DinoV3–LSTM for Early-Stage Classification of Paroxysmal Atrial Fibrillation in ECGs Toward Prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Detecting precursors of paroxysmal atrial fibrillation in long-term ECGs is crucial for timely intervention but remains difficult because precursor signals are subtle and patient physiology varies widely. We present a streamlined framework that adapts a distillation-based vision foundation model, DinoV3, to multi-channel two-dimensional ECG encodings and combines it with an LSTM to form an indicator model for precursor classification. We use three complementary 2D representations—short-time Fourier transform, Gramian angular field, and progressive moving-average transform—to capture time–frequency structure, temporal relationships, and multiscale trends, respectively. Evaluated as a precursor classification study on public long-term ECG datasets, the DinoV3–LSTM pipeline achieves competitive performance despite the absence of ECG-specific fine-tuning. Finally, we outline two concrete next steps to move from precursor classification to clinically useful prediction: ECG-domain fine-tuning with end-to-end optimization, and rigorous lead-time and prospective validation to quantify true prediction capability.

1 Introduction

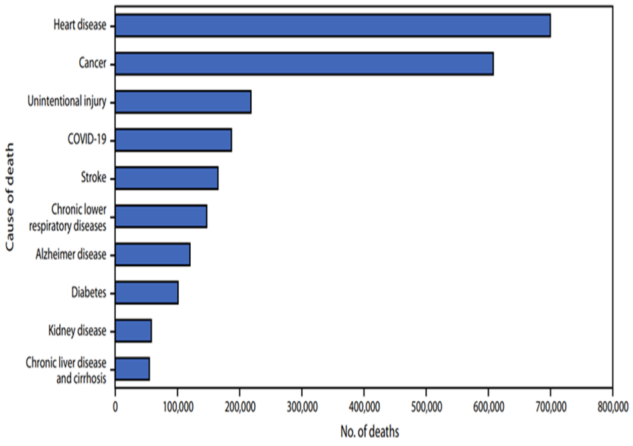

Figure 1: Global mortality statistics related to diseases

Cardiovascular diseases (CVDs) remain the leading cause of global mortality, a figure projected to rise with the accelerating global aging population [1, 2]. Among these, cardiac arrhythmias represent

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

a significant and often life-threatening subclass. Paroxysmal atrial fibrillation (PAF), a common type of irregular heart rhythm, is particularly dangerous as it can lead to severe complications such as stroke, heart failure, and sudden cardiac death if not detected and managed promptly [3]. While often asymptomatic, PAF can manifest as palpitations, shortness of breath, or dizziness, and its transient nature makes timely diagnosis and intervention challenging. For patients with underlying cardiac conditions or advanced heart failure, the sudden onset of PAF can be critically destabilizing, necessitating proactive and early warning systems.

The advancement of artificial intelligence (AI), particularly in deep learning, has spurred numerous research efforts in ECG-based arrhythmia classification, including PAF detection [4, 5]. These studies have demonstrated impressive accuracy in identifying various arrhythmias from ECG signals. However, a significant limitation of most existing AI-ECG approaches is their focus on classification, determining the current cardiac state based on a contemporaneous ECG recording. While valuable for diagnosis, this retrospective analysis falls short for patients where anticipation of an event is paramount. For vulnerable individuals, such as those with advanced heart failure or other severe cardiac comorbidities, a real-time predictive capability for impending PAF is not merely beneficial but essential for enabling proactive medical intervention and preventing catastrophic outcomes.

In response to this critical need, a growing body of research has explored predictive models for arrhythmias. These studies can generally be categorized into a few types: (1) **Feature-based Predictive Models** that rely on hand-crafted features from heart rate variability (HRV) or P-wave morphology extracted from preceding ECG segments, often utilizing machine learning classifiers like SVM or decision trees [6, 7]. (2) **End-to-End Deep Learning Classification** approaches that use neural networks to directly classify ECG segments into "normal" or "pre-AF" states within a fixed temporal window before onset [8]. (3) More recently, **Self-Supervised Learning (SSL) enhanced models** that leverage unlabeled ECG data to pre-train robust feature extractors. A notable example is a recent study utilizing an MAE-based 1D-ViT [9] for feature extraction, followed by an LSTM network for modeling state transitions and issuing early warnings for both PAF and ventricular fibrillation (VF) [10]. This pipeline demonstrated strong performance.

