50 trained end-to-end on raw waveforms as an upper-bound reference.

### 3.5. Experimental Setup

The world model is pretrained on all consecutive same-patient ECG pairs from the training folds using AdamW with learning rate $10^{-3}$, weight decay $5 \times 10^{-4}$, and batch size 512, for 50 epochs with a 5% linear warmup followed by cosine annealing to $10^{-5}$. The SIGReg trade-off is set to $\lambda = 0.1$ and the projection dimension to $d_z = 256$. Pretraining runs on a single NVIDIA RTX 4090.

The downstream MLP probe is trained on the frozen embeddings for up to 5,000 epochs with early stopping (patience 50), using AdamW with learning rate $10^{-4}$ and weight decay $10^{-2}$. Augmented samples are down-weighted by $w_{\text{aug}} = 0.3$ in the loss. All experiments use a fixed random seed for reproducibility.

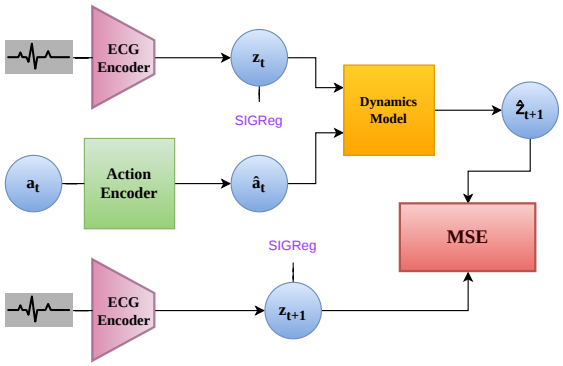


*Figure 1.* Overview of the LeJEPA world model pretraining. Consecutive ECG recordings from the same patient are encoded independently by a shared ECG encoder into embeddings $z_t$ and $z_{t+1}$. The action $a_t = y_{t+1} - y_t$ encodes the multi-label ICD transition and is mapped to latent space by the Action Encoder. The Dynamics Model predicts the future embedding $\hat{z}_{t+1}$ from $z_t$ and the action embedding, supervised by MSE against the real $z_{t+1}$. SIGReg is applied to both encoder outputs to prevent representational collapse.

*Table 1.* Macro-averaged AUROC on MIMIC-IV-ECG test fold (76 conditions). † End-to-end training on raw waveforms. All other methods operate on a LeJEPA-pretrained encoder; the dynamics model is not used at inference.

| Method | AUROC |
|---|---|
| Supervised xResNet1d-50† | 0.772 |
| Fine-tuned (no augmentation) | **0.789** |
| Linear probe, real data only | 0.687 |
| Linear probe + dynamics aug. (ours) | 0.743 |

## 4. Results

Table 1 reports macro-averaged AUROC on the held-out test fold (fold 19) for all methods. All probing and fine-tuning experiments share the same LeJEPA-pretrained xResNet1d-50 encoder; the dynamics model is discarded at inference for every baseline.

Fine-tuning the LeJEPA-pretrained encoder surpasses the fully supervised baseline (0.789 vs. 0.772), suggesting that

*Figure 2.* t-SNE projection of 7,000 embeddings from the LeJEPA encoder. Augmented embeddings (orange diamonds) generated by the frozen dynamics model overlap broadly with both real normal (gray) and real abnormal (blue) embeddings, confirming that dynamics-based synthesis operates within the encoder's learned representation space rather than producing out-of-distribution samples.

the world-model pretraining objective yields representations that transfer more effectively than training from scratch on the classification task alone. Among frozen-encoder methods, the real-only linear probe (0.687) lags considerably behind fine-tuning, as expected given that no gradient flows through the encoder. Our dynamics-augmented probe recovers a substantial portion of this gap (0.743, +5.6 points over real-only), demonstrating that the action-conditioned predictor generates label-consistent synthetic embeddings that meaningfully improve generalisation on rare condition combinations — without any additional waveforms, encoder updates, or waveform-level generation. The remaining gap to fine-tuning (0.789 vs. 0.743) is expected, since augmented embeddings are ultimately filtered through a frozen representation not optimised for the downstream task. Closing this gap further, for instance through richer action encodings or augmentation-aware encoder training, is a promising direction for future work.

