Prune or Quantize? Layer-wise Compression of Time-Series ECG Foundation Networks

Tushar Shinde[1,*], Sudhanshu Gaurhar[2,*], Anil Kumar Tiwari[2]
[1]Indian Institute of Technology Madras, Zanzibar, Tanzania
[2]Indian Institute of Technology Jodhpur, Rajasthan, India
shinde@iitmz.ac.in, sudhanshu.1@iitj.ac.in, akt@iitj.ac.in
[*]Equal contribution

Abstract

Foundation models for biosignals, such as wearable ECG monitors, face challenges in resource-constrained settings due to high memory and computational demands. We propose an adaptive layer-wise compression framework that combines quantization and pruning to reduce model size while preserving predictive performance. Layer importance, estimated via parameter contribution and weight variance, guides fine-grained assignment of bit-widths and pruning thresholds, balancing efficiency and accuracy across high- and low-sensitivity layers. Experiments on Chapman and CPSC ECG datasets show that our method consistently outperforms fixed global compression schemes, achieving up to 12.10× compression with no loss in performance. Our architecture-agnostic framework scales from lightweight residual networks to large foundation models, enabling real-time, low-resource ECG monitoring. By efficiently deploying foundation models on edge devices, this work advances scalable, physiology-aware biosignal AI for mobile health and clinical applications.

1 Introduction and Related Work

Electrocardiography (ECG) is a cornerstone of cardiac health assessment, capturing the heart's electrical activity through body-surface electrodes to reveal characteristic waveforms Trobec et al. [2018]. These signals enable detection of arrhythmias, ranging from asymptomatic to life-threatening conditions like sudden cardiac death Srinivasan and Schilling [2018]. Traditional rule-based diagnostics struggle with the scale and complexity of physiological data, driving demand for automated, cost-effective ECG monitoring Ebrahimi et al. [2020].

Deep learning (DL) has transformed arrhythmia detection, with convolutional neural networks (CNNs) and recurrent architectures achieving high accuracy Kiranyaz et al. [2015], Alzubaidi et al. [2021]. Recent innovations include transforming ECGs into images de Santana et al. [2021], CNN-LSTM hybrids Tan et al. [2018], and transformer-based models El-Ghaish and Eldele [2024], Jin et al. [2021]. However, these approaches often lack generalization across diverse populations or robustness to class imbalance. Foundation models, pretrained on large-scale unlabeled data, have revolutionized NLP, vision, and audio by enabling robust generalization across tasks and domains Radford et al. [2018], He et al. [2022], Hsu et al. [2021]. In medicine, models like CheXzero Tiu et al. [2022], MedSAM Ma et al. [2024], and ECGFounder Li et al. [2024] leverage large-scale biosignal data for improved transferability. However, their computational complexity and reliance on supervised pretraining with limited cohorts hinder deployment in resource-constrained settings, such as wearable ECG devices.

Deploying DL models on wearables is limited by memory, energy, and latency constraints Chen and Ran [2019]. Compression techniques like pruning Frankle and Carbin [2018], quantization

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from Time Series for Health.

Hubara et al. [2018], Krishnamoorthi [2018], and binarization Courbariaux et al. [2015] enable lightweight deployment. Quantization reduces parameter precision, acting as a regularizer that preserves discriminative capacity in noisy biosignals Liu et al. [2022b]. Recent advances, such as nonuniform-to-uniform quantization (N2UQ) Liu et al. [2022b], adapt bin widths to data distributions, achieving near full-precision accuracy. Knowledge distillation Hinton et al. [2015] and neural architecture search Tan [2019] further optimize model efficiency, but their application to biosignal foundation models remains underexplored Shinde et al., Gaurhar et al..

Our contributions are:

- The first adaptive compression framework for ECG foundation models, enabling edge deployment with up to $10\times$ size reduction for both traditional and foundational models.
- A ResNet1D achieving state-of-the-art arrhythmia classification with high compression.
- A demonstration that compressed foundation models maintain clinical accuracy while reducing computational costs by an order of magnitude, advancing scalable biosignal AI.

