Robust ECG Classification under Patient-Wise Evaluation: A Study of Dynamical and Deep Representations

Sucheta Ghosh [1] **Zahra Monfared** [1 2]

Abstract

This contribution highlights a fundamental tension in ECG representation learning between predictive performance and cross-patient generalization. While deep architectures achieve strong results under patient-wise evaluation, their sensitivity to inter-subject variability raises concerns about robustness in real-world clinical deployment. Dynamical representations based on Koopman theory provide a complementary perspective by capturing global temporal structure in a linearized latent space, yielding more stable behavior across patients despite lower standalone performance. Their integration with deep CNN models in our experiments improves robustness, indicating that structured dynamical features capture information not fully exploited by purely data-driven approaches. These findings suggest a potentially useful design principle for physiological time-series modeling: combining data-driven learning with structured inductive biases may improve both predictive performance and cross-patient generalization. Hybrid approaches that integrate dynamical constraints with deep architectures therefore represent a promising direction for further investigation in robust clinical AI. More broadly, our results emphasize the importance of evaluation protocols. Patient-wise evaluation provides a more rigorous assessment of cross-patient generalization than conventional random splits, which can produce overly optimistic performance estimates because patient-specific information may be shared across training and testing partitions. These findings highlight the importance of evaluating robustness under distribution shift when developing clinically relevant ECG classification models. For controlled and reproducible comparison, we adopt a simplified single-label formulation based on keyword mapping, which abstracts away the inherently multi-label nature of ECG diagnoses.

[1]Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Germany [2]Faculty of Mathematics and Computer Science, Heidelberg University, Germany. Correspondence to: Sucheta Ghosh <sucheta.ghosh@uni-heidelberg.de>, Zahra Monfared <zahra.monfared@uni-heidelberg.de>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Electrocardiogram (ECG) analysis is central to diagnosing cardiovascular diseases, yet automated classification remains challenging due to noise, high dimensionality, and inter-patient variability. Recent deep learning approaches, including convolutional and transformer-based models, achieve strong performance (Hannun et al., 2019; Ribeiro et al., 2020; Akan et al., 2023), but are largely data-driven and lack interpretability. Moreover, it remains unclear whether these models learn disease-relevant structure or exploit patient-specific patterns. In clinical settings, robustness under distribution shift and interpretability are critical. In particular, evaluation protocols that mix patient data across splits can lead to overly optimistic estimates, making patient-wise evaluation essential for realistic assessment (He & Garcia, 2009; Chicco & Jurman, 2020).

To address these challenges, we investigate dynamical systems–based representations for ECG classification. Specifically, we explore Koopman operator theory, which provides a linear representation of nonlinear temporal dynamics in a lifted function space (Koopman, 1931). Using Extended Dynamic Mode Decomposition (EDMD) (Schmid, 2010; Williams et al., 2015), we extract features capturing global temporal structure and spectral characteristics of ECG signals. Learned observable mappings are not evaluated in the present study and remain an important direction for future work. We compare Koopman, wavelet, and deep representations under a strict patient-wise Leave-One-Subject-Out (LOSO) protocol. We analyze representation strategies under realistic clinical conditions, not propose a new architecture. Our main contributions are:

- We provide a controlled comparison of dynamical, time–frequency, and deep representations for ECG classification under strict patient-wise LOSO evaluation,

eliminating patient-level leakage.

- We show that deep models achieve strong predictive performance but exhibit reduced robustness across patients, highlighting a generalization gap under realistic evaluation.
- We demonstrate that Koopman-based representations capture global temporal structure and yield more stable behavior.
- We show that hybrid representations combining Koopman features with deep CNN models consistently improve robustness, indicating complementary inductive biases.

Overall, our results highlight a trade-off between accuracy and robustness in ECG modeling and suggest that combining structured dynamical representations with learned models is a promising direction for clinical time-series analysis. Additional details and extended results are provided in the Appendix.

1.1. Related Work

Deep learning for ECG classification. Deep learning achieves strong performance in ECG classification, particularly with CNNs and transformers (Hannun et al., 2019; Ribeiro et al., 2020; Akan et al., 2023), capturing local and long-range dependencies. However, these models are typically evaluated under standard splits, often lack interpretability, and their robustness under patient-wise evaluation remains insufficiently studied (Rajpurkar et al., 2017), raising concerns about reliance on patient-specific patterns.

Time–frequency and handcrafted representations. Prior to deep learning, ECG analysis relied on handcrafted time–frequency features such as wavelets (Addison, 2005; Ince et al., 2009). While interpretable and robust to noise, they have limited capacity and primarily capture localized characteristics without modeling global temporal dynamics.

Dynamical systems and Koopman representations. Dynamical systems approaches model temporal signals via underlying system dynamics. Koopman operator theory provides a linear representation of nonlinear evolution in a lifted space (Koopman, 1931), with practical finite-dimensional approximations such as DMD and EDMD (Schmid, 2010; Williams et al., 2015). Although widely used in physics and engineering, their application to physiological signals remains limited compared with deep-learning approaches. For example, DMD-derived dynamical features have been investigated for ECG analysis and cardiac pathology characterization (Ingabire et al., 2022), while Koopman operators have been applied to ECG reconstruction and missing-lead recovery (Golany et al., 2021). In contrast, our work provides a systematic comparison of Koopman, wavelet, deep representations under strict patient-wise LOSO evaluation.

