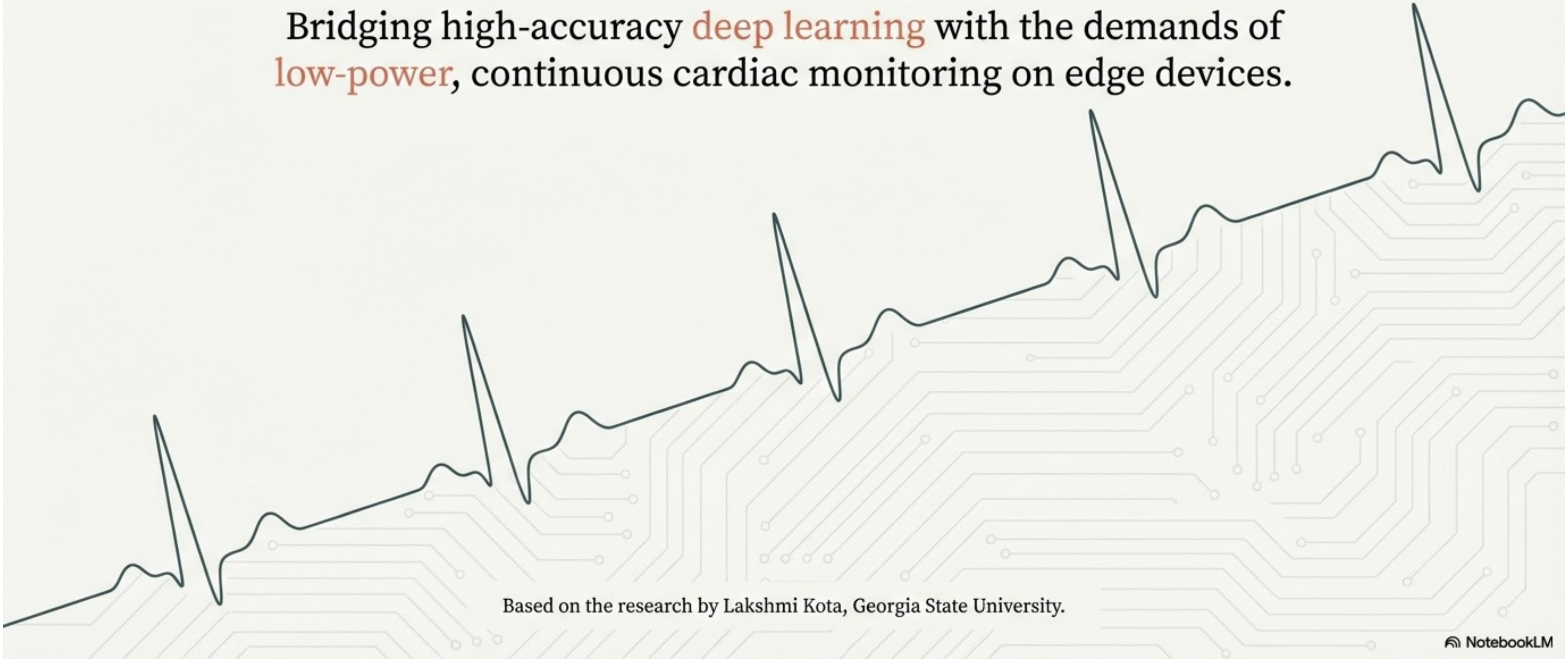
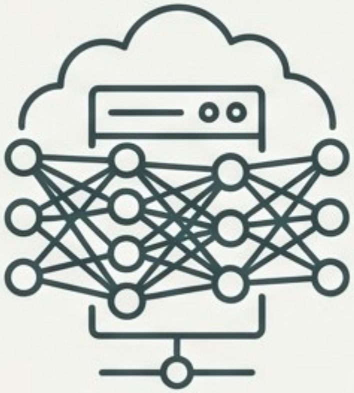
Energy-Efficient AI for Wearable Health: A Compression Framework for ECG Arrhythmia Detection

Bridging high-accuracy deep learning with the demands of low-power, continuous cardiac monitoring on edge devices.



Based on the research by Lakshmi Kota, Georgia State University.