2 Method

We aim to design scalable, interpretable biosignal foundation models that balance physiological fidelity with edge deployment efficiency. Our framework integrates morphology-aware convolutional models with self-supervised transformers, enhanced by adaptive compression to address heterogeneous biosignals.

ResNet1D Architecture. The ResNet1D processes ECG signals $\mathbf{X} \in \mathbb{R}^{C\times L}$, where C is the number of leads and L is the sequence length. The initial convolution maps $\mathbf{X}$ to a feature space: $\mathbf{H}_0 = \mathrm{BN}(\mathrm{Conv1d}(\mathbf{X}; W_0))$. Each residual block applies:

$$\mathbf{Z}_k = \sigma(\mathrm{BN}(\mathrm{Conv1d}(\mathbf{H}_{k-1}; W_{k,1}))), \quad \mathbf{Z}'_k = \sigma(\mathrm{BN}(\mathrm{Conv1d}(\mathbf{Z}_k; W_{k,2}))), \tag{1}$$

with output $\mathbf{H}_k = \mathbf{Z}'_k + \mathcal{S}(\mathbf{H}_{k-1})$. The final output is $\hat{\mathbf{y}} = \mathrm{Softmax}(W_f\,\mathrm{vec}(\mathbf{H}_K) + b_f)$. ResNet1D captures local ECG morphology (e.g., QRS complexes), complementing the global temporal modeling of foundation models.

ECG-HuBERT Architecture. The HuBERT-ECG model Coppola et al. [2024], pretrained on large-scale unlabeled ECG data, extracts Mel-spectrogram features: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T], \quad \mathbf{x}_t \in \mathbb{R}^F$. Clustering assigns pseudo-labels $c_t = \arg\min_k \|\mathbf{x}_t - \mu_k\|_2^2$. Masked frames are embedded by a convolutional encoder f_{conv}, producing $\mathbf{z}_t$, contextualized by a Transformer: $\mathbf{h}_t = \mathcal{T}(\mathbf{z}_t + \mathbf{p}_t)$. The loss is:

$$\mathcal{L}_{\text{HuBERT}} = -\frac{1}{\sum_t m_t} \sum_{t=1}^{T} m_t \log p_\theta(c_t|\mathbf{h}_t). \tag{2}$$

This self-supervised pretraining enables robust, generalizable representations for ECG tasks, embodying biosignal foundation model principles.

Adaptive Model Compression. The layer-wise adaptive compression capitalizes on the heterogeneous sensitivity of network layers: more critical layers are pruned and quantized conservatively, whereas less important layers undergo more aggressive compression Shinde [2024], Shinde and Tukaram Naik [2024], Shinde [2025].

Let a neural network $\mathcal{M}$ have L layers, each with a weight tensor W_l. The goal is to determine the bit-width b_l and pruning factor p_l for each layer, minimizing model size while ensuring minimal accuracy loss: $\min_{\{b_l,p_l\}_{l=1}^L} \text{Size}(\mathcal{M}) \quad \text{s.t.} \quad \text{Accuracy}(\mathcal{M}_{\text{comp}}) \geq A_0 - \Delta$, where $\mathcal{M}_{\text{comp}}$ is the compressed model, A_0 is the baseline accuracy, and Δ is the allowable accuracy degradation.

Layer Importance Estimation. To guide the compression process, we assign an importance score to each layer based on two factors: **Parameter Density Index (PDI)** reflects the proportion of parameters in layer l relative to the total parameters: $PDI_l = \dim(W_l)/\sum_{k=1}^{L} \dim(W_k)$. **Parameter Variability Index (PVI)** captures the variability of weights within a layer, which influences its sensitivity to quantization. Specifically, the PVI (PVI_l) is computed based on the variance of the weights in each layer, normalized relative to the maximum variance across all layers. This helps in assessing how much a layer's weight distribution varies, affecting its ability to maintain accuracy after quantization. **Kurtosis:** Kurtosis captures the presence of outliers in the weight or activation

