Optimizing Reduced-Lead ECG Diagnosis: An Interpretable Pipeline for Lead Selection and Model Adaptation

**Luisa G. Porfírio[1]*, Guilherme H. G. Evangelista[1]*, Pedro B. Rigueira[1]*,
Caio Souza Grossi[1]*, Artur Xavier[1]*, Victoria Andrade Flores de Mello[1]*,
Raquel Teodoro[1]*, Pedro Dutenhefner[1], Gabriela M. M. Paixão[2],
Gisele L. Pappa[1], Antonio Ribeiro[2], Wagner Meira Jr.[1]**

[1] Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
[2] Centro de Telessaúde – Hospital das Clínicas da UFMG

Belo Horizonte – MG – Brazil

{luisagontijo, guilherme.evangelista}@dcc.ufmg.br

{pedrobacelar.rigueira, caio.grossi, arturxavier}@dcc.ufmg.br

{victoriaflores, raquel.teodoro, glpappa, meira}@dcc.ufmg.br

gabimiana@gmail.com, tom@hc.ufmg.br

Abstract. *The 12-lead electrocardiogram (ECG) is vital for heart diagnosis but mostly limited to clinics due to its complexity. Wearable devices use fewer electrical viewpoints (leads), democratizing cardiac care by enabling monitoring at home or in low-resource settings. This shift raises two key challenges: selecting the most informative leads and adapting AI models to keep accuracy. We tackle this with a data-driven pipeline that ranks leads by combining multiple model-interpretability methods. Our evaluation shows that a model architecturally adapted to use only the top two leads (V1, I) achieves a macro F1-score of 0.885, matching the full 12-lead baseline. This work provides a framework for efficient, powerful AI systems, advancing accessible cardiac diagnostics.*

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of death worldwide, responsible for over 17 million deaths each year [World Health Organization 2024]. Early and accurate diagnosis is key to reducing this burden, and the 12-lead electrocardiogram (ECG) remains the clinical gold standard for initial cardiac assessment. An ECG records the heart's electrical activity from multiple perspectives, known as leads, which are virtual viewpoints derived from voltage differences between electrodes placed on the patient's limbs and chest. This standard 12-lead setup offers a rich, multi-angled view of cardiac function, enabling the detection of a wide range of conditions. Despite clinical effectiveness, this configuration depends on specialized equipment and trained personnel, limiting access in low-resource environments.

This paradigm of hospital-centric diagnosis is now being challenged by the rapid proliferation of wearable health technologies and remote patient monitoring (RPM) systems. The global market for smart wearable ECG monitors, projected to surpass USD 3.5

*These authors contributed equally to this work.

billion by 2030 [Grand View Research 2025], signals a fundamental shift towards decentralized healthcare. Such technologies have the potential to extend diagnostics from the clinic into community and home settings. By favoring simpler devices with one to six leads, they offer a practical path to continuous, real-world monitoring, especially for underserved populations. This transformation, however, hinges on a critical question: which leads are truly indispensable for a reliable diagnosis?

The promise of accessible ECG monitoring is supported by a well-known property of the standard 12-lead layout: its inherent redundancy. Many leads are linear combinations or spatially correlated projections of the same underlying cardiac vector. Recent work has not only quantified this redundancy [Ramirez et al. 2024] but also demonstrated that the full 12-lead signal can be accurately reconstructed from as few as three or four inputs [Gradowski and Buchner 2025]. While this confirms that fewer leads can suffice in theory, a significant practical challenge remains: how to maintain, or even enhance, diagnostic accuracy and clinical trust when operating with a strategically reduced set of leads. Answering this is the key to unlocking the full potential of portable ECG devices.