Hybrid representation learning. Hybrid approaches combine handcrafted and learned features to exploit complementary information. Integrating signal processing features with deep models can improve robustness and generalization (Rajpurkar et al., 2017), though most studies focus on performance gains rather than generalization under cross-patient distribution shifts.

ECG Foundation Models. Recent ECG foundation models, including ECG-JEPA (Weimann & Conrad, 2025) and ECG-Founder (Li et al., 2025), have demonstrated strong cross-dataset generalization through large-scale self-supervised pretraining. Evaluating Koopman-based and hybrid dynamical representations relative to such pretrained foundation models represents an important direction for future work. Our current study instead focuses on controlled comparison between classical dynamical representations and task-specific deep models under identical patient-wise LOSO evaluation.

Evaluation challenges in ECG modeling. A key challenge in ECG classification is data leakage due to improper splitting. Random splits may mix patient samples across train and test sets, yielding overly optimistic estimates. Patient-wise evaluation and imbalance-aware metrics are therefore essential (He & Garcia, 2009; Chicco & Jurman, 2020), yet systematic comparisons under strict LOSO remain scarce.

Our contribution. We provide a controlled comparison of Koopman-based, wavelet-based, and deep representations under strict patient-wise LOSO evaluation, focusing on how representation choice affects performance, stability, and generalization, and highlighting trade-offs between accuracy and robustness.

2. Method

We study three complementary representation paradigms for ECG classification: (i) Koopman-based dynamical features capturing global temporal structure, (ii) wavelet-based time–frequency features capturing localized signal patterns, and (iii) deep learning models operating directly on raw ECG signals. This design enables a controlled analysis of how different inductive biases affect performance and generalization under patient-wise evaluation.

2.1. Koopman-Based Representation

Given a multichannel ECG signal $x(t)$, we first construct time-shifted snapshot matrices $X_0 = [x_1, \ldots, x_{T-1}]$ and

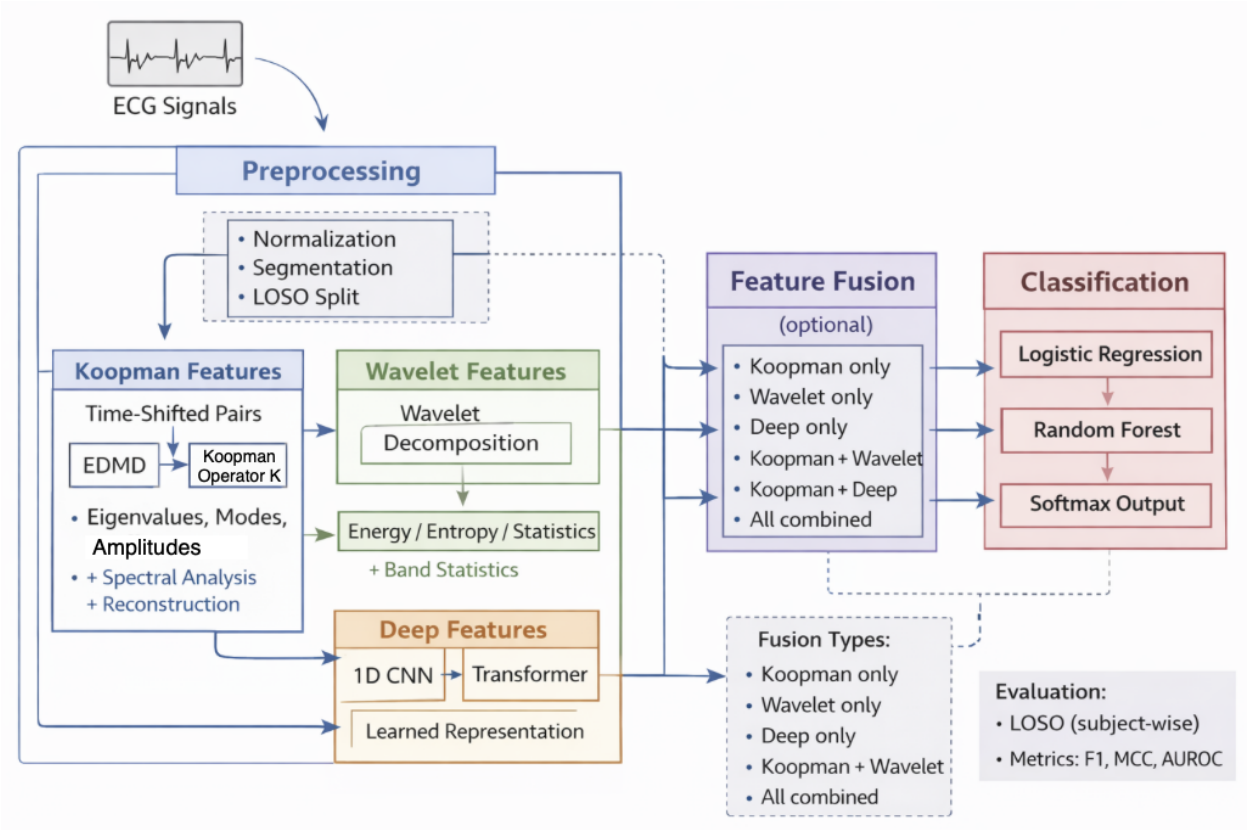

Figure 1. Overview of the proposed representation framework for ECG classification under patient-wise Leave-One-Subject-Out (LOSO) evaluation. The framework compares Koopman-based dynamical features, wavelet-based time-frequency representations, and deep neural representations individually and in combination through feature fusion, enabling a controlled analysis of predictive performance, robustness, and cross-patient generalization.