Despite these advances, the MAE-LSTM pipeline, while effective, presents a key limitation. It typically involves a decoupled training process in which the feature extractor (MAE-based 1D-ViT) and the temporal warning model (LSTM) are trained separately. In many cases, the pre-trained encoder's weights are frozen or only partially adapted during downstream supervised fine-tuning. This separation can restrict task-specific discriminability, as the features learned from a generic masking task may not be optimally tuned for capturing the subtle, evolving patterns of pre-arrhythmic states.

To address this limitation, we propose a novel end-to-end AF early warning framework that replaces the MAE-based feature extraction module with DinoV3, a state-of-the-art self-supervised learning backbone. We adapt DinoV3 to process 2D ECG patch representations (PMAT [11], STFT, GAF) and integrate it seamlessly with an LSTM warning head, enabling joint optimization across the entire pipeline. This end-to-end approach allows DinoV3 to refine its representation learning directly for the early detection of precursor electrophysiological changes, while also benefiting from the strong general priors inherited from its self-supervised pre-training.

It is noteworthy that, in our current study, DinoV3 was employed without additional fine-tuning on large-scale ECG datasets. As a result, its performance is slightly lower than the best-reported MAE-LSTM pipeline. Nonetheless, achieving comparable early warning accuracy without domain-specific fine-tuning underscores DinoV3's potential as a powerful backbone for arrhythmia prediction. We hypothesize that further fine-tuning and fully end-to-end optimization will unlock superior predictive performance. Moreover, DinoV3's attention mechanisms provide interpretable patch-level saliency maps, offering clinicians valuable insights into which ECG segments drive the early warning. This work thus lays the foundation for more accurate, interpretable, and clinically actionable AF early warning systems.

2 Related Work

This section reviews existing research pertinent to AF prediction and early warning systems, highlighting the progression from traditional methods to advanced deep learning and foundation models, ultimately setting the stage for our proposed DinoV3-based end-to-end framework.

2.1 AF Prediction and Early Warning Research

Early and accurate prediction of AF is paramount for improving patient outcomes and preventing severe complications such as stroke and heart failure [3]. Historically, AF prediction has largely relied on manual feature engineering from electrocardiogram (ECG) signals, combined with conventional machine learning classifiers.

Traditional ECG Features and Shallow Models. Classical approaches to AF prediction predominantly utilized hand-crafted features extracted from ECG recordings. These included various heart rate variability (HRV) metrics, which quantify beat-to-beat variations in heart rate, and morphological features such as P-wave duration, QRS complex characteristics, and T-wave abnormalities [6, 7]. Machine learning models like Support Vector Machines (SVMs), Logistic Regression, and XGBoost were then employed to classify these features as indicative of impending AF. While these methods offer a degree of interpretability and have shown promise in certain contexts, they suffer from significant limitations. Their performance is highly sensitive to the accuracy of fiducial point detection (e.g., P-wave onset/offset), which can be challenging in noisy or atypical ECG recordings. Furthermore, these models often require relatively long ECG windows (e.g., 5 minutes or more of R-R intervals) to extract stable and reliable features, which can delay timely warnings [12].

Deep Learning for AF Classification and Prediction. The advent of deep learning has revolutionized ECG analysis by enabling automatic feature extraction and classification directly from raw or minimally preprocessed signals [4, 5]. Convolutional Neural Networks (CNNs) have proven highly effective for learning complex spatio-temporal patterns in ECG data, achieving expert-level performance in arrhythmia detection and classification [13]. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, are well-suited for modeling the sequential nature of ECG data, capturing temporal dependencies crucial for predicting future events [14]. However, many early deep learning methods primarily focused on classification of the current state rather than prediction of future events. While valuable for diagnosis, this approach often falls short in providing the necessary lead time for proactive intervention, especially for patients requiring continuous monitoring for paroxysmal events.