This challenge involves two closely linked problems: first, which subset of leads preserves the most critical diagnostic information; and second, how to adapt deep learning models, both architecturally and procedurally, to maintain performance with such limited input. To address these, we propose a data-driven pipeline based on the LGA-ECG model [Buzelin et al. 2025], a state-of-the-art deep learning architecture for multi-label ECG classification. Our pipeline has two main stages. The first introduces a systematic lead selection methodology: treating the trained 12-lead model as an expert, we interrogate its decision process from multiple complementary angles to robustly assess lead importance. We combine four analytical perspectives: measuring prediction impact when a lead is masked; quantifying changes in overall model loss; analyzing disruptions in Attention Maps caused by perturbations; and attributing features to source signals via Integrated Gradients (IG). These results are fused into a unified importance ranking for all 12 leads.

This ranking allows us to transition into the second stage, where we construct reduced-lead subsets of varying granularity, specifically the Top-1, Top-2, and Top-3 leads. To empirically quantify the trade-off between the number of input channels and diagnostic performance, we then examine four different model adaptation strategies for handling these reduced-lead inputs: zero-shot inference on masked inputs; fine-tuning the original model on masked data; training the full 12-lead model from scratch with masked inputs; and training a new, smaller model with a reduced input layer tailored to the selected leads. This approach enables a thorough assessment of model performance under constrained input conditions.

By integrating lead selection with a systematic exploration of model adaptation techniques, this work presents a comprehensive framework to address a critical challenge: maintaining diagnostic accuracy when moving from the 12-lead standard to the fewer leads used in wearable and low-resource settings. Our results demonstrate that a strategically chosen two-lead subset can, with proper model adaptation, match the overall diagnostic performance of the full 12-lead ECG. Furthermore, our work establishes an empirical basis for navigating the crucial trade-off between hardware simplicity for broad screening and the need for additional leads to achieve high precision in specific diagnoses. By providing a guide to optimize this balance, our findings aim to accelerate the deployment of effective AI-powered systems in portable devices, contributing to more equitable and accessible cardiac care.

2. Related Works

Our work is situated at the intersection of three key research areas. We begin with the foundation of Deep Learning for ECG Classification, then narrow our focus to the specific challenges and existing approaches in Reduced-Lead ECG diagnostics. Finally, we leverage techniques from Interpretability in ECG Deep Learning, not merely for explanation, but as the core engine for our lead selection pipeline.

Deep Learning for ECG Classification. The traditional automated analysis of ECGs has transitioned from signal processing with handcrafted features to deep learning models that learn directly from raw signals. Convolutional Neural Networks (CNNs) excel at extracting morphological features, with models like the 34-layer CNN by Hannun et al. achieving cardiologist-level performance in arrhythmia detection [Hannun et al. 2019]. While recurrent networks have been used to model temporal dependencies, recent hybrid architectures and Transformers have established new performance benchmarks on large-scale datasets like PTB-XL [Wagner et al. 2020].

Reduced-Lead ECG. While the 12-lead ECG is the clinical standard, its use is impractical in ambulatory and wearable contexts, motivating research into reduced-lead diagnostics. The central challenge is to maintain diagnostic accuracy with fewer leads. Methodologies for lead selection include clinical heuristics based on electrophysiological principles [Tsouri and Ostertag 2014] and data-driven approaches. The latter category encompasses techniques such as quantifying the information loss from lead removal through signal reconstruction [Gradowski and Buchner 2025] or evaluating performance on downstream tasks.

Interpretability in ECG Deep Learning. The clinical adoption of deep learning models is often hindered by their ”black box” nature, making interpretability essential for building trust and ensuring safe deployment. Explainable AI techniques provide intuition into model predictions by identifying influential features in the input signal. In ECG analysis, methods like Grad-CAM have proven effective at highlighting clinically relevant waveform segments [Vijayarangan et al. 2020] and often outperform other attribution methods in localization tasks. Perturbation-based approaches, which systematically mask leads or time segments to measure their impact on model output, have also been used to assess feature importance [Vijayarangan et al. 2020]. More recently, techniques that measure shifts in a model’s internal attention distributions upon lead masking have been proposed to quantify each lead’s contribution [Oh et al. 2022]. Our work utilizes such interpretability methods not merely for post-hoc explanation but as an integral component of our data-driven lead selection pipeline.