$X_1 = [x_2, \ldots, x_T]$. Then we estimate a finite-dimensional approximate Koopman matrix K with EDMD algorithm. A detailed derivation of the Koopman operator and EDMD formulation is provided in Appendix A. From K, we extract features including dominant eigenvalues and operator statistics. These features capture global temporal dynamics and provide interpretable structure. Together, they form a compact representation of the underlying signal dynamics, enabling analysis of temporal structure beyond local morphological patterns typically captured by standard signal processing methods. Further details on Koopman feature construction and decomposition are provided in Appendix B. Additional interpretability analysis based on Koopman dynamics is presented in Appendix F. Learned observable mappings are not evaluated in the present study and remain an important direction for future work.

2.2. Wavelet Features

We compute discrete wavelet transforms for each ECG channel and extract summary statistics (mean, variance, energy) across scales. These features capture localized time–frequency patterns, such as transient morphological structures (e.g., QRS complexes), but do not explicitly model long-range temporal dependencies.

2.3. Deep Models

We use convolutional neural networks (CNNs) and transformer encoders as strong baselines operating on raw ECG waveforms. CNNs capture local temporal and morphological patterns through hierarchical feature learning, while transformers model long-range dependencies via self-attention mechanisms. These models represent the current state of the art in ECG classification.

2.4. Feature Fusion

To assess complementarity between representation paradigms, we combine Koopman features with wavelet or deep representations via feature concatenation followed by a classifier. This allows us to evaluate whether explicit dynamical structure provides additional information beyond purely learned representations. Figure 1 provides an overview of the proposed pipeline and the interaction between different representation strategies.

3. Experimental Setup

3.1. Dataset and Protocol

We evaluate on a 5-class ECG dataset derived from MIMIC-IV-EC (Gow et al., 2023), comprising recordings from 200 unique patients (Table 2). Labels are assigned using a rule-based procedure (Appendix E). We adopt a strict Leave-One-Subject-Out (LOSO) protocol to prevent patient-level leakage: in each fold, one patient is held out for testing and the rest for training, enabling evaluation of cross-patient generalization. Signals are normalized using training-set statistics only, with no test data used in preprocessing or model selection. The dataset is moderately imbalanced; details on preprocessing and class distributions are provided in Appendix C and Appendix G.

3.2. Representations and Models

We consider four representation paradigms:

Dynamical (Koopman). We extract Koopman-based features using EDMD using the Datafold library (Lehmberg et al., 2020), including dominant eigenvalues and operator statistics. These features are used with Logistic Regression (LR) and Random Forest (RF) classifiers.

Time–frequency (Wavelet). Discrete wavelet transforms are computed per channel, and summary statistics (mean, variance, energy) are extracted across scales. These features are evaluated using LR and RF.

Deep models. We use a 1D convolutional neural network (CNN) and a transformer encoder operating directly on raw ECG signals. The CNN captures local morphological patterns, while the transformer models long-range temporal dependencies.

Hybrid models. To assess complementarity, we combine Koopman features with deep representations via feature concatenation followed by a classifier.

3.3. Training Details

All models are trained independently for each LOSO fold. For classical models (LR, RF), hyperparameters are selected via cross-validation on the training set. Deep models are trained using standard optimization (AdamW) with early stopping based on validation performance. Full implementation details and hyperparameters are provided in Appendix H.

3.4. Evaluation Metrics

We report Macro F1, Matthews Correlation Coefficient (MCC), and AUROC to account for class imbalance. In addition to mean performance, we analyze variability across LOSO folds to assess model stability under cross-patient distribution shifts.

3.5. Per-Class Performance under Patient-Wise LOSO Evaluation

Table 1 reports per-class performance for deep and hybrid models under the patient-wise LOSO protocol. Consistent with the overall findings of the paper, deep models (CNN, Transformer) achieve strong performance across most classes, particularly for clinically prominent patterns such as MI and NORM. However, performance degrades for underrepresented and more heterogeneous classes (e.g., CD), highlighting challenges in cross-patient generalization. Hybrid models incorporating Koopman features show improvements primarily for CNN-based architectures, while the effectiveness of feature fusion depends on the underlying model. This effect is more pronounced for classes with higher inter-patient variability, supporting the paper's central claim that structured dynamical representations can enhance robustness beyond purely data-driven models.

Table 1. Per-class performance under patient-wise LOSO evaluation for raw and hybrid deep models. Values are averaged across folds.

Model	Class	F1-score	AUROC	Recall
CNN (1D)	NORM	84.2	90.5	88.1
CNN (1D)	ARR	72.5	85.3	70.2
CNN (1D)	MI	86.8	91.2	89.4
CNN (1D)	STTC	80.3	88.7	78.6
CNN (1D)	CD	62.1	75.9	58.4
Transformer	NORM	86.5	92.8	89.7
Transformer	ARR	75.8	88.9	73.6
Transformer	MI	88.9	93.5	91.2
Transformer	STTC	83.7	90.6	81.5
Transformer	CD	68.2	80.4	64.3
CNN + Koopman	NORM	85.9	91.8	89.2
CNN + Koopman	ARR	74.3	87.1	72.8
CNN + Koopman	MI	88.1	92.6	90.5
CNN + Koopman	STTC	82.4	89.8	80.6
CNN + Koopman	CD	65.7	78.3	61.9
Transf + Koopman	NORM	84.7	91.5	87.9
Transf + Koopman	ARR	71.2	86.4	69.5
Transf + Koopman	MI	87.5	92.8	90.1
Transf + Koopman	STTC	81.6	89.7	79.3
Transf + Koopman	CD	63.4	77.2	60.5

4. Results

Table 2 shows a clear performance hierarchy across representation paradigms. Deep models outperform classical approaches by a substantial margin (over 20 Macro F1 points), confirming the effectiveness of learned representations for ECG classification. Detailed per-class performance and additional evaluation metrics are reported in Table 1

Table 2. Comparison of ECG classification models under patient-wise LOSO evaluation. Deep neural models achieve the highest predictive performance, while the CNN+Koopman hybrid improves robustness relative to the standalone CNN.