Pathway Modeling and Early Warning Frameworks. To overcome the limitations of static classification, more recent research has shifted towards pathway modeling or state-evolution modeling. This paradigm frames the early warning task as monitoring the continuous progression of cardiac states: from a normal state (S_0), through a precursor state (S_1) characterized by subtle electrophysiological changes, to an eventual event state (S_2) like AF onset [15, 16]. Systems built on this concept typically combine powerful feature encoders with recurrent models (like LSTMs) to track these transitions over time. For instance, a recent work successfully utilized a Masked Autoencoder (MAE)-based 1D Vision Transformer (ViT) as a feature extractor, paired with an LSTM for modeling temporal evolution. This pipeline achieved remarkable performance in both PAF and VF early warning tasks, delivering clinically relevant lead times while maintaining high accuracy [15]. This approach optimized a combined metric (e.g., a harmonic metric, HM) that balances accuracy and lead time, demonstrating the feasibility of effective early warning systems. However, a common practice in such SSL-enhanced frameworks is that the pre-trained encoder's weights are often fixed or only lightly fine-tuned during the subsequent warning training phase. This decoupling can potentially hinder the optimal specialization of the feature extractor for the nuanced demands of precursor detection, leaving some discriminative power untapped.

2.2 Foundation Models and Self-Supervised Learning for ECG

The concept of foundation models, trained on broad data at scale and adaptable to a wide range of downstream tasks, has gained significant traction across various AI domains [17]. Self-Supervised Learning (SSL) plays a crucial role in developing these models by enabling them to learn powerful, general-purpose representations from unlabeled data, thereby addressing the data scarcity and labeling challenges prevalent in medical fields such as ECG analysis [18].

Self-Supervised Learning Strategies in ECG. Common SSL strategies applied to ECG include:

- **Generative Models**: Methods like Masked Autoencoders (MAE) learn to reconstruct masked or corrupted portions of the input signal. This forces the model to capture meaningful underlying data structures [19]. For example, MAE-based pretraining on large ECG corpora using 1D-ViT encoders has demonstrably improved downstream arrhythmia classification performance [9].
- **Contrastive Learning**: Approaches such as SimCLR [20], CPC [21], and BYOL [22] learn representations by maximizing the similarity between different augmented views of the same data sample while minimizing similarity with other samples. These methods have also been successfully applied to learn robust features from ECG data [23].

Despite their effectiveness, a common challenge for these SSL methods, especially when integrated into multi-stage pipelines, is that the pre-trained feature extractors are often treated as fixed components after initial pre-training. This limits their ability to fully adapt to the specific nuances of the downstream task during supervised fine-tuning, potentially restricting their optimal discriminative power for subtle precursor detection.

Distillation-based Foundation Models. More recently, distillation-based self-supervised learning, exemplified by models like DINO [24], DINOv2 [25], and its latest iteration, DinoV3[26], has shown immense promise. These models learn rich, dense, and highly transferable features from unlabeled data through a teacher-student knowledge distillation framework. Unlike generative or contrastive approaches, DINO-style models inherently produce robust attention maps and learn powerful representations without relying on explicit masking strategies or complex data augmentations. This paradigm offers a compelling direction for developing general-purpose encoders that can be more seamlessly integrated and adapted within end-to-end learning frameworks, such as for the complex task of ECG-based AF early warning.

3 Method

This section details our proposed end-to-end early warning framework for paroxysmal atrial fibrillation (PAF) using a DinoV3-based feature extractor. We outline the overall architecture, describe the datasets utilized for training and evaluation, explain the model's structure.