Table 1: Comparison of the performance and compression ratio (CR) of the proposed ResNet1D and HuBERT-ECG models on the Chapman dataset under different quantization and pruning settings. The best, second-best, and third-best values in each column are highlighted. The optimal results are achieved for $\alpha = 0.1$, $\beta = 0.1$, and $\gamma = 0.8$ under both LAQ and LAP

Method	Proposed ResNet1D Model					HuBERT ECG Model				
	Acc	Prec	Rec	F1	CR	Acc	Prec	Rec	F1	CR
Full-Precision	0.9542	0.9483	0.9467	0.9475	1.00x	0.9674	0.9636	0.9636	0.9636	1.00x
Quantized (8-bit)	0.9542	0.9483	0.9467	0.9475	3.99x	**0.9684**	**0.9646**	**0.9646**	**0.9646**	3.98x
Quantized (7-bit)	0.9547	0.9489	0.9473	0.9481	4.56x	0.9669	0.9630	0.9630	0.9630	4.55x
Quantized (6-bit)	0.9542	0.9482	0.9470	0.9475	5.31x	0.9665	0.9623	0.9623	0.9623	5.30x
Quantized (5-bit)	0.9537	0.9477	0.9467	0.9472	6.37x	0.9655	0.9614	0.9614	0.9614	6.35x
Quantized (4-bit)	0.9542	*0.9490*	*0.9480*	*0.9485*	7.95x	0.9608	0.9561	0.9561	0.9561	7.91x
Quantized (3-bit)	0.9108	0.8962	0.8920	0.8941	10.58x	0.6577	0.6526	0.6526	0.6526	*10.50*x
Quantized (2-bit)	0.2158	0.0887	0.1310	0.1310	15.79x	0.2101	0.0868	0.0868	0.0868	15.63x
Quantized (1-bit)	0.2191	0.0899	0.1310	0.1310	**31.16x**	0.2101	0.0868	0.0868	0.0868	**30.49x**
Pruned (10%)	0.9481	0.9413	0.9399	0.9406	1.11x	*0.9679*	*0.9640*	*0.9640*	*0.9640*	1.11x
Pruned (30%)	0.7002	0.7029	0.7001	0.7015	1.43x	0.8966	0.8851	0.8851	0.8851	1.43x
Pruned (50%)	0.2427	0.1412	0.1412	0.1412	2.00x	0.3343	0.2700	0.2700	0.2700	2.00x
Pruned (70%)	0.2158	0.0887	0.1310	0.1310	3.33x	0.2191	0.0899	0.0899	0.0899	3.32x
Pruned (90%)	0.3551	0.1310	0.1310	0.1310	9.92x	0.2384	0.1514	0.1514	0.1514	9.86x
Proposed LAQ	**0.9551**	**0.9499**	**0.9488**	**0.9494**	*12.10*x	0.9674	0.9639	0.9633	0.9636	9.80x
Proposed LAP	0.9547	0.9497	0.9487	0.9492	1.67x	**0.9684**	**0.9644**	0.9644	0.9644	2.41x

distributions; accordingly, the layer-wise kurtosis importance index is defined as $\mathcal{K}I_l = \frac{\mathrm{Kur}(W_l)}{\max_k \mathrm{Kur}(W_k)}$, which quantifies the relative sensitivity of each layer to quantization errors.

Combined Importance. The final importance score for layer l is a weighted sum of the PDI, PVI, and KI: $I_l = \alpha \cdot PDI_l + \beta \cdot PVI_l + \gamma \cdot KI_l$, where α, β, and γ are hyperparameters controlling the emphasis on parameter density and sensitivity to quantization.