The Core Challenge: High-Performance AI is Too Demanding for Wearable Devices

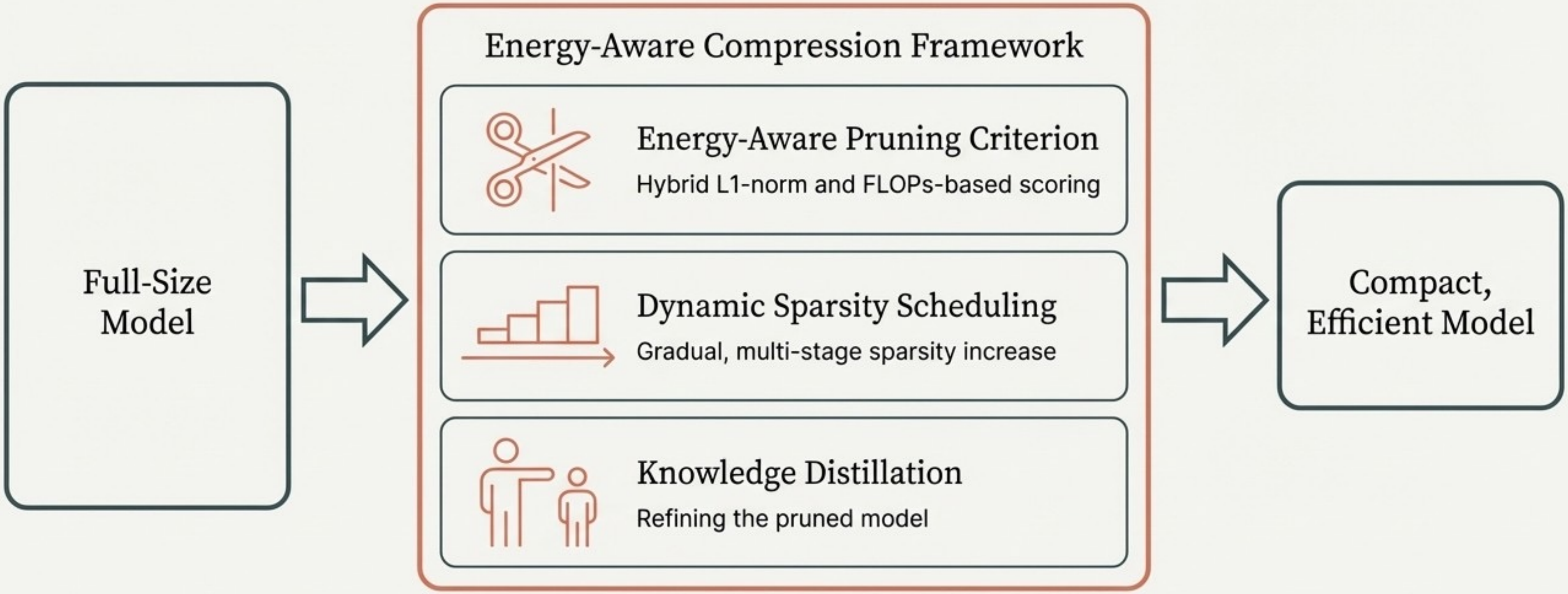


State-of-the-art deep learning models like ResNet1D and Transformers excel at ECG diagnosis but are computationally expensive. Their high energy and memory requirements are a major barrier to deployment on battery-powered, resource-constrained wearable and edge devices.

- The Need: Continuous, real-time cardiac surveillance is critical for patient health.
- The Problem: High computational load directly impacts battery life, inference latency, and system cost.
- The Goal: Reduce computational load and inference energy without compromising the clinical reliability and diagnostic accuracy of the models.

Our Solution: An Energy-Aware Framework for Model Compression

We developed a compression framework that systematically reduces model complexity by explicitly targeting computational energy, not just parameter count.



Full-Size Model

Energy-Aware Compression Framework

Energy-Aware Pruning Criterion

Hybrid L1-norm and FLOPs-based scoring

Dynamic Sparsity Scheduling

Gradual, multi-stage sparsity increase

Knowledge Distillation

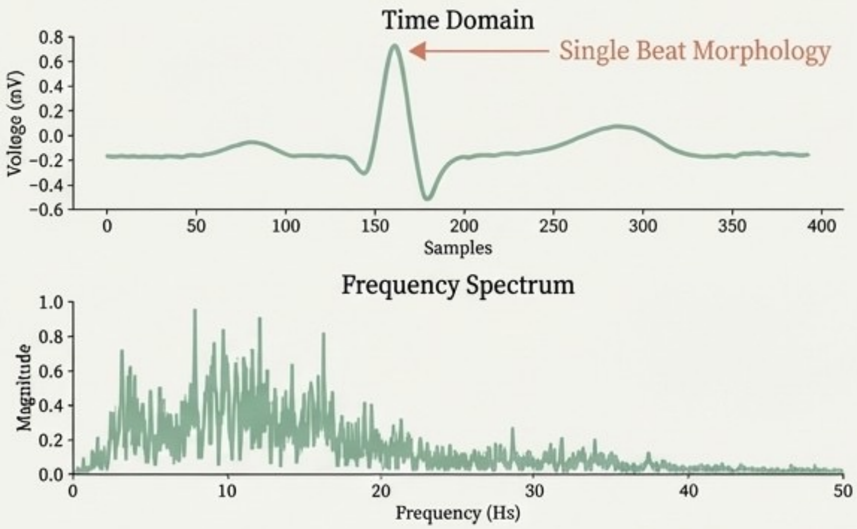
Refining the pruned model

Compact, Efficient Model

Selecting the Right Battlegrounds: Beat-Level vs. Record-Level ECG Analysis

Rationale: We used two complementary datasets to evaluate the framework's performance on different temporal granularities of ECG data, ensuring the solution is robust and generalizable.

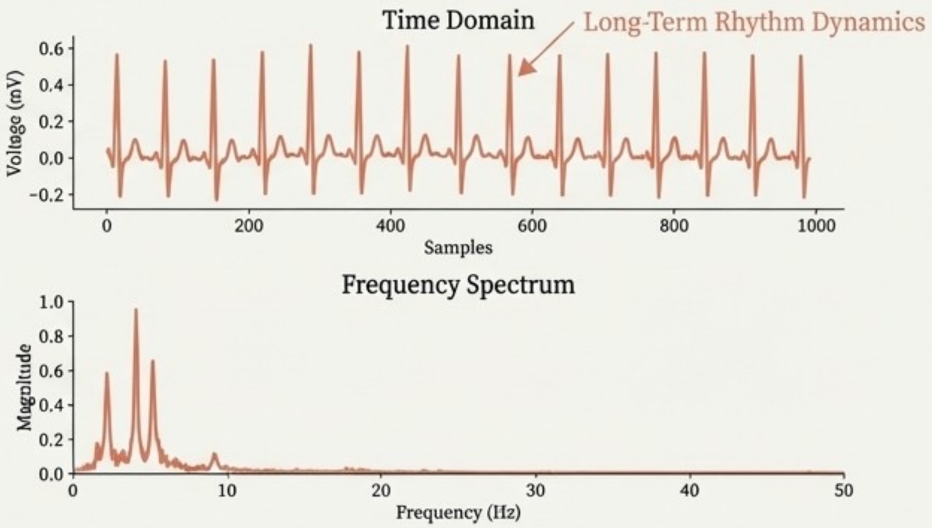
Dataset 1: MIT-BIH Arrhythmia (Beat-Level)



Single Beat Morphology

Focus: Captures local morphological features of individual heartbeats.
Task: Single-label classification (N, S, V).
Signal: Single-lead, 256-sample windows.

Dataset 2: PTB-XL (Record-Level)



Long-Term Rhythm Dynamics

Focus: Represents long-term cardiac dynamics and inter-lead correlations.
Task: Multi-label diagnosis (NORM, MI, STTC, etc.).
Signal: 12-lead, 10-second recordings.