3.1 Overview of the Proposed Framework

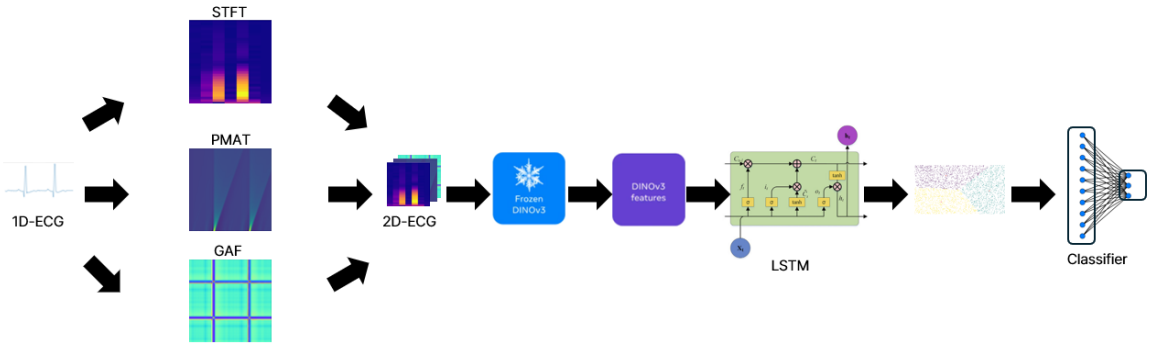

Figure 2: Overall Structure

Our goal is to evaluate whether a vision Foundation Model pre-trained on large-scale image data (DinoV3) can provide robust, transferable representations for ECG state classification without ECG-specific pretraining. The system is organized in two functional blocks for the current study: (1) **Feature extraction** via a DinoV3 backbone adapted to 2D ECG patch representations (PMAT, STFT, GAF), and (2) **Indicator (classification) head** — a temporal model (LSTM) and classifier that maps sequences of extracted features to one of three ECG states (S0: normal, S1: precursor, S2: event). While the overall architecture is designed to support future end-to-end fine-tuning and prospective prediction, here we experimentally isolate and evaluate the classification-stage performance.

3.2 Datasets

We used publicly available long-term ECG datasets that are widely used in AF research to ensure reproducibility and comparability.

- **PAF 2001 Dataset**[27]: This database was created for the Computers in Cardiology Challenge 2001 to facilitate the development of automated methods for predicting paroxysmal atrial fibrillation (PAF).It comprises a learning set and a test set, each containing 50 record sets from different subjects. Each record set includes two 30-minute, two-channel ECG recordings. Notably, for subjects with PAF, one 30-minute record immediately precedes a PAF episode, while the other is distant from any PAF activity. This structure is critical for training models to distinguish between pre-PAF states and normal rhythm. The dataset also includes corresponding 5-minute continuation records for verification and QRS annotations.
- **MIT-BIH Normal Sinus Rhythm (NSR) Database**[28]: This database consists of 18 long-term ECG recordings from subjects referred to Boston's Beth Israel Hospital, who were found to have no significant arrhythmias. It includes recordings from 5 men (aged 26 to 45) and 13 women (aged 20 to 50). This dataset provides valuable healthy control data, essential for robustly training the model to recognize and distinguish normal cardiac activity from pre-arrhythmic states.
- **IRIDIA-AF Database**[29]: This dataset comprises 167 Holter monitoring records from 152 patients with paroxysmal AF, collected from an outpatient cardiology clinic in Belgium between 2006 and 2017. Records vary in length from 19 to 95 hours, divided into 24-hour files, and are sampled at 200 Hz. All AF episodes were manually annotated and reviewed by an expert cardiologist and a specialist cardiac nurse. This extensive and expertly annotated dataset offers a rich resource for training and validating models for AF detection and prediction, particularly in real-world long-term monitoring scenarios.

These datasets were preprocessed consistently (normalization, 128Hz resampling, noise reduction filter) to produce training and evaluation sets for the indicator (classification) task.

3.3 Model Structure

Our early warning framework is designed with three main components that collaboratively extract discriminative ECG features and capture states.

- **DinoV3-based Feature Extractor**: To enhance representation learning, we replace the original MAE-based 1D Vision Transformer with DinoV3 as the backbone. Raw 1D ECG signals are converted into multi-channel 2D representations using Progressive Moving Average Transform (PMAT), Short-Time Fourier Transform (STFT), and Gramian Angular Field (GAF). These representations are fed into DinoV3, a self-supervised vision transformer pre-trained on large-scale image datasets. DinoV3 provides robust high-level semantic features and interpretable attention maps, enabling the detection of subtle electrophysiological variations related to arrhythmia onset.
- **LSTM-based Classification indicator Model**: An LSTM network processes the sequential state indicators to model the temporal evolution of ECG states. By capturing long-range dependencies, the LSTM effectively learns the progression pathway from stable rhythm (S0), through precursor states (S1), to arrhythmic events (S2). This enables accrately classification for impending arrhythmias.