Quantization. Quantization reduces the model's memory and computational requirements by converting continuous weights to discrete levels. In *Fixed Quantization*, all weights are uniformly quantized to a global bit-width b_{fixed}: $\hat{w} = \text{Quantize}(w, b_{\text{fixed}})$. *Layer-wise Adaptive Quantization (LAQ).* Bit-widths are assigned to layers based on importance scores. For layer l, the optimal bit-width b_l^* is selected by:

$$b_l^* = \min\{b_l : \text{Accuracy}(\mathcal{M}_{\text{quant}}) \geq A_0 - \gamma \cdot I_l\}, \tag{3}$$

where γ is a global accuracy tolerance, and b_l is selected greedily to minimize accuracy degradation.

Pruning. Pruning removes low-magnitude weights, inducing sparsity and reducing model size. In *Fixed Pruning*, a uniform pruning factor p_{fixed} is applied across the network. *Layer-wise Adaptive Pruning (LAP).* Pruning is performed based on layer-specific importance, with a pruning factor p_l for each layer. For each weight:

$$\hat{W}_{l,ij} = \{ \ W_{l,ij}, \quad \text{if } |W_{l,ij}| > \theta_l, \quad 0, \quad \text{otherwise.} \tag{4}$$

The pruning threshold θ_l is computed as the p_l-th percentile of $|W_l|$, where layers with higher importance are pruned conservatively. This adaptive, importance-guided compression achieves a trade-off between model size and performance, enabling efficient deployment.

3 Experimental Setup

All experiments were conducted on the Kaggle platform equipped with an NVIDIA Tesla P100 GPU, leveraging PyTorch for deep learning operations.

Datasets: *CPSC 2018 Challenge Dataset.* Contains 6,877 twelve-lead ECG recordings (6–60 seconds, 500 Hz) with nine rhythm categories Liu et al. [2018]. *Chapman Clinical Dataset.* Includes $\sim$10,000 subjects with 10-second, twelve-lead ECGs (500 Hz), aggregated into four rhythm groups Zheng et al. [2020], Murat et al. [2021]. CPSC and Chapman enable rigorous evaluation of benchmark performance and cross-domain generalization under class imbalance and noise.

Model Architectures and Training. We evaluate: (1) ResNet1D and (2) HuBERT-ECG models. Inputs are zero-padded to 5,120 samples. Training uses Adam (lr=10^{-3} for ResNet1D, 10^{-4} for HuBERT, following Kiranyaz et al. [2015]), weight decay 10^{-3}, dropout (0.2–0.3), and `ReduceLROnPlateau`. Categorical cross-entropy loss ensures robust rhythm classification.

4 Results and Discussion

We conduct an ablation study to evaluate the effects of quantization and pruning on the proposed ResNet1D and HuBERT-ECG models. Table 1 reports classification metrics (Accuracy, Precision, Recall, F1) and compression ratio (CR) across various settings.

Fixed Quantization. Uniform quantization (8-bit to 1-bit) reveals distinct sensitivities. For ResNet1D, 4-bit quantization achieves the highest accuracy (0.9542) and F1 (0.9485) with 7.95× compression, likely due to quantization noise acting as implicit regularization. Analysis of weight distributions shows reduced variance in feature activations, mitigating overfitting on ECG waveforms. Performance collapses at $\leq$3 bits due to excessive information loss. In contrast, HuBERT-ECG is more sensitive to low-precision quantization, as its self-attention layers require fine-grained weight resolution to capture global temporal dependencies Vaswani et al. [2017]. While 8–7 bits maintain accuracy $\approx$0.96, performance drops sharply at 4-bit (0.9561), unlike the robust ResNet1D.

Layer-wise Adaptive Quantization (LAQ). Our LAQ strategy achieves near-baseline accuracy with high compression. By allocating precision based on layer importance, LAQ preserves critical layers (e.g., convolutional filters capturing QRS complexes) while aggressively compressing redundant ones, optimizing for noisy biosignals. ResNet1D reaches 0.9551 accuracy at 12.10× CR, while HuBERT-ECG attains 0.9674 at 9.80×, consistently outperforming fixed quantization schemes.