NotebookLM

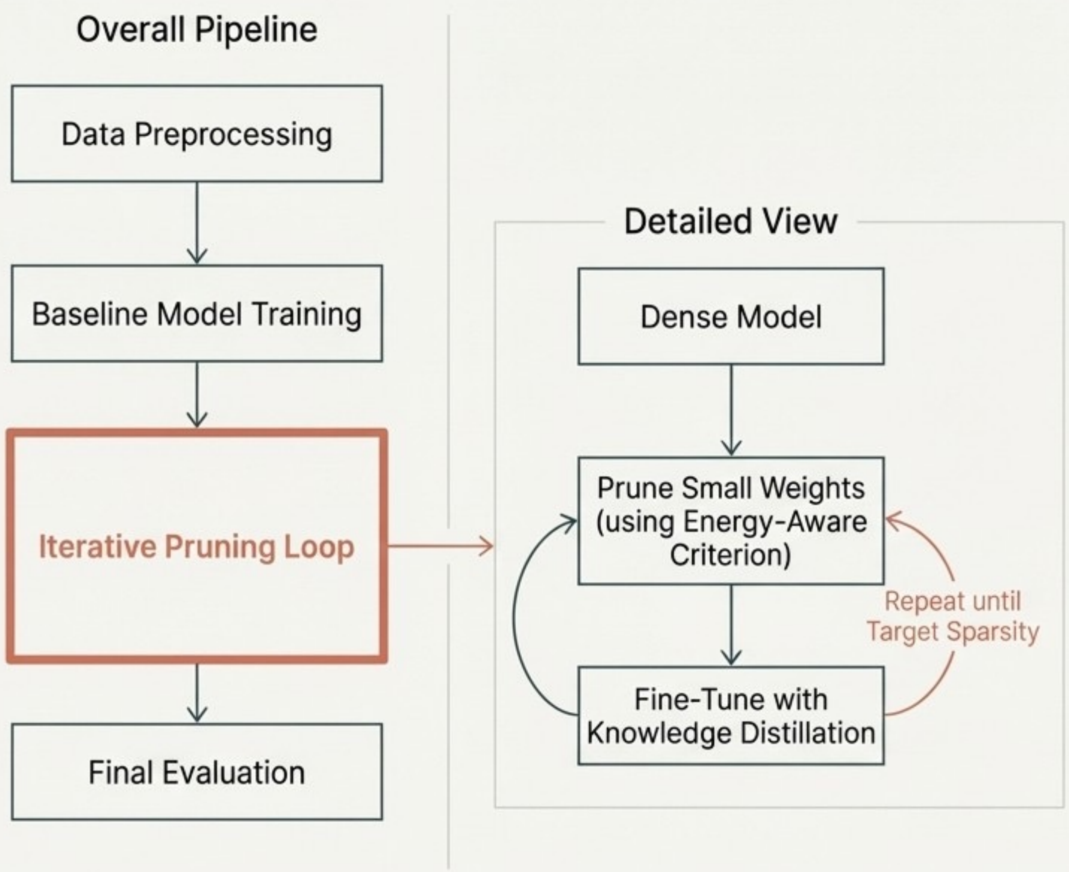
Assembling a Diverse Portfolio of Architectures for a Robust Benchmark

We selected six distinct architectures to evaluate our framework's effectiveness across different model families, from lightweight CNNs to complex Transformers. This ensures our findings are not specific to a single model type.

Model	Total Params	Suited For	Key Characteristic
MIT-BIH Models		**MIT-BIH Models**	
CNN1D 3Class	188,035	MIT-BIH	Lightweight, captures localized morphological features.
CNN LSTM Attn	349,860	MIT-BIH	Hybrid spatial-temporal modeling.
BiLSTM ECG	530,947	MIT-BIH	Captures long-term dependencies across cardiac cycles.
ResNet1D Attn	417,892	MIT-BIH	Multi-scale feature aggregation via residual connections.
PTB-XL Models		**PTB-XL Models**	
ResNet1D SE Attn	1,661,126	PTB-XL	Adaptive channel weighting for multi-lead signals.
ConvTransformer	1,711,878	PTB-XL	Integrates CNN feature extraction with global self-attention.

Key Insight: The choice of model is not arbitrary. MIT-BIH models focus on morphology and intra-beat patterns, while PTB-XL models must handle multi-lead correlations and long-range dependencies. Direct transfer between datasets is not feasible due to these structural differences.

The Compression Engine: A Multi-Stage Pruning and Refinement Pipeline



Overall Pipeline

Data Preprocessing

Baseline Model Training

Iterative Pruning Loop

Final Evaluation

Detailed View

Dense Model

Prune Small Weights (using Energy-Aware Criterion)

Fine-Tune with Knowledge Distillation

Repeat until Target Sparsity

1. Energy-Aware Importance Score

$$I_{l,i} = \alpha|w_{l,i}| + (1-\alpha)\frac{E_l}{\max(E)}$$

Balances a weight's magnitude $|w|$ with its layer's energy cost E' , allowing the framework to prune computationally expensive layers more aggressively.

2. Iterative Pruning & Fine-Tuning

Sparsity is increased gradually over `$K$` steps, with fine-tuning after each step to allow the network to recover and adapt.