This Two-state annotation is used solely to train and evaluate the classification (indicator) model in this study; moving from accurate indicator classification to reliable prospective event prediction is left to future work.

3.4 ECG State Annotation

Accurate labeling of sequential segments is critical for training the indicator model. We adopt a sliding-window scheme with a 10-second window and 5-second stride; labels are assigned based on the window endpoint relative to the annotated event onset T_{onset}.

- **Normal State (S0)**: segments whose window endpoint lies more than 20 minutes before T_{onset}. These segments are considered to represent stable, baseline rhythm.
- **Precursor State (S1)**: segments whose window endpoint falls within the 20-minute interval immediately preceding T_{onset}. These segments may contain subtle alterations that precede the AF episode.
- **Event State (S2)**: segments whose window endpoint is at or after T_{onset} and which contain the arrhythmic event.

These states are meticulously annotated to serve as ground truth for training the LSTM-based indicator model, enabling it to recognize and classify the ECG state of individual segments. In the future, the warning model utilizes these classifications to identify the critical transition pathways from S0 to S1 and ultimately to S2.

4 Experiments

This section presents the experimental setup and results for evaluating the classification performance of the DinoV3-based indicator model. Our primary goal is to demonstrate the effectiveness of using DinoV3 as the feature extractor.

4.1 Evaluation of Indicator Model Classification Performance

We compare the DinoV3-based indicator model against representative baselines, including classical deep ECG classifiers and recent SSL-enhanced pipelines adapted to ECG.

Model	Acc	Spe	Sen	F1
FCN$_{wang}$[30]	**0.6124**	**0.7567**	0.5944	0.5673
ConvNeXt[31]	0.4967	0.6667	0.3333	0.2212
ViT1D[32]	0.5325	0.7169	0.3968	0.1541
Trs[9]	0.5900	0.7561	**0.6416**	**0.6208**
DinoV3 (Proposed)	0.5950	0.7479	0.5368	0.5489

Table 1: Comparison of classification performance by changing feature extractor

Table 1 illustrates the classification performance for the paroxysmal Atrial fibrillation event (PAF). When trained for classification, the DinoV3-based indicator model achieves competive results across all metrics (Acc, Spe, Sen, MCC, F1) compared to other baselines, including traditional ECG classification models and contrastive learning models.This demonstrates that DinoV3 exhibits enhanced capabilities in extracting discriminative features for ECG state recognition even without the ECG dataset fine tuning, thereby establishing a robust foundation for the subsequent warning model. These results suggest that DinoV3's powerful feature extraction abilities can effectively contribute to detecting various ECG abnormalities.

5 Discussion and Limitations

In this study we evaluated the utility of a publicly available Foundation Model (DinoV3) — pretrained on large-scale image data and applied *without* ECG-specific pretraining — as a feature extractor for an ECG-state **classification** (indicator) model targeting paroxysmal atrial fibrillation (PAF). Our principal finding is that DinoV3-derived representations, when applied to 2D ECG patch transforms (PMAT, STFT, GAF) and coupled with a temporal classification head (LSTM), yield competitive indicator-stage performance (Table 1). This result highlights notable cross-domain transferability: visual representations learned at scale retain structure that can be exploited for biological time-series classification, at least at the upstream classification (indicator) level.

Crucially, we reiterate that the present work focuses on the **classification/indicator stage** — i.e., recognizing and distinguishing ECG segments that are normal (S0), precursor (S1), or event (S2). We did not perform prospective time-to-event prediction experiments in this manuscript. Translating accurate indicator classification into reliable early-warning or prediction requires additional modeling

choices (e.g., explicit temporal forecasting, survival analysis, calibration over time) and rigorous temporal validation; these steps are outside the scope of the current paper and are discussed below as future directions.