Fixed Pruning. Light pruning (10%) has minimal impact, but pruning beyond 30% rapidly degrades accuracy by disrupting structural connectivity essential for ECG pattern recognition. ResNet1D accuracy falls to 0.2158 at 50% pruning, and HuBERT-ECG shows similar degradation, confirming that indiscriminate weight removal is detrimental.

Layer-wise Adaptive Pruning (LAP). LAP offers stable trade-offs. For ResNet1D, LAP retains 0.9547 accuracy at 1.67× CR. HuBERT-ECG achieves 0.9684 accuracy with 2.41× CR, surpassing fixed pruning. By prioritizing layers with high importance (e.g., self-attention heads in HuBERT), LAP preserves expressive capacity, crucial for generalizable biosignal modelling.

Comparison with Existing Work. On the Chapman dataset, ResNet1D+LAQ achieves 0.9551 accuracy with 12.10× compression, and HuBERT-ECG+LAQ reaches 0.9674 with 9.80× (see Table 2). Unlike prior methods optimizing solely for accuracy Murat et al. [2021], our approach sets a new state-of-the-art by balancing clinical fidelity and edge deployability. On CPSC 2018, ResNet1D achieves 95.78% accuracy with 10× compression, outperforming baselines Dhyani et al. [2023]. These results provide the first evidence of compact ECG models achieving superior performance while enabling real-time deployment on resource-constrained devices. To assess generalization, a key property of biosignal foundation models, we compare performance across CPSC and Chapman datasets. HuBERT-ECG's self-supervised pretraining enhances robustness to Chapman's class imbalance, achieving 0.9674 accuracy despite fewer training samples. ResNet1D excels on CPSC (95.78%) due to its focus on local morphology but shows slightly lower generalization on Chapman's heterogeneous clinical data. These findings underscore the complementary strengths of convolutional and transformer-based foundation models for biosignals.

Discussion. Adaptive, layer-aware compression (LAQ/LAP) achieves Pareto-optimal trade-offs between accuracy and efficiency, enabling real-time ECG monitoring on edge devices. ResNet1D's robustness to quantization and pruning makes it ideal for lightweight applications, while HuBERT-ECG benefits from adaptive strategies to preserve self-supervised features. The framework's ability to generalize across datasets and handle noisy biosignals aligns with the scalability and robustness goals of foundation models, advancing clinical deployment of AI-driven health monitoring.

5 Conclusion

We present an adaptive compression framework for biosignal foundation models, enabling efficient ECG monitoring on edge devices with up to 12.10× compression without any loss. Layer importance guides conservative compression of critical layers and aggressive optimization of redundant ones. The framework's architecture-agnostic design generalizes across datasets and modalities, supporting real-time health monitoring. Future work will explore multimodal biosignal integration (e.g., EEG, EMG), dynamic inference, and ethical considerations for clinical adoption, enhancing the framework's impact on scalable, reliable biosignal AI.

Table 2: Classification performance comparisons on Chapman and CPSC 2018 datasets. Best, second-best, and third-best values per column are highlighted.

Dataset	Author	Classes	#Lead	Method	Acc.	Prec.	Rec.	F1	CR
Chapman	Yildirim et al. [2020]	4	12	Deep neural network	96.13	95.78	95.43	95.57	–
	Baygin et al. [2021]	4	1	HIT pattern SVM	97.18	97.07	96.77	96.91	–
	Murat et al. [2021]	4	1	DNN + feature fusion	**98.00**	**97.76**	**97.70**	**97.72**	–
	Domazetoski et al. [2022]	3	12	XGBoost	89.37	–	–	–	–
	Venkatesh et al. [2024]	5	1	1D-CNN-BiLSTM	93.97	93.96	98.49	93.95	–
	ResNet1D + LAQ	4	1	Residual Network	95.51	94.99	94.88	94.94	**12.10x**
	HuBERT ECG + LAQ	4	1	Foundational Network	*96.74*	*96.39*	*96.33*	*96.36*	9.80x
CPSC 2018	Zhang et al. [2020]	9	12	CNN+Attention+BiGRU	86.83	84.18	82.93	83.51	–
	Ge et al. [2021]	9	1	SEBlock(CNN)	–	83.00	82.70	82.80	–
	Liu et al. [2022a]	9	12	CRT-Net	87.20	87.30	*87.20*	86.90	–
	Li and Zhang [2023]	9	12	KNN+CNN	*88.50*	*87.77*	87.08	*87.37*	–
	Dhyani et al. [2023]	9	12	ResNet+RNN	93.29	93.38	93.10	93.09	–
	Ji et al. [2024]	9	12	Multi-scale grid transformer	87.34	85.67	86.21	85.90	–
	Proposed ResNet1D	9	1	Residual Network	**95.78**	**95.61**	**95.81**	**95.68**	**10.44x**