3. Knowledge Distillation (KD)

$$L_{\mathrm{KD}} = (1-\alpha)L_{\mathrm{CE}}(y,\hat{y}_s) + \alpha T^{2}{}_{\mathrm{KL}}(\sigma(\hat{y}_t/T)\|\sigma(\hat{y}_s/T))$$

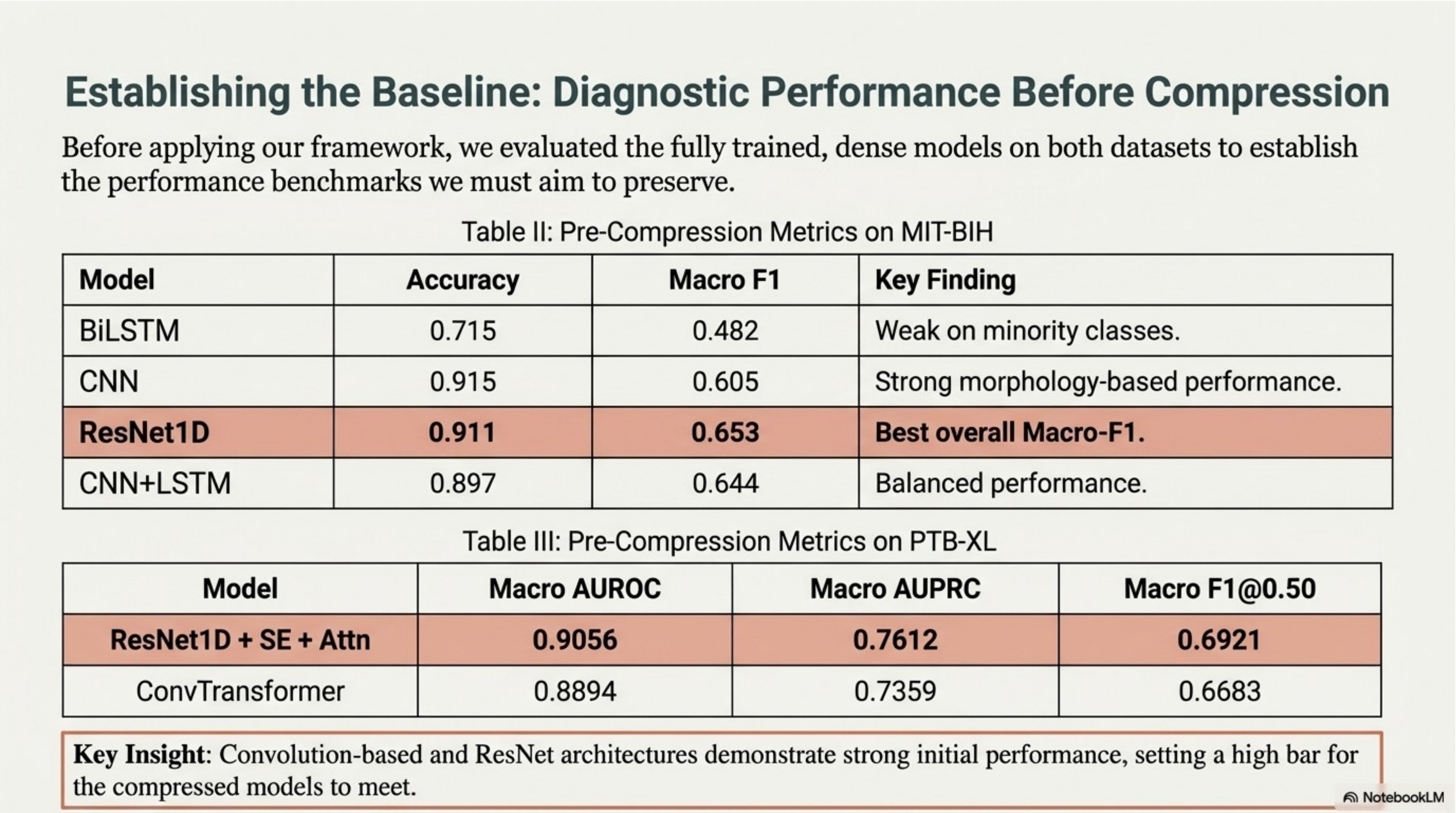
The pruned "student" model learns from the original "teacher" model's logits, helping to preserve nuanced decision boundaries.

4. Dynamic Sparsity Scheduling

$$s_l(t) = S_{\mathrm{target}}(1-e^{-\beta_l t})$$

Allows for adaptive pruning, preserving critical shallow layers while more aggressively pruning redundant deeper layers.

NotebookLM



Establishing the Baseline: Diagnostic Performance Before Compression

Before applying our framework, we evaluated the fully trained, dense models on both datasets to establish the performance benchmarks we must aim to preserve.

Table II: Pre-Compression Metrics on MIT-BIH

Model	Accuracy	Macro F1	Key Finding
BiLSTM	0.715	0.482	Weak on minority classes.
CNN	0.915	0.605	Strong morphology-based performance.
ResNet1D	**0.911**	**0.653**	**Best overall Macro-F1.**
CNN+LSTM	0.897	0.644	Balanced performance.

Table III: Pre-Compression Metrics on PTB-XL

Model	Macro AUROC	Macro AUPRC	Macro F1@0.50
ResNet1D + SE + Attn	**0.9056**	**0.7612**	**0.6921**
ConvTransformer	0.8894	0.7359	0.6683

Key Insight: Convolution-based and ResNet architectures demonstrate strong initial performance, setting a high bar for the compressed models to meet.

A Standard Approach: Performance Under Global L1-Norm Pruning

To set a reference point, we first applied standard iterative L1 pruning. This method removes the smallest magnitude weights, a common but less sophisticated approach to model compression.

TABLE IV: Summary of L1 Pruning Results

Model	Dataset	Params Before (k)	Params After (k)	Sparsity (%)	Test Score (Acc/F1 or AUROC/AUPRC)
CNN1D 3C	MIT-BIH	188.0	38.4	79.5%	0.913 / 0.610
LSTM1D 3C	MIT-BIH	530.9	109.4	79.4%	0.80 / 0.45
CNNLSTM Attn	MIT-BIH	349.9	183.2	47.6%	0.867 / 0.638
ResNet1D Attn	**MIT-BIH**	**417.9**	**86.0**	**79.4%**	**0.916 / 0.656**
ResNet1D SE Attn	PTB-XL	1661.1	623.6	47.8%	0.874 / 0.654
ConvTrans	**PTB-XL**	**1711.9**	**1177.7**	**31.2%**	**0.890 / 0.736**

Key Finding: Standard L1 pruning can achieve high sparsity (up to ~80%) on MIT-BIH models while maintaining strong performance. Transformer-based models on PTB-XL show less tolerance for pruning due to their dense attention mechanisms.