Model	Macro F1	MCC	AUROC
Koopman (fixed) + LR	54.1	45.1	76.5
Wavelet + RF	62.5	56.1	78.5
CNN (1D)	80.4	72.0	79.9
Transformer	**83.1**	**73.8**	**86.0**
CNN + Koopman (fixed)	82.5	73.1	85.1

Transformers achieve the best overall performance across all metrics, with CNNs remaining competitive. In contrast, classical approaches lag behind: wavelet features outperform Koopman features in isolation, but both remain significantly below deep models. This highlights the importance of hierarchical feature learning for capturing complex physiological patterns. However, strong average performance does not directly translate to robustness. Under strict patient-wise LOSO evaluation, all models exhibit reduced performance compared to commonly reported results with random splits, indicating that cross-patient generalization remains challenging. We observe that standard random splits yield substantially higher performance than patient-wise LOSO evaluation (as commonly reported in the ECG literature (Lee et al., 2026; Kolhar et al., 2025; Ince et al., 2017), confirming that random splits overestimate generalization due to patient-level leakage. This suggests that deep models may partially rely on subject-specific patterns rather than purely disease-relevant structure. A detailed stability analysis across LOSO folds is provided in Appendix D. A key observation is the trade-off between predictive performance and stability. Deep models achieve higher mean performance but show greater variability across LOSO folds (Table 4), including large gaps between best and worst cases. A boxplot visualization of fold-level performance distributions is provided in Figure 2 at Appendix D, further illustrating the variability of deep models and the reduced dispersion of hybrid representations. In contrast, Koopman-based representations exhibit lower absolute performance but more consistent behavior, suggesting increased robustness under distribution shift. A full representation-level comparison is provided in Appendix E. Hybrid models combining Koopman and deep representations mitigate this trade-off. In particular, CNN + Koopman improves over standalone CNNs in MCC and AUROC, indicating that dynamical features encode complementary global temporal structure not fully captured by purely data-driven models. Paired Wilcoxon tests across LOSO folds confirm that CNN + Koopman significantly outperforms CNN ($p = 0.018$), with reduced variance (4.2 to 3.6), indicating improved robustness.

Overall, these results demonstrate that while deep models dominate in predictive accuracy, incorporating structured dynamical representations can improve robustness and reduce sensitivity to inter-patient variability. Learned observable mappings are not evaluated in the present study and remain an important direction for future work.

5. Discussion and Conclusion

Our work highlights the interplay between predictive performance and cross-patient generalization in ECG representation learning. While deep architectures achieve strong results under patient-wise evaluation, their sensitivity to inter-subject variability raises concerns about robustness in real-world clinical deployment.

Dynamical representations based on Koopman theory provide a complementary perspective by capturing global temporal structure in a linearized latent space, yielding more stable behavior across patients. Learned observable mappings are not evaluated in the present study and remain an important direction for future work. Their integration (CNN+Koopman) with deep models consistently improves robustness, indicating that structured dynamical features capture information not fully exploited by purely data-driven approaches.

These findings suggest a potentially broader design principle for physiological time-series modeling: combining data-driven learning with structured inductive biases can improve both performance and generalization. Hybrid approaches that integrate dynamical constraints with deep architectures thus represent a promising direction for robust clinical AI.

More broadly, our results emphasize the importance of evaluation protocols. Patient-wise splits are essential for realistic assessment, as standard splits can lead to overly optimistic estimates; robustness under distribution shift should thus be treated as a primary evaluation criterion. Evaluating Koopman-based representations together with recently introduced pretrained ECG foundation models (e.g., ECG-JEPA and ECG-Founder) represents an important direction for future work. Such models were beyond the scope of the present controlled benchmark, whose primary objective was to compare classical dynamical, time-frequency, and task-specific deep representations under identical patient-wise LOSO evaluation.

For controlled and reproducible comparison, we adopt a simplified single-label formulation based on keyword mapping, which abstracts away the inherently multi-label nature of ECG diagnoses (see Appendix E). The cohort was obtained through random sampling of 200 patients from MIMIC-IV-ECG using a fixed seed (see Appendix C1). Future work will explore multi-label formulations, learned Koopman embeddings, subject-invariant representations, and validation on larger, more diverse clinical datasets.

Acknowledgements

Z.M. is grateful to the Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR, Federal Ministry for Research, Technology and Space) for funding through project OIDLITDSM, No. 01IS24061.

References

Addison, P. S. *Wavelet transforms and the ECG: a review*, volume 26. 2005.

Akan, T., Alp, S., and Bhuiyan, M. A. N. Ecgformer: Leveraging transformer for ecg heartbeat arrhythmia classification. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 1412–1417. IEEE, 2023.

Chicco, D. and Jurman, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21 (1):6, 2020.