We also note that the inverse-frequency allocation of synthetic samples plays an important role: rare condition combinations — which appear fewer than 100 times in the training set — receive up to 13,000 synthetic samples each, directly counteracting the long-tail distribution without any manual resampling strategy. The t-SNE projection in Figure 2 provides geometric evidence that augmented embeddings occupy the same region of representation space as their real counterparts, validating that the dynamics model synthesises plausible rather than out-of-distribution embeddings.

## 5. Conclusion

We presented an action-conditioned latent transition approach to long-tail multi-label ECG classification, repurposing the action-conditioned dynamics of a LeJEPA-pretrained encoder to synthesise rare-condition embeddings directly in representation space. Our method requires no additional waveforms, no generative model, and no encoder updates at augmentation time, yet recovers 55% of the gap between a real-data linear probe and a fully fine-tuned encoder on 76 ICD-coded conditions from MIMIC-IV-ECG.

These results suggest that the temporal structure of longitudinal clinical data — typically treated as a nuisance or ignored entirely — carries exploitable signal about how pathological states deform the representation space. We view this as an early proof of concept for a broader class of methods that use patient-level dynamics as a free source of supervision for data augmentation in clinical settings.

**Limitations and future work.** The quality of synthetic embeddings is bounded by the fidelity of the dynamics model. As an early proof of concept, our evaluation does not yet compare against simpler imbalance remedies (e.g., random oversampling, class-balanced loss, or embedding-space SMOTE), which is necessary to isolate the specific value of the transition dynamics. Furthermore, the augmentation strategy assumes a normal embedding is a valid starting point for any target condition, ignoring strong patient-specific anatomical confounders. Future work must include causal ablations (e.g., shuffled label deltas, random target combinations) to confirm the model learns clinically meaningful transitions rather than exploiting label frequencies. We will also expand validation beyond t-SNE to include per-label AUROC, calibration, and external evaluation on cohorts like PTB-XL, alongside exploring richer action representations and augmentation-aware encoder fine-tuning.

## References

Alcaraz, J. M. L. and Strodthoff, N. Diffusion-based conditional ecg generation with structured state space models. *Computers in biology and medicine*, 163:107115, 2023.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

Balestriero, R. and LeCun, Y. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. Revisiting feature

prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Delaney, A. M., Brophy, E., and Ward, T. E. Synthesis of realistic ecg using generative adversarial networks. *arXiv preprint arXiv:1909.09150*, 2019.

Gopal, B., Han, R., Raghupathi, G., Ng, A., Tison, G., and Rajpurkar, P. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Machine learning for health*, pp. 156–167. PMLR, 2021.

Gow, B., Pollard, T., Nathanson, L. A., Johnson, A., Moody, B., Fernandes, C., Greenbaum, N., Waks, J. W., Eslami, P., Carbonati, T., et al. Mimic-iv-ecg: Diagnostic electrocardiogram matched subset*Type: dataset*, 2023.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Ha, D. and Schmidhuber, J. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse control tasks through world models. *Nature*, 640 (8059):647–653, 2025.

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., and Ng, A. Y. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019.

Kiyasseh, D., Zhu, T., and Clifton, D. A. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International conference on machine learning*, pp. 5606–5615. PMLR, 2021.

McKeen, K., Masood, S., Toma, A., Rubin, B., and Wang, B. Ecg-fm: An open electrocardiogram foundation model. *Jamia Open*, 8(5):ooaf122, 2025.

Mehari, T. and Strodthoff, N. Self-supervised representation learning from 12-lead ecg data. *Computers in biology and medicine*, 141:105114, 2022.

Na, Y., Park, M., Tae, Y., and Joo, S. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *arXiv preprint arXiv:2402.09450*, 2024.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P., Andersson, C. R., Macfarlane, P. W., Meira Jr, W., et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1760, 2020.

Strodthoff, N., Wagner, P., Schaeffter, T., and Samek, W. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE journal of biomedical and health informatics*, 25(5):1519–1528, 2020.

Strodthoff, N., Alcaraz, J. M. L., and Haverkamp, W. Prospects for ai-enhanced ecg as a unified screening tool for cardiac and non-cardiac conditions–an explorative study in emergency care. *arXiv preprint arXiv:2312.11050*, 2023.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., and Schaeffter, T. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154, 2020.