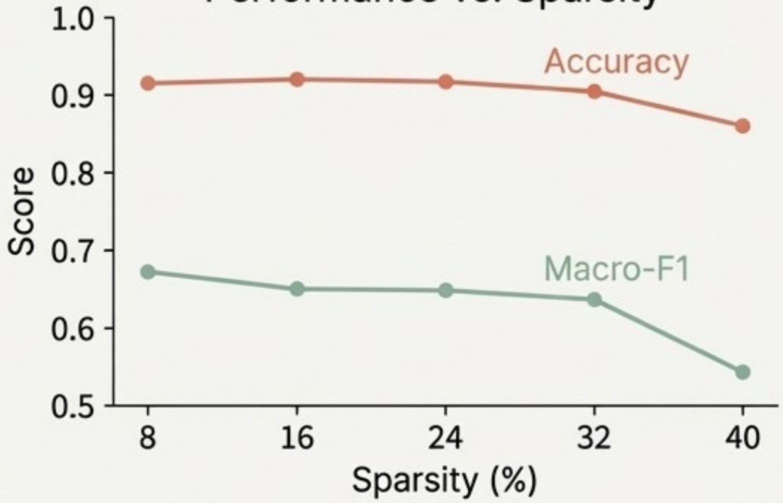
Energy-Aware Pruning on MIT-BIH: Performance Remains Stable Under Sparsity

We applied our energy-aware framework to the ResNet1D_Attn model, progressively increasing sparsity over five steps. All results are the mean ± standard deviation over three runs to ensure statistical robustness.

TABLE V: ResNet1D_Attn on MIT-BIH

Pruning Step	Sparsity (%)	Accuracy	Macro-F1
1	8	0.915 ± 0.002	0.671 ± 0.004
2	16	0.920 ± 0.001	0.650 ± 0.003
3	24	0.917 ± 0.002	0.648 ± 0.003
4	32	0.904 ± 0.002	0.637 ± 0.004
5	40	0.860 ± 0.003	0.543 ± 0.005

Performance vs. Sparsity



Key Insight: Diagnostic performance remains highly stable up to ~30% sparsity. The minor degradation at 40% is primarily due to reduced sensitivity in the minority supraventricular (S) class, while performance on Normal (N) and Ventricular (V) beats remains consistent.

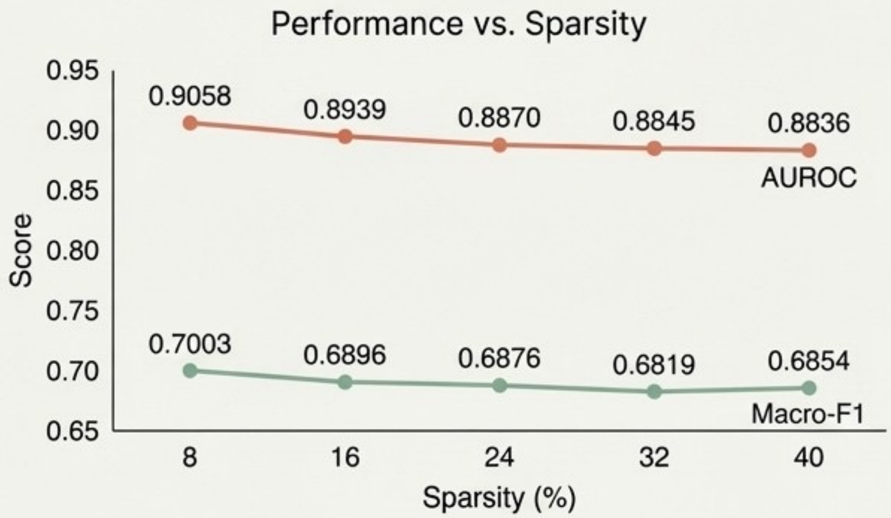
NotebookLM

Scaling to Complex Data: ConvTransformer on PTB-XL Shows High Robustness

The same energy-aware framework was applied to the ConvTransformer model on the 12-lead, record-level PTB-XL dataset. The baseline model has a computational complexity of 230.56 MFLOPs.

TABLE VI: ConvTransformer on PTB-XL

Pruning Step	Sparsity (%)	AUROC	Macro-F1
1	8.0	0.9058 ± 0.002	0.7003 ± 0.004
2	16.0	0.8939 ± 0.002	0.6896 ± 0.003
3	24.0	0.8870 ± 0.002	0.6876 ± 0.004
4	32.0	0.8845 ± 0.002	0.6819 ± 0.003
5	40.0	0.8836 ± 0.003	0.6854 ± 0.004



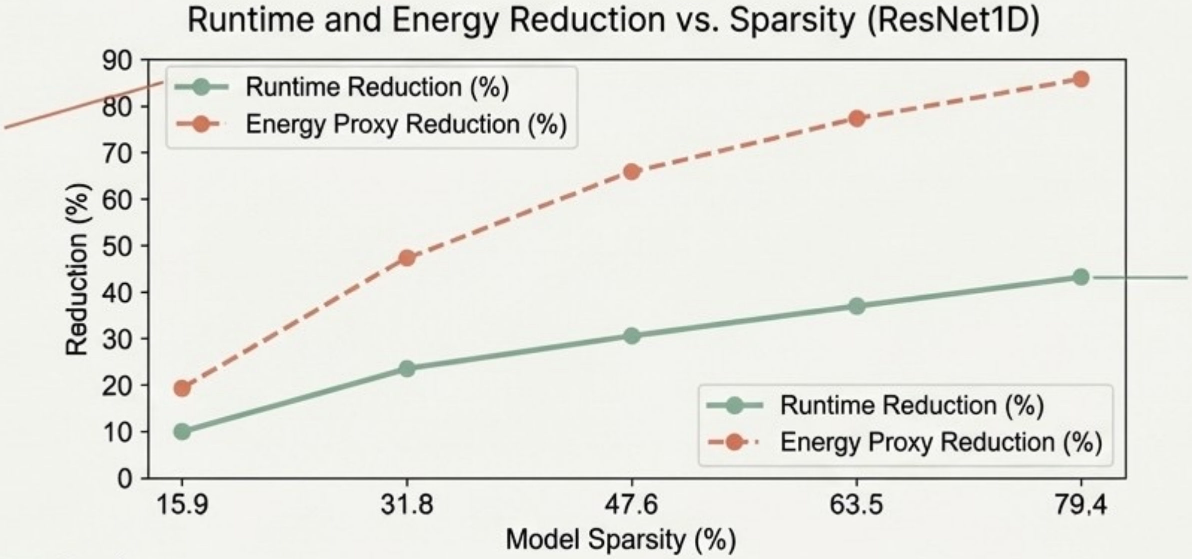
Key Insight: Even at 40% sparsity in its convolutional layers, the ConvTransformer shows minimal degradation (less than 2%) in AUROC and Macro-F1. This indicates that the hybrid attention mechanisms effectively preserve diagnostic information, making the model robust to weight reduction.