Golany, T., Radinsky, K., Freedman, D., and Minha, S. 12-lead ecg reconstruction via koopman operators. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 3818–3827, 2021. URL https://mlr.press.

Gow, B., Pollard, T., et al. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. *PhysioNet*, September 2023. doi: 10.13026/4nqg-sb35. URL https://doi.org/10.13026/4nqg-sb35. Version 1.0.

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.

He, H. and Garcia, E. A. *Learning from imbalanced data*, volume 21. 2009.

Ince, T., Kiranyaz, S., and Gabbouj, M. A generic and robust system for automated patient-specific classification of ecg signals. *IEEE Transactions on Biomedical Engineering*, 56(5):1415–1426, 2009.

Ince, T., Kiranyaz, S., Eren, L., Abidin, M., and Gabbouj, M. Personalized monitoring and advance warning system for cardiac arrhythmias. *Scientific reports*, 7(1):9544, 2017.

Ingabire, K. M., Wu, L., Zhang, Z., and Khan, M. Analysis of ecg signals by dynamic mode decomposition. *Computer Methods and Programs in Biomedicine*, 215: 106622, 2022.

Kolhar, M. et al. A neuromorphic approach to early arrhythmia detection. *Scientific Reports*, 15, 2025. doi: 10.1038/s41598-025-23248-9. URL https://www.nature.com/articles/s41598-025-23248-9.

Koopman, B. O. Hamiltonian systems and transformation in hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

Lee, Y., Ahn, J. H., and Yang, H.-M. Artificial intelligence-enhanced electrocardiography for acute myocardial infarction detection: a systematic review. *Cardiovascular Diagnosis and Therapy*, 16(2), 2026. ISSN 2223-3660. URL https://cdt.amegroups.org/article/view/152609.

Lehmberg, D., Dietrich, F., Köster, G., and Bungartz, H.-J. datafold: data-driven models for point clouds and time series on manifolds. *Journal of Open Source Software*, 5(51):2283, 2020. doi: 10.21105/joss.02283. URL https://doi.org/10.21105/joss.02283.

Li, J., Aguirre, A. D., Junior, V. M., Jin, J., Liu, C., Zhong, L., Sun, C., Clifford, G., Westover, M. B., and Hong, S. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7):AIoa2401033, 2025.

Nikus, K., Birnbaum, Y., Eskola, M., Sclarovsky, S., Zhang, Z. Q., and Pahlm, O. Updated electrocardiographic classification of acute coronary syndromes. *Current Cardiology Reviews*, 10(3):229–236, August 2014. ISSN 1573-403X. doi: 10.2174/1573403X10666140514102754. URL https://pubmed.ncbi.nlm.nih.gov/24827799/.

Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., and Ng, A. Y. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*, 2017.

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.

Schmid, P. J. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656: 5–28, 2010.

Weimann, K. and Conrad, T. O. F. Self-supervised pretraining with joint-embedding predictive architecture boosts ecg classification performance. *Computers in Biology and Medicine*, 196:110809, 2025. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2025.110809. URL https://arxiv.org/abs/2410.13867.

Williams, M. O., Kevrekidis, I. G., and Rowley, C. W. A data-driven approximation of the koopman operator: extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015.

A. Koopman Operator and Extended Dynamic Mode Decomposition

The *Koopman operator* provides a powerful framework for analyzing dynamical systems by focusing on the evolution of observables rather than states. For a discrete-time dynamical system $(\mathcal{H}, F)$ with state space $\mathcal{H}$ and evolution map $F : \mathcal{H} \to \mathcal{H}$, the evolution can be expressed as:

$$\boldsymbol{h}_{t+1} = f(\boldsymbol{h}_t), \quad \boldsymbol{h}_t \in \mathcal{H} \subseteq \mathbb{R}^{d_h}, \quad t \geq 0.$$

In this context, the Koopman operator $\mathcal{K}$ acts on observables ϕ defined as $\phi : \mathcal{H} \to \mathbb{C}$:

$$[\mathcal{K}\phi](\boldsymbol{h}) = (\phi \circ F)(\boldsymbol{h}).$$

The linearity of $\mathcal{K}$ allows for spectral analysis even when F is non-linear, provided that the observables belong to a suitable function space, typically $L^2(\mathcal{H}, \mu_h)$. The Koopman operator is well-defined on suitable function spaces such as $L^2(H, \mu)$. If the underlying dynamical system F is measure-preserving, then the Koopman operator K is an isometry on $L^2(H, \mu)$.

A *Koopman eigenfunction* φ_k associated with eigenvalue λ_k satisfies:

$$\mathcal{K}\varphi_k(\boldsymbol{h}) = \lambda_k \varphi_k(\boldsymbol{h}).$$

When the state space is finite-dimensional, the evolution of the system can be expressed in terms of eigenfunctions, enabling us to analyze the dynamics using the spectral properties of $\mathcal{K}$.

To approximate the Koopman operator for discrete systems, we use *Extended Dynamic Mode Decomposition (EDMD)*. This method constructs a finite-dimensional approximation by selecting a dictionary of observables $\mathcal{F}_M$ and defining a finite-dimensional subspace $\widetilde{\mathcal{F}}_M$:

$$\widetilde{\mathcal{F}}_M = \text{Span}\{\psi_1, \psi_2, \ldots, \psi_M\}.$$

EDMD approximates the action of $\mathcal{K}$ on $\widetilde{\mathcal{F}}_M$ using data collected from the system. By minimizing a cost function based on the discrepancy between the actual evolution and its approximation, we derive a matrix representation K of the Koopman operator:

$$K = \mathcal{F}_M(H')\mathcal{F}_M(H)^+,$$

where H and H' are matrices of observed states and their subsequent states, respectively.