3. Experimental Setup

Following the original LGA-ECG methodology, we used a 90/10 split for training and validation on CODE-15%, a publicly available, expert-labelled ECG dataset from the Telehealth Network of Minas Gerais (TNMG) [Ribeiro et al. 2021]. Our focus is on multi-label classification of six diagnostic conditions in the test sets: Sinus Tachycardia (ST), Sinus Bradycardia (SB), Atrial Fibrillation (AF), Right Bundle Branch Block (RBBB), Left

Bundle Branch Block (LBBB), and 1st Degree Atrioventricular Block (1dAVb). For evaluation, performance was assessed on a combined test set merging the public CODE-TEST dataset [Ribeiro et al. 2020] with AUDIT, a companion set of 432 annotated recordings from [Rigueira et al. 2024]. Both originate from CODE-15% and share identical formats and review protocols, forming a consistent test pool. The set was split 50/50 stratified by class: one half for interpretability, the other for final metrics, preventing information leakage and ensuring a reliable benchmark.

All models were trained under a unified protocol to ensure fair comparison. We used the AdamW optimizer with a cosine annealing learning rate schedule, starting at 10^{-3} and decreasing to 10^{-6}. To prevent overfitting, the checkpoint with lowest validation loss was used for evaluation. Training ran on an NVIDIA 4090 GPU. Diagnostic performance was measured by Accuracy, Recall, and F1-Score, with macro-averaged F1 as the main metric due to class imbalance.

4. Methodology

This section details the complete methodology employed to select the most informative ECG leads and adapt the diagnostic models. The entire process, illustrated in Figure 1, is divided into three main parts. First, we describe the baseline model architecture (LGA-ECG) used as our starting point. Next, we present our multi-perspective interpretability framework for lead selection and ranking. Finally, we detail the four model adaptation strategies that were systematically evaluated.

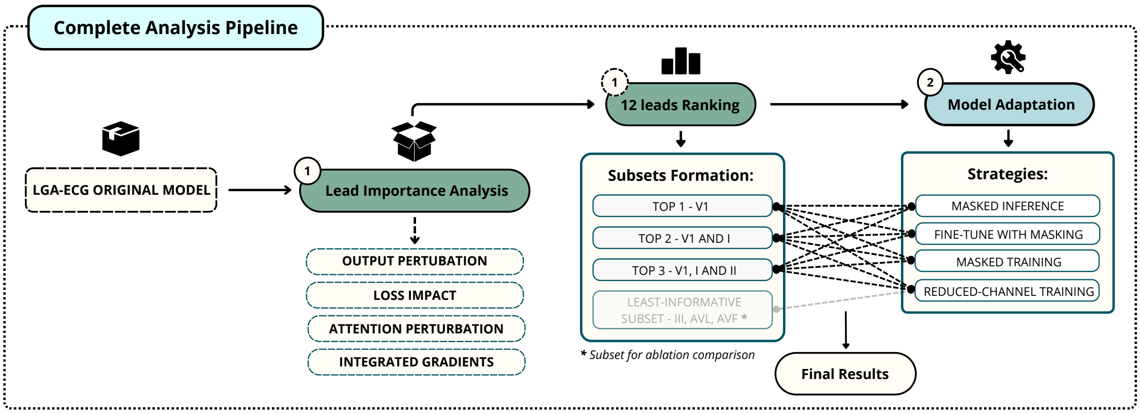

Figure 1. Overview of the proposed pipeline, from lead importance analysis to model adaptation and evaluation.

4.1. Baseline Model Architecture

We adopt a state-of-the-art architecture for ECG analysis as our baseline, founded on an innovative Local-Global (LG) self-attention mechanism. Its core strength is its ability to efficiently capture both fine-grained local morphological features (i.e., ECG waves) and the global temporal context (i.e., heart rhythm). This dual-focus capability is designed to overcome a fundamental limitation of traditional transformers in capturing the nested, hierarchical dependencies inherent in biomedical time-series analysis.