The Tangible Impact: Drastic Energy Reduction in Convolutional Models

For the convolution-dominant ResNet1D model on MIT-BIH, pruning directly translates to significant reductions in computational load, latency, and energy consumption.

~86%

Energy Proxy Reduction



~43%

Runtime Reduction

~75%

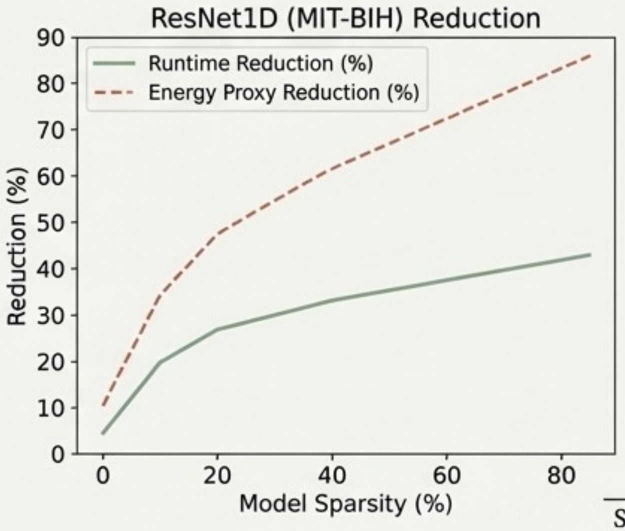
FLOPs Reduction

Sparsity (%)	FLOPs (M)	Runtime (ms)	Energy Proxy (FLOPs×ms)
15.9	15.60	2.48	38.67
79.4	3.82	1.42	**5.43**

These results confirm the framework's suitability for deploying real-time ECG inference on embedded and wearable devices.

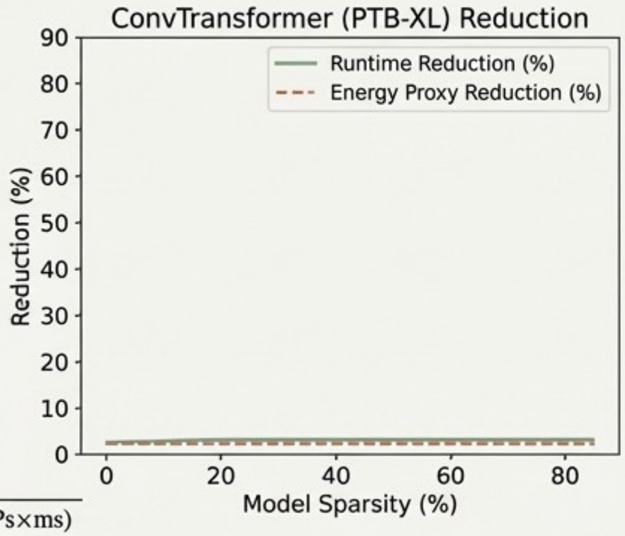
A Tale of Two Architectures: Why Transformers Behave Differently

In contrast to ResNet1D, the ConvTransformer's computational metrics remained largely static despite increasing weight sparsity. This highlights a fundamental difference between architecture types.