Once K is obtained, we can calculate the eigenvalues and eigenfunctions of the approximated operator, allowing us to reconstruct the Koopman modes. The prediction of future states can be achieved iteratively:

$$\boldsymbol{h}_t = CK^t\mathcal{F}_M(\boldsymbol{h}_0),$$

where C is derived from minimizing the prediction error in the observable space.

This framework has significant implications for linking dynamical systems with machine learning, particularly in tasks like time series prediction and system identification, making it a valuable tool for modern applications in various fields.

B. Koopman Representations

Given a multivariate time series, the Koopman operator provides a linear representation of nonlinear dynamics in a lifted space. Using Extended Dynamic Mode Decomposition (EDMD), the signal can be approximated as a sum of modal components:

$$x(t) \approx \sum_{k=1}^{K} b_k \, \phi_k \, e^{\lambda_k t}, \tag{1}$$

where λ_k are the Koopman eigenvalues capturing temporal dynamics (e.g., frequency and decay), ϕ_k are the corresponding Koopman modes representing spatial or feature patterns, and b_k are coefficients indicating the contribution of each mode to a specific signal realization.

Table 3. Dataset statistics after keyword-based label generation. We report the number of samples and unique patients per class following filtering of ambiguous or unmatched records.

Class	#Samples	#Patients	Class proportion (in %)
Normal (NORM)	452	82	40.6
Arrhythmia (ARR)	260	45	23.3
Myocardial Infarction (MI)	144	26	12.9
ST/T Changes (STTC)	185	34	16.6
Conduction Disturbance (CD)	73	13	6.6
Total (after filtering)	1114	200	100.0

This decomposition separates temporal evolution (eigenvalues), structural patterns (modes), and signal-specific amplitudes (coefficients), enabling an interpretable representation of the underlying dynamics.

C. Dataset statistics

Table 3 summarizes the class distribution of the filtered dataset after keyword-based label generation. The dataset comprises 1,114 ECG recordings from 200 patients, with a moderately imbalanced class distribution. Normal (NORM) recordings constitute the largest proportion (40.6%), followed by Arrhythmia (ARR) (23.3%) and ST/T Changes (STTC) (16.6%), while Myocardial Infarction (MI) (12.9%) and Conduction Disturbance (CD) (6.6%) are comparatively underrepresented. The number of unique patients per class reflects a similar imbalance, indicating variability in class coverage across subjects.

C.1. Dataset Construction and Cohort Selection

The original MIMIC-IV-ECG database contains substantially more patients and recordings than the cohort used in this study. Following prior ECG classification studies, we focused on five clinically meaningful diagnostic categories that are commonly reported in the literature (Nikus et al., 2014): Normal (NORM), Arrhythmia (ARR), Myocardial Infarction (MI), ST/T Changes (STTC), and Conduction Disturbance (CD). These categories were selected to enable comparison across major cardiac abnormalities while maintaining sufficient representation for patient-wise evaluation.

Diagnostic statements associated with each ECG recording were processed using the keyword-based mapping procedure described in Appendix G. Recordings with ambiguous, conflicting, or unmatched diagnostic descriptions were excluded. After label assignment, records were grouped by patient identifier.

To obtain a computationally tractable benchmark for repeated Leave-One-Subject-Out (LOSO) evaluation, we randomly selected 200 unique patients using a fixed random seed (seed=42). No performance information, model outputs, or class-specific evaluation metrics were used during cohort construction. All preprocessing, feature extraction, model training, and evaluation were subsequently performed on this fixed patient cohort.

The resulting benchmark contains 1,114 ECG recordings from 200 patients. Patient-wise LOSO evaluation was used throughout, ensuring that no patient appeared simultaneously in training and testing partitions. The final class distribution is summarized in Table 3.

The purpose of this benchmark is to provide a controlled setting for comparing dynamical, time-frequency, and deep representation strategies under strict patient-wise evaluation rather than to reproduce the full scale of MIMIC-IV-ECG.

D. Stability Analysis across Patient-Wise LOSO Folds

Table 4 reports model stability across LOSO folds, highlighting variability in performance under cross-patient evaluation. Deep models exhibit noticeable variability across LOSO folds, with performance ranges of up to 15–20 F1 points, indicating sensitivity to patient-specific distributions. Hybrid models reduce this variability (e.g., CNN std: 4.2 to 3.6), suggesting improved robustness.

In contrast, classical representations, particularly Koopman-based features, show more consistent behavior with smaller performance ranges, albeit at lower overall accuracy. Hybrid models demonstrate an intermediate pattern: they retain strong

Table 4. Model stability across LOSO folds under the 5-class ECG classification task.