Despite promising indicator-stage results, there are important limitations:

- **No ECG-specific pretraining in main experiments.** The DinoV3 backbone was used primarily in its publicly released image-pretrained form. While this allowed us to probe cross-domain transferability, we expect that ECG-specific pretraining and/or systematic fine-tuning would further improve discriminative power for precursor patterns.
- **Classification vs. predictive utility.** Strong performance on S0/S1/S2 classification does not automatically imply robust prospective prediction of PAF onset. Predictive performance must be evaluated under temporal holdout schemes that respect event timing, and with metrics appropriate for forecasting (e.g., lead-time distribution, time-dependent AUROC, calibration of predicted risk).
- **Dataset diversity and generalization.** Public long-term ECG datasets with high-quality, temporally precise annotations for precursor intervals are limited. Broader external validation across multiple cohorts, devices, and recording conditions is necessary to assess generalizability and to avoid dataset-specific biases.
- **Clinical validation and interpretability.** Although DinoV3 attention maps provide patch-level saliency that may aid interpretability, these visualizations require formal evaluation by clinicians to determine clinical relevance and to avoid misleading explanations. Human-in-the-loop validation and prospective pilot studies are needed before clinical deployment.

Based on the above, we identify concrete next steps to bridge the gap between the present indicator-stage findings and clinically useful predictive systems:

1. **ECG-specific fine-tuning and ablation studies**: systematically fine-tune DinoV3 on large ECG corpora, compare frozen vs. partially fine-tuned vs. fully end-to-end settings, and report corresponding gains in indicator and (eventually) prediction metrics.
2. **Temporal/predictive modeling**: extend the pipeline to explicit time-to-event or sequence-forecasting models and evaluate under temporally strict validation (e.g., by-record or by-patient splits, prospective simulation).
3. **External and prospective validation**: test on heterogeneous external datasets and, where possible, conduct prospective pilot evaluations to measure real-world utility, false alarm rates, and clinician acceptance.
4. **Clinician-centered interpretability studies**: present attention/saliency maps and decision traces to expert cardiologists for qualitative and quantitative assessment.
5. **Robustness and calibration**: evaluate model calibration over lead time and implement uncertainty-aware decision thresholds suitable for clinical workflows.

Taken together, the current results establish a promising proof-of-concept: Foundation Models can serve as effective upstream feature extractors for the *classification* stage in AF early-warning pipelines, but additional work (fine-tuning, temporal prediction modeling, and clinical validation) is required to realize reliable prospective warning systems.

6 Conclusion

We investigated the feasibility of using a vision Foundation Model (DinoV3) as a feature extractor for an ECG-state classification (indicator) model relevant to paroxysmal atrial fibrillation (PAF). By converting ECG signals into 2D patch-based representations and coupling DinoV3 embeddings with a temporal LSTM classifier, we demonstrated competitive indicator-stage performance despite using DinoV3 without ECG-specific pretraining. These findings indicate meaningful cross-domain transferability and position Foundation Models as a viable starting point for building upstream components of AF early-warning systems.

It is important to emphasize the scope of this work: the experiments target the **classification (indicator)** stage that precedes prospective prediction. We do not claim to have developed a validated

predictive early-warning system in this manuscript. Rather, our contribution is to show that (1) Foundation Models can yield useful representations for ECG-state discrimination without domain-specific pretraining, and (2) these representations provide a strong foundation for subsequent work on fine-tuning, end-to-end optimization, and time-to-event prediction.

Future work will prioritize systematic ECG fine-tuning of the backbone, extension to explicit predictive modeling with temporally rigorous evaluation, comprehensive external validation, and clinician-in-the-loop interpretability studies. Addressing these steps is essential to translate the present proof-of-concept into clinically actionable early-warning tools that improve patient outcomes.

References

[1] World Health Organization. *Global Atlas on Cardiovascular Disease Prevention and Control*. World Health Organization, 2011. URL: https://apps.who.int/iris/handle/10665/44754.

[2] GBD 2020 CVD Collaborators. "Global, regional, and national burden of cardiovascular diseases, 1990–2020: a systematic analysis for the Global Burden of Disease Study 2020". In: *The Lancet* 397.10269 (2021), pp. 126–155. DOI: 10.1016/S0140-6736(20)32556-5.

[3] Gerhard Hindricks et al. "2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS)". In: *European Heart Journal* 42.5 (2021), pp. 373–498. DOI: 10.1093/eurheartj/ehaa612.

[4] J. Liu et al. "A review of arrhythmia detection based on electrocardiogram with artificial intelligence". In: *Expert Review of Medical Devices* 19.7 (2022), pp. 549–560. DOI: 10.1080/17434440.2022.2083184.