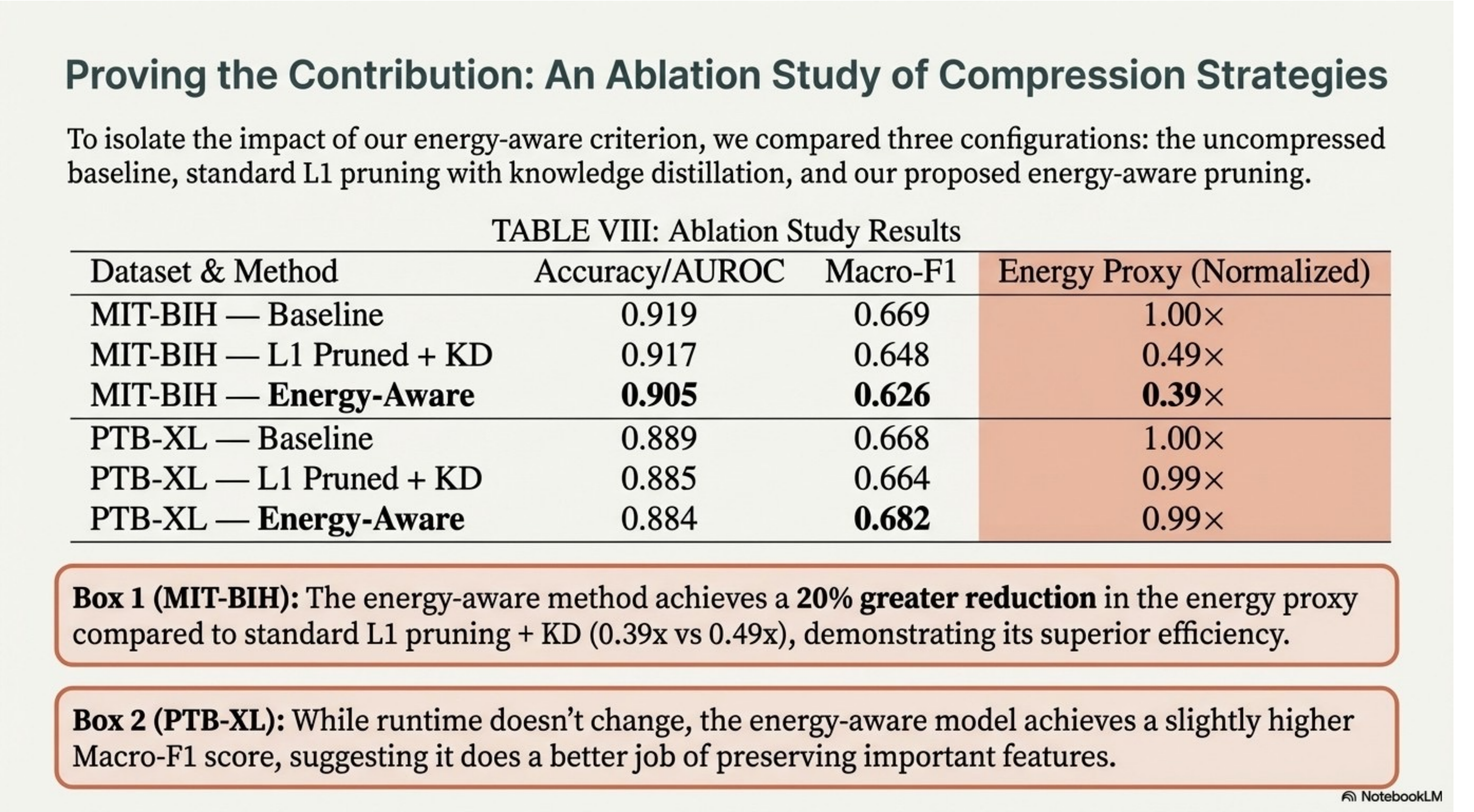
Why

- Pruning was applied **only** to the **convolutional** layers.
- The model's total FLOPs are dominated by the large, dense self-attention blocks, which were not pruned.
- Unstructured pruning does not typically lead to hardware-level speedups on standard GPUs without specialized sparse kernels.



Sparsity (%)	FLOPs (M)	Runtime (ms)	Energy Proxy (FLOPs×ms)
8.0	**230.56**	30.84	7111.40
40.0	**230.56**	30.44	7018.00

For transformer-based models, achieving tangible runtime acceleration requires **structured pruning** (removing entire channels or blocks), not just unstructured weight removal.



Proving the Contribution: An Ablation Study of Compression Strategies

To isolate the impact of our energy-aware criterion, we compared three configurations: the uncompressed baseline, standard L1 pruning with knowledge distillation, and our proposed energy-aware pruning.

TABLE VIII: Ablation Study Results

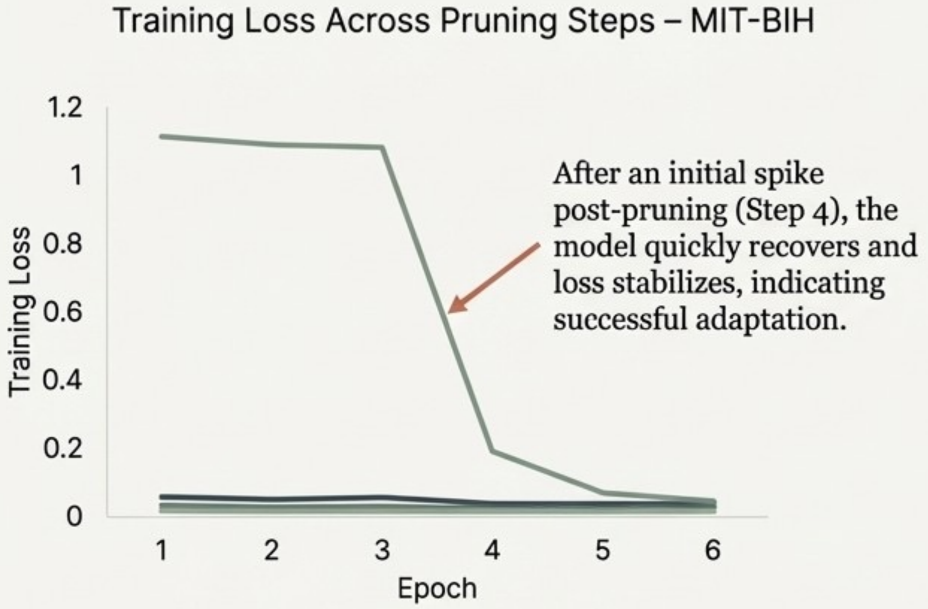
Dataset & Method	Accuracy/AUROC	Macro-F1	Energy Proxy (Normalized)
MIT-BIH — Baseline	0.919	0.669	1.00×
MIT-BIH — L1 Pruned + KD	0.917	0.648	0.49×
MIT-BIH — **Energy-Aware**	**0.905**	**0.626**	**0.39**×
PTB-XL — Baseline	0.889	0.668	1.00×
PTB-XL — L1 Pruned + KD	0.885	0.664	0.99×
PTB-XL — **Energy-Aware**	0.884	**0.682**	0.99×

Box 1 (MIT-BIH): The energy-aware method achieves a **20% greater reduction** in the energy proxy compared to standard L1 pruning + KD (0.39x vs 0.49x), demonstrating its superior efficiency.

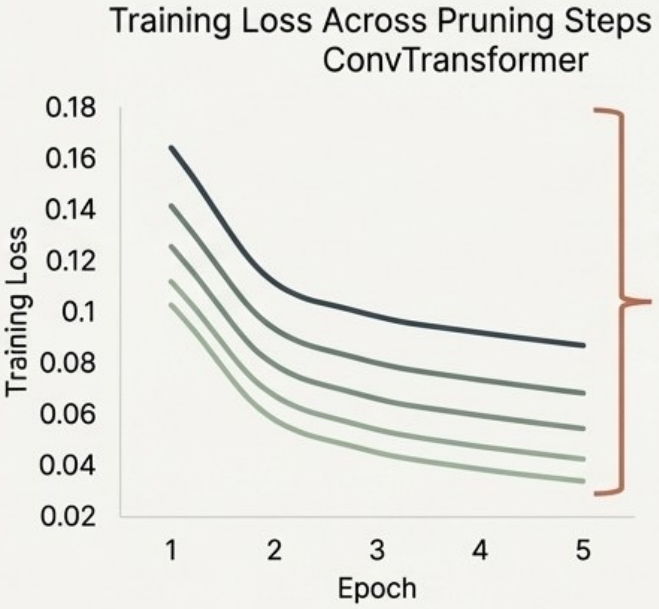
Box 2 (PTB-XL): While runtime doesn't change, the energy-aware model achieves a slightly higher Macro-F1 score, suggesting it does a better job of preserving important features.