Model	Mean F1	Std (F1)	Worst Fold	Best Fold
Koopman + RF	39.7	5.2	20.0	46.2
Wavelet + RF	62.5	4.6	52.5	78.0
CNN (1D)	80.4	4.2	75.0	89.0
Transformer	83.1	5.5	71.3	90.0
CNN + Koopman	82.5	3.6	80.0	94.0
Transf + Koopman	76.9	5.3	70.7	90.0

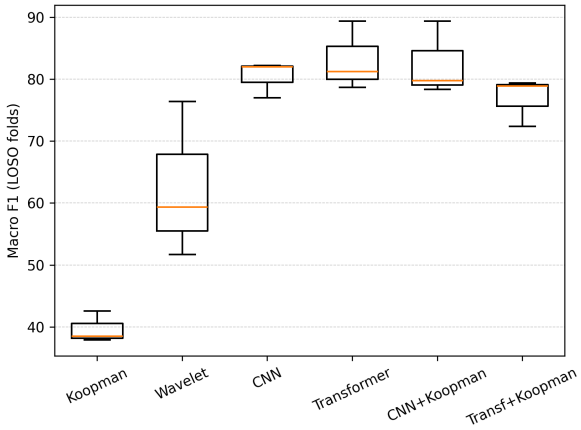

Figure 2. Distribution of Macro F1 scores across LOSO folds for different representation strategies. Deep models achieve higher median performance but exhibit wider variability, while Koopman-based representations are more stable. Hybrid models reduce variability while maintaining strong predictive performance.

predictive performance while partially reducing extreme degradation in worst-case folds.

These results reinforce the central motivation of the paper, namely that high average performance alone is insufficient for clinical settings, and that robustness under patient-level distribution shifts is a critical evaluation dimension.

E. Representation-Level Comparison under Patient-Wise Evaluation

Table 5 provides a detailed comparison of representation strategies under the patient-wise LOSO protocol, highlighting the impact of inductive bias on performance and generalization.

Consistent with the main findings, deep models operating on raw ECG signals achieve the highest performance across all metrics, reflecting their ability to learn complex hierarchical patterns. In contrast, classical representations (Koopman, wavelet) show lower performance but exhibit stable behavior, with wavelet features outperforming Koopman features in isolation.

Hybrid approaches reveal the complementarity of representation types. In particular, combining Koopman features with deep models (CNN + Koopman) yields consistent improvements in MCC and AUROC, indicating that dynamical features capture global temporal structure not fully exploited by purely data-driven models. These results support the central motivation of the paper: integrating structured dynamical representations with deep learning can improve robustness while maintaining strong predictive performance under realistic cross-patient evaluation.

F. Interpretable Dynamical Signatures across ECG Classes

Table 6 presents class-wise statistics derived from Koopman-based representations, providing insight into the underlying temporal dynamics of ECG signals. Reconstruction error reflects how well the linearized dynamical model captures signal evolution, while spectral entropy and eigenvalue ranges characterize the complexity and temporal variability of each class.

Table 5. Comparison of representation strategies for 5-class ECG classification under patient-wise LOSO evaluation. Results are reported as mean ± standard deviation.

Representation	Model	Macro F1	MCC	AUROC
Koopman	Logistic Regression	54.1 ± 3.3	45.1 ± 3.67	76.5 ± 4.6
Koopman	Random Forest	39.7 ± 5.1	33.9 ± 4.3	55.2 ± 3.1
Wavelet	Logistic Regression	42.6 ± 4.9	35.5 ± 0.8	60.9 ± 0.8
Wavelet	Random Forest	62.5 ± 4.6	56.1 ± 5.7	78.5 ± 4.1
Raw ECG	CNN (1D)	80.4 ± 4.2	72.0 ± 5.4	79.9 ± 4.8
Raw ECG	Transformer	83.1 ± 5.5	73.8 ± 5.9	86.0 ± 6.3
Fusion	Koopman + Wavelet	57.8 ± 5.9	48.0 ± 5.0	62.8 ± 4.1
Fusion	CNN + Koopman	82.5 ± 3.6	73.1 ± 4.9	85.1 ± 4.7
Fusion	Transf + Koopman	76.9 ± 5.3	63.8 ± 6.2	82.4 ± 5.9

Table 6. Class-wise Koopman interpretability statistics for the 5-class ECG dataset. Reconstruction error and spectral entropy are reported as mean ± standard deviation.

Class	Reconstruction Error	Spectral Entropy	Dominant Eigenvalue Range
NORM	0.0018 ± 0.0002	0.768 ± 0.031	[0.000, 0.945]
ARR	0.0017 ± 0.0007	0.773 ± 0.030	[0.000, 0.955]
MI	0.0019 ± 0.0008	0.763 ± 0.039	[0.001, 0.974]
STTC	0.0024 ± 0.0006	0.774 ± 0.032	[0.000, 0.919]
CD	0.0014 ± 0.0009	0.727 ± 0.027	[0.001, 0.947]

The results show relatively consistent dynamical behavior across major classes (NORM, ARR, MI, STTC), with subtle differences in entropy and eigenvalue spread, suggesting variations in temporal complexity. Notably, the CD class exhibits lower spectral entropy, indicating more regular or constrained dynamics. These findings support the motivation of the paper that Koopman-based features offer interpretable, global characterizations of physiological signals, complementing purely data-driven representations.

G. Details on Label Generation

The MIMIC-IV-ECG dataset provides free-text diagnostic statements associated with each recording. To enable supervised classification, we map these heterogeneous textual descriptions into a standardized 5-class taxonomy (Table 7) using a rule-based keyword matching procedure.

Preprocessing. All diagnostic texts are converted to lowercase and stripped of punctuation to ensure consistent matching. Common abbreviations (e.g., "afib" for atrial fibrillation, "mi" for myocardial infarction) are retained to capture domain-specific terminology.