References

Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8(1):53, 2021.

Mehmet Baygin, Turker Tuncer, Sengul Dogan, Ru-San Tan, and U Rajendra Acharya. Automated arrhythmia detection with homeomorphically irreducible tree technique using more than 10,000 individual subject ecg records. *Information Sciences*, 575:323–337, 2021.

Jiasi Chen and Xukan Ran. Deep learning with edge computing: A review. *Proceedings of the IEEE*, 107(8): 1655–1674, 2019.

Edoardo Coppola, Mattia Savardi, Mauro Massussi, Marianna Adamo, Marco Metra, and Alberto Signoroni. Hubert-ecg as a self-supervised foundation model for broad and scalable cardiac applications. *medRxiv*, pages 2024–11, 2024.

Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. *Advances in neural information processing systems*, 28, 2015.

JR G de Santana, Marly GF Costa, and Cicero Ferreira Fernandes Costa Filho. A new approach to classify cardiac arrythmias using 2d convolutional neural networks. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 566–570. IEEE, 2021.

Shikha Dhyani, Adesh Kumar, and Sushabhan Choudhury. Arrhythmia disease classification utilizing resrnn. *Biomedical Signal Processing and Control*, 79:104160, 2023.

Viktor Domazetoski, Goran Gligoric, Milan Marinkovic, Alexei Shvilkin, Jelena Krsic, Ljupco Kocarev, and Marija D Ivanovic. The influence of atrial flutter in automated detection of atrial arrhythmias-are we ready to go into clinical practice?". *Computer Methods and Programs in Biomedicine*, 221:106901, 2022.

Zahra Ebrahimi, Mohammad Loni, Masoud Daneshtalab, and Arash Gharehbaghi. A review on deep learning methods for ecg arrhythmia classification. *Expert Systems with Applications: X*, 7:100033, 2020.

Hany El-Ghaish and Emadeldeen Eldele. Ecgtransform: Empowering adaptive ecg arrhythmia classification framework with bidirectional transformer. *Biomedical Signal Processing and Control*, 89:105714, 2024.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

Sudhanshu Gaurhar, Tushar Shinde, and Anil Kumar Tiwari. Resource-efficient ecg foundation networks via layer-wise adaptive compression. In *NeurIPS 2025 Workshop on Foundation Models for the Brain and Body*.

Rongjun Ge, Tengfei Shen, Ying Zhou, Chengyu Liu, Libo Zhang, Benqiang Yang, Ying Yan, Jean-Louis Coatrieux, and Yang Chen. Convolutional squeeze-and-excitation network for ecg arrhythmia detection. *Artificial Intelligence in Medicine*, 121:102181, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *journal of machine learning research*, 18(187):1–30, 2018.

Changqing Ji, Liyong Wang, Jing Qin, Lu Liu, Yue Han, and Zumin Wang. Msgformer: A multi-scale grid transformer network for 12-lead ecg arrhythmia detection. *Biomedical Signal Processing and Control*, 87: 105499, 2024.

Yanrui Jin, Jinlei Liu, Yunqing Liu, Chengjin Qin, Zhiyuan Li, Dengyu Xiao, Liqun Zhao, and Chengliang Liu. A novel interpretable method based on dual-level attentional deep neural network for actual multilabel arrhythmia detection. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2021.