Verifying Stability: Training Loss Converges After Each Pruning Step

- The iterative fine-tuning process is crucial for model recovery. We tracked the training loss after each pruning step to ensure the models maintained their learning capacity and did not become unstable.



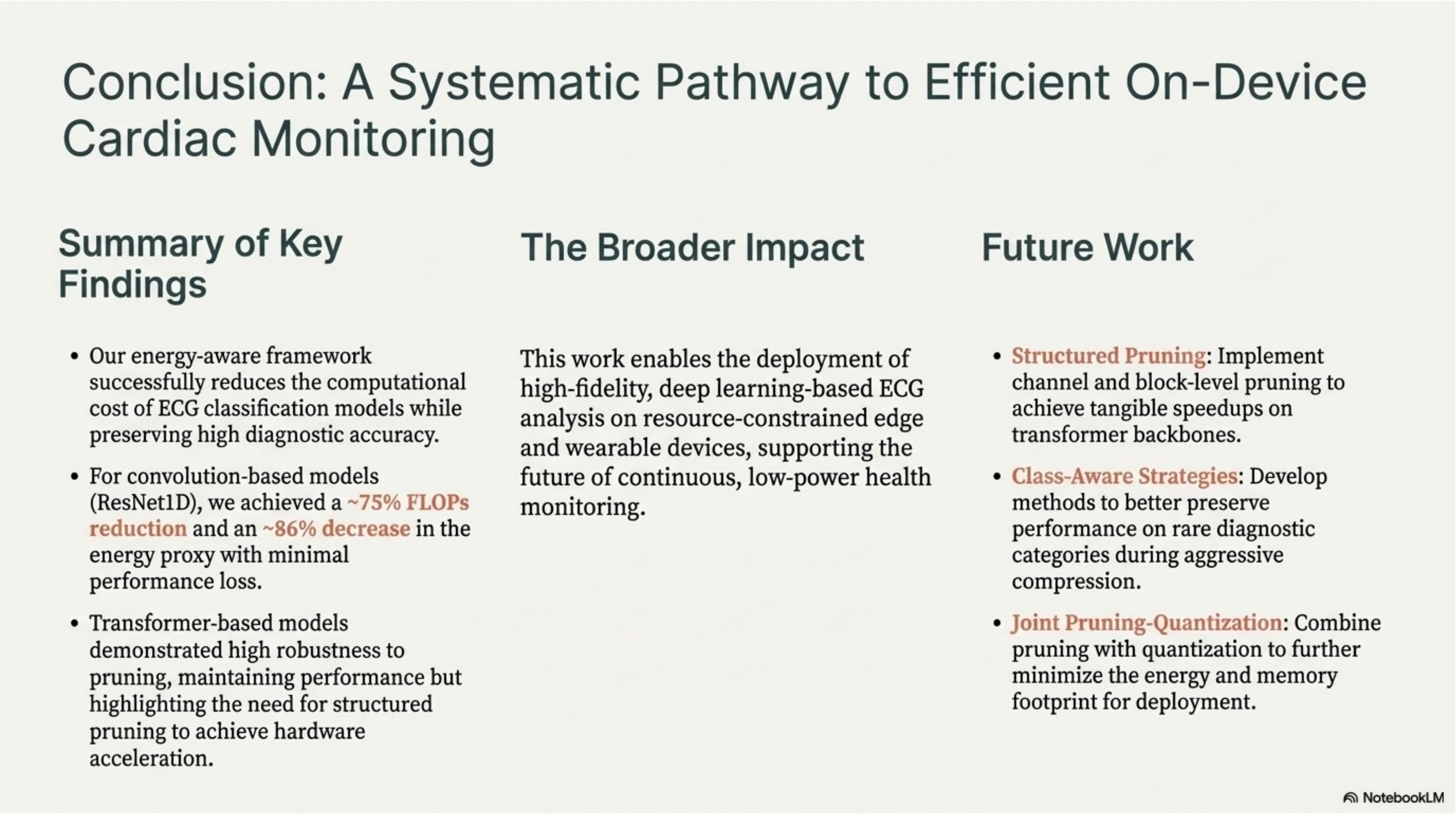
After an initial spike post-pruning (Step 4), the model quickly recovers and loss stabilizes, indicating successful adaptation.



Each curve shows consistent and smooth loss reduction during fine-tuning, demonstrating stable convergence across all pruning stages.

The gradual stabilization of loss across epochs and steps confirms that the framework allows the models to retain generalization capacity even at high sparsity levels.

NotebookLM



Conclusion: A Systematic Pathway to Efficient On-Device Cardiac Monitoring

Summary of Key Findings

- Our energy-aware framework successfully reduces the computational cost of ECG classification models while preserving high diagnostic accuracy.
- For convolution-based models (ResNet1D), we achieved a **~75% FLOPs reduction** and an **~86% decrease** in the energy proxy with minimal performance loss.
- Transformer-based models demonstrated high robustness to pruning, maintaining performance but highlighting the need for structured pruning to achieve hardware acceleration.

The Broader Impact

This work enables the deployment of high-fidelity, deep learning-based ECG analysis on resource-constrained edge and wearable devices, supporting the future of continuous, low-power health monitoring.

Future Work

- **Structured Pruning**: Implement channel and block-level pruning to achieve tangible speedups on transformer backbones.
- **Class-Aware Strategies**: Develop methods to better preserve performance on rare diagnostic categories during aggressive compression.
- **Joint Pruning-Quantization**: Combine pruning with quantization to further minimize the energy and memory footprint for deployment.