The model's innovation lies in its asymmetric attention computation. Unlike standard transformers, the queries are generated from local windows of the signal to focus on specific morphological details. In contrast, the keys and values are derived from the entire sequence by applying one-dimensional convolutional projections, which provide the global context. This allows the model to efficiently relate local patterns to the overall rhythmic context of the signal using a standard scaled dot-product attention operation. By progressively reducing sequence length in deeper layers, the architecture builds a hierarchical signal representation. Its convolutional components inherently encode positional information, thus eliminating the need for explicit positional encodings.

We chose the LGA-ECG model not only for its state-of-the-art performance, with key metrics detailed in Table 1, but more critically for its architectural suitability for our research. Its hybrid convolutional and attention structure provides a transparent framework for applying our multi-perspective interpretability suite including occlusion analysis, Integrated Gradients, and attention-perturbation metrics. This property is fundamental to our pipeline, enabling the data-driven lead selection that directly addresses the challenge of building robust models for resource-constrained scenarios.

4.2. Lead Selection via Multi-Perspective Interpretability

This section details our method for generating a unified lead importance ranking. We fuse results from four interpretability techniques (three occlusion-based and one gradient-based) to create a final ranking that directly guides the selection of our reduced-lead subsets.

4.2.1. Occlusion-Based Analyses

The central idea of occlusion analysis is to quantify a lead's importance by measuring the impact of its removal on the model's behavior. For all perturbation-based methods below, we systematically generate a modified input $X^{(-l)}$ by zeroing out the channel corresponding to each lead $l \in \{1, \ldots, 12\}$. The final importance score for each method is averaged across all samples in the test set.

Output Perturbation Analysis. This first method quantifies a lead's importance based on its direct impact on the model's final prediction confidence [Vijayarangan et al. 2020]. The underlying principle is that an important lead is one whose removal causes a significant drop in the model's ability to make a correct prediction. To measure this, we calculate the change in the model's output probability, $P(c|X)$, for a pathology c when lead l is occluded. The score captures both correct detections (a drop in confidence for a present condition) and correct rejections (an increase in confidence for an absent condition), as formalized by:

$$\text{Importance}(l, c) = \begin{cases} P(c|X) - P(c|X^{(-l)}) & \text{if } c \text{ is present} \\ P(c|X^{(-l)}) - P(c|X) & \text{if } c \text{ is absent} \end{cases}$$

A higher positive score, averaged across the dataset, indicates that the lead is critical for correctly classifying a specific pathology.

Loss Impact Analysis. As a complementary perspective, this technique assesses a lead's importance by its effect on the model's overall prediction error. It operates on the principle that a lead is crucial if its absence significantly increases the model's loss function.

We formalize this by computing the increase in the multi-label cross-entropy loss when comparing the prediction from the original input with that from the occluded input. The importance score is defined as:

$$\text{Importance}(l) = L(f(X^{(-l)}), Y) - L(f(X), Y)$$

where f is the model, X is the input, Y are the true labels, and L is the loss function. A larger score signifies that the model relies heavily on that lead to minimize its prediction error.

Attention Perturbation Analysis. This final architecture-aware method evaluates a lead's importance by measuring its influence on the model's internal reasoning process. It is based on the principle that a critical lead is one whose removal substantially disrupts the model's internal attention patterns. We quantify this disruption by measuring the Kullback-Leibler (KL) divergence between the model's original attention distribution (A) and the perturbed one (A_{-l}):

$$\text{Importance}(l) = \text{KL}(A||A_{-l}) = \sum_{i,j} A_{ij} \log \left(\frac{A_{ij}}{A_{-l,ij} + \epsilon} \right)$$

A high KL divergence value suggests the lead is fundamental for the model to form stable internal representations for its decisions.

4.2.2. Gradient-Based Saliency (Integrated Gradients)

Following the analysis, our employed