Serkan Kiranyaz, Turker Ince, and Moncef Gabbouj. Real-time patient-specific ecg classification by 1-d convolutional neural networks. *IEEE transactions on biomedical engineering*, 63(3):664–675, 2015.

Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018.

Jun Li, Aaron Aguirre, Junior Moura, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains. *arXiv preprint arXiv:2410.04133*, 2024.

Zicong Li and Henggui Zhang. Fusing deep metric learning with knn for 12-lead multi-labelled ecg classification. *Biomedical Signal Processing and Control*, 85:104849, 2023.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

Jingyi Liu, Zhongyu Li, Xiayue Fan, Xuemeng Hu, Jintao Yan, Bolin Li, Qing Xia, Jihua Zhu, and Yue Wu. Crt-net: A generalized and scalable framework for the computer-aided diagnosis of electrocardiogram signals. *Applied Soft Computing*, 128:109481, 2022a.

Zechun Liu, Kwang-Ting Cheng, Dong Huang, Eric P Xing, and Zhiqiang Shen. Nonuniform-to-uniform quantization: Towards accurate quantization via generalized straight-through estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4942–4952, 2022b.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Fatma Murat, Ozal Yildirim, Muhammed Talo, Yakup Demir, Ru-San Tan, Edward J Ciaccio, and U Rajendra Acharya. Exploring deep features and ecg attributes to detect cardiac rhythm classes. *Knowledge-Based Systems*, 232:107473, 2021.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Tushar Shinde. Adaptive quantization and pruning of deep neural networks via layer importance estimation. In *Workshop on Machine Learning and Compression, NeurIPS 2024*, 2024.

Tushar Shinde. Towards optimal layer ordering for efficient model compression via pruning and quantization. In *2025 25th International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2025.

Tushar Shinde and Sukanya Tukaram Naik. Adaptive quantization of deep neural networks via layer importance estimation. In *International Conference on Computer Vision and Image Processing*, pages 220–233. Springer, 2024.

Tushar Shinde, Sudhanshu Gaurhar, and Anil Kumar Tiwari. qhubert: Quantized model for ecg classification. In *Recent Advances in Time Series Foundation Models Have We Reached the'BERT Moment'?*

Neil T Srinivasan and Richard J Schilling. Sudden cardiac death and arrhythmias. *Arrhythmia & electrophysiology review*, 7(2):111, 2018.

Jen Hong Tan, Yuki Hagiwara, Winnie Pang, Ivy Lim, Shu Lih Oh, Muhammad Adam, Ru San Tan, Ming Chen, and U Rajendra Acharya. Application of stacked convolutional and long short-term memory network for accurate identification of cad ecg signals. *Computers in biology and medicine*, 94:19–26, 2018.

Mingxing Tan. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, pages 6105–6114, 2019.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022.

Roman Trobec, Ivan Tomašić, Aleksandra Rashkovska, Matjaž Depolli, and Viktor Avbelj. *Body sensors and electrocardiography*. Springer, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

N Prasanna Venkatesh, R Pradeep Kumar, Bala Chakravarthy Neelapu, Kunal Pal, and J Sivaraman. Automated atrial arrhythmia classification using 1d-cnn-bilstm: A deep network ensemble model. *Biomedical Signal Processing and Control*, 97:106703, 2024.

Ozal Yildirim, Muhammed Talo, Edward J Ciaccio, Ru San Tan, and U Rajendra Acharya. Accurate deep neural network model to detect cardiac arrhythmia on more than 10,000 individual subject ecg records. *Computer methods and programs in biomedicine*, 197:105740, 2020.

Jing Zhang, Aiping Liu, Min Gao, Xiang Chen, Xu Zhang, and Xun Chen. Ecg-based multi-class arrhythmia detection using spatio-temporal attention-based convolutional recurrent neural network. *Artificial Intelligence in Medicine*, 106:101856, 2020.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific data*, 7(1): 48, 2020.