[5] Shenda Hong et al. *Opportunities and challenges of deep learning methods for electrocardiogram data: a systematic review*. 2020. arXiv: 2001.01550 [cs.LG]. URL: https://arxiv.org/abs/2001.01550.

[6] Kok Han Boon, Mohamad Khalil-Hani, and Tin Kee Swee. "Paroxysmal atrial fibrillation prediction based on HRV analysis and non-dominated sorting genetic algorithm III". In: *Computers in Biology and Medicine* 72 (2016), pp. 149–158. DOI: 10.1016/j.compbiomed.2016.03.001.

[7] Hsin-An Tzou, C. Chen, and H. Lin. "Paroxysmal atrial fibrillation prediction based on morphological variant P-wave analysis with wideband ECG and deep learning". In: *Computers in Biology and Medicine* 132 (2021), p. 104313. DOI: 10.1016/j.compbiomed.2021.104313.

[8] Hsin-An Tzou, H. Lin, and C. Chen. "A deep learning approach for early detection of atrial fibrillation from short-term ECG signals". In: *IEEE Journal of Biomedical and Health Informatics* 25.4 (2021), pp. 1330–1339. DOI: 10.1109/JBHI.2020.3026321.

[9] Zhiyuan Li et al. "An early warning method for arrhythmias in long-term ECGs based on self-supervised learning and LSTM". In: *Knowledge-Based Systems* (2025), p. 114137.

[10] Zhiyuan Li et al. "An early warning method for arrhythmias in long-term ECGs based on self-supervised learning and LSTM". In: *Knowledge-Based Systems* 327 (2025), p. 114137. DOI: 10.1016/j.knosys.2025.114137.

[11] Rabah Mokhtari et al. "ECG heartbeat classification using progressive moving average transform". In: *Scientific Reports* 15.1 (2025), p. 4285.

[12] Jelmer J. M. Verwijs and et al. "Limitations of current heart rate variability analysis for clinical prediction: a review". In: *Journal of the American Heart Association* 6.7 (2017), e005953.

[13] Awni Hannun and et al. "Cardiologist-level arrhythmia detection and classification in ambulatory ECG records using deep neural networks". In: *Nature Medicine* 25 (2019), pp. 65–69.

[14] J. Huang and et al. "An LSTM-based model for atrial fibrillation prediction". In: *Computers in Biology and Medicine* 99 (2018), pp. 253–261.

[15] Zhiyuan Li et al. "An early warning method for arrhythmias in long-term ECGs based on self-supervised learning and LSTM". In: *Knowledge-Based Systems* 327 (2025), p. 114137.

[16] Q. Chang and et al. “A novel deep learning model for atrial fibrillation prediction based on pathway modeling”. In: *IEEE Transactions on Biomedical Engineering* 67.12 (2020), pp. 3421–3430.

[17] Rishi Bommasani and et al. “On the opportunities and risks of foundation models”. In: *arXiv preprint arXiv:2108.07258* (2021).

[18] Jian Zeng and et al. “Self-supervised learning in electrocardiogram analysis: a systematic review”. In: *IEEE Journal of Biomedical and Health Informatics* 25.12 (2021), pp. 4453–4467.

[19] R. Cao and et al. “MAECG: Masked Autoencoders for ECG Signal Processing”. In: *arXiv preprint arXiv:2208.06822* (2022).

[20] Ting Chen and et al. “A simple framework for contrastive learning of visual representations”. In: *International Conference on Machine Learning* (2020), pp. 159–170.

[21] Aaron van den Oord, Yazhe Li, and Igor Babakhin. “Representation learning with contrastive predictive coding”. In: *arXiv preprint arXiv:1807.03748* (2018).

[22] Jean-Bastien Grill and et al. “Bootstrap your own latent: A new approach to self-supervised learning”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 21271–21284.

[23] Aditya Sharma and et al. “Contrastive learning for ECG signal analysis”. In: *IEEE Transactions on Biomedical Engineering* 68.8 (2021), pp. 2519–2528.

[24] Mathilde Caron and et al. “Emerging properties in self-supervised vision transformers”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 965–975.

[25] Maxime Oquab and et al. “DINOv2: Learning Robust Visual Features without Supervision”. In: *arXiv preprint arXiv:2304.07193* (2023).