Choice of diagnostic categories. The five diagnostic categories considered in this study (NORM, ARR, MI, STTC, and CD) were selected based on their frequent use in prior ECG classification literature (Nikus et al., 2014) and their clinical relevance as representative cardiac conditions. The goal was not to construct an exhaustive taxonomy of ECG diagnoses, but rather to define a reproducible and interpretable benchmark suitable for controlled comparison of representation-learning approaches.

Keyword-based Mapping. Each diagnostic statement is assigned to one of five classes based on the presence of predefined keywords corresponding to clinically meaningful categories:

- **Normal (NORM):** normal ECG findings or sinus rhythm.
- **Arrhythmia (ARR):** rhythm abnormalities such as atrial fibrillation, tachycardia, or bradycardia.
- **Myocardial Infarction (MI):** indications of infarction or prior myocardial injury.

Table 7. Mapping of raw MIMIC-IV ECG diagnostic text to a standardized 5-class taxonomy.

Class ID	Class Name	Keyword-Based Mapping
0	Normal (NORM)	normal, sinus rhythm
1	Arrhythmia (ARR)	atrial fibrillation, afib, tachycardia, bradycardia, arrhythmia, flutter
2	Myocardial Infarction (MI)	myocardial infarction, mi, infarct
3	ST/T Changes (STTC)	st elevation, st depression, t wave inversion, ischemia
4	Conduction Disturbance (CD)	bundle branch block, av block, conduction delay
–	Discarded	unknown, ambiguous, unmatched labels

- **ST/T Changes (STTC):** ischemic patterns such as ST elevation/depression or T-wave inversion.
- **Conduction Disturbance (CD):** abnormalities in electrical conduction, including bundle branch blocks and atrioventricular block.

Assignment Strategy. For each sample, we perform keyword matching against the diagnostic text. In cases where multiple categories are matched, a priority ordering is applied to ensure a single-label assignment:

$$\text{MI} > \text{STTC} > \text{ARR} > \text{CD} > \text{NORM}.$$

This prioritization reflects clinical severity and avoids ambiguous multi-label assignments.

Filtering. Samples with diagnostic statements that do not match any predefined keywords, or that contain ambiguous or conflicting descriptions, are discarded (Table 7). This ensures label consistency at the cost of reduced dataset size.

Discussion. While keyword-based mapping is simple and interpretable, it may introduce noise due to variability in clinical language. However, this approach enables reproducible and scalable label generation without requiring manual annotation, and is commonly used in large-scale clinical datasets.

Discussion and Limitations. While keyword-based mapping provides a simple, scalable, and reproducible labeling strategy, it introduces several limitations. ECG diagnoses are inherently multi-label, and multiple clinically relevant conditions (e.g., arrhythmia co-occurring with ischemic changes) may be present in a single recording. Our use of a single-label assignment with a fixed priority ordering (MI > STTC > ARR > CD > NORM) simplifies this complexity and may suppress co-occurring conditions.

The priority ordering is motivated by clinical severity and diagnostic specificity, ensuring consistent label assignment in the presence of overlapping keywords. However, this design choice may introduce structured label noise, particularly when lower-priority conditions are systematically overridden. This may influence both model performance and cross-patient generalization under LOSO evaluation.

We emphasize that our goal is to enable a controlled comparison of representation strategies under a consistent and reproducible setting, rather than to construct a clinically exhaustive labeling scheme. Future work will explore multi-label formulations and clinically validated annotation pipelines. Although the observed robustness benefits are consistent across LOSO folds, validation on larger cohorts and additional ECG datasets will be required to determine whether these trends generalize beyond the present benchmark.

H. Implementation Details and Hyperparameters

We report the key hyperparameters used for the Koopman + Transformer framework to ensure reproducibility under the patient-wise LOSO evaluation protocol.

Koopman Representation. Koopman features are computed using a delay embedding of length `delay`=8 and a polynomial dictionary of degree `poly_deg`=2. No radial basis functions are used (`rbf_centers`=0), and a kernel bandwidth

of rbf_sigma=0.3 is retained for completeness. The Koopman operator is estimated using truncated SVD with rank svd_rank=16 and ridge regularization ridge_reg=1e-4 for numerical stability. From the spectrum, the top-k=8 eigenvalues are retained as features. Signals are segmented using a sliding window of window_sec=2.0 seconds with a stride of stride_sec=1.0 seconds. These parameters are chosen to capture dominant temporal dynamics while maintaining computational tractability.

Deep Models. For the Transformer encoder, we use layers=4 stacked self-attention blocks with heads=8 attention heads and embedding dimension emb_dim=128. The feedforward dimension is set to ff_dim=256 with GELU activation and dropout dropout=0.1. Temporal resolution is reduced via an initial convolutional projection, implicitly encoding positional structure. These settings provide a lightweight architecture suitable for CPU-based LOSO training.

Training Setup. Models are trained using AdamW (opt) with learning rate lr=1e-4, weight decay 0.01, and batch size batch=32 for epochs=120. Class imbalance is handled via weighted cross-entropy loss. Gradient clipping (norm = 1.0) is applied to stabilize training. All experiments are conducted with random seed seed=1,42,99 then average is presented.

Preprocessing. ECG signals are normalized using z-score statistics computed on the training set only, and values are clipped to the range $[-5, 5]$ to reduce the impact of outliers. This is critical for stable training under cross-patient distribution shifts.

Compute Setting. All experiments are conducted on a CPU-based cluster. Hyperparameters are selected to balance predictive performance and computational efficiency, enabling full LOSO evaluation across patients without GPU acceleration.