[26] Oriane Siméoni et al. “Dinov3”. In: *arXiv preprint arXiv:2508.10104* (2025).

[27] Ary L Goldberger et al. “PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals”. In: *circulation* 101.23 (2000), e215–e220.

[28] George Moody et al. “Predicting the onset of paroxysmal atrial fibrillation: The Computers in Cardiology Challenge 2001”. In: *Computers in Cardiology 2001. Vol. 28 (Cat. No. 01CH37287)*. IEEE. 2001, pp. 113–116.

[29] Cédric Gilon et al. “IRIDIA-AF, a large paroxysmal atrial fibrillation long-term electrocardiogram monitoring database”. In: *Scientific data* 10.1 (2023), p. 714.

[30] Zhiguang Wang, Weizhong Yan, and Tim Oates. “Time series classification from scratch with deep neural networks: A strong baseline”. In: *2017 International joint conference on neural networks (IJCNN)*. IEEE. 2017, pp. 1578–1585.

[31] Zhuang Liu et al. “A convnet for the 2020s”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11976–11986.

[32] Shunxiang Yang, Cheng Lian, and Zhigang Zeng. “Masked autoencoder for ECG representation learning”. In: *2022 12th International Conference on Information Science and Technology (ICIST)*. IEEE. 2022, pp. 95–98.

Agents4Science AI Involvement Checklist

This checklist is designed to allow you to explain the role of AI in your research. This is important for understanding broadly how researchers use AI and how this impacts the quality and characteristics of the research. **Do not remove the checklist! Papers not including the checklist will be desk rejected.** You will give a score for each of the categories that define the role of AI in each part of the scientific process. The scores are as follows:

- **[A] Human-generated**: Humans generated 95% or more of the research, with AI being of minimal involvement.
- **[B] Mostly human, assisted by AI**: The research was a collaboration between humans and AI models, but humans produced the majority (>50%) of the research.
- **[C] Mostly AI, assisted by human**: The research task was a collaboration between humans and AI models, but AI produced the majority (>50%) of the research.
- **[D] AI-generated**: AI performed over 95% of the research. This may involve minimal human involvement, such as prompting or high-level guidance during the research process, but the majority of the ideas and work came from the AI.

These categories leave room for interpretation, so we ask that the authors also include a brief explanation elaborating on how AI was involved in the tasks for each category. Please keep your explanation to less than 150 words.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "Agents4Science AI Involvement Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[TODO]** [B]

   Explanation:**[TODO]** AI informed trend analysis, but the research direction and core ideas were proposed by the researcher

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[TODO]** [C]

   Explanation: **[TODO]** Researcher designs experiments; AI handles coding and execution, with researcher refining errors or logic gaps.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[TODO]** [C]

   Explanation: **[TODO]** Researcher designs experiments; AI handles coding and execution, with researcher refining errors or logic gaps.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[TODO]** [D]

   Explanation: **[TODO]** AI drafted the overall framework; I manually revised the logical flow and experimental-result interpretations to correct overinterpretation.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: **[TODO]** AI often risks overinterpretation in experiments and lacks creativity in topic selection, yet proves useful when merging diverse fields.

Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: **[TODO]**That's generally correct.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: **[TODO]**The discussion mentions future directions but does not address limitations related to AI itself, such as computational efficiency.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: **[TODO]**It does not contain any theoretical assumptions. Therefore, the answer is marked as NA.

   Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: **[TODO]**This is entirely feasible simply because we used publicly available models, data and weights.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: **[TODO]**This is entirely feasible simply because we used publicly available models, data and weights.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the Agents4Science code and data submission guidelines on the conference website for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: **[TODO]**The training procedure follows the same details as in Reference 9, making it fully implementable.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: **[TODO]**This paper focuses solely on the results of the Foundation Model and does not discuss those details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: **[TODO]**Reference 9 occupies about 1–2 GB of VRAM, and since Dino's model size is adjustable, we deemed it unnecessary to specify.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

   Answer: [Yes]

   Justification: **[TODO]**Yes

   Guidelines:

   - The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: **[TODO]**The paper only presents an optimistic outlook, suggesting that a future prediction model could be developed and applied meaningfully in clinical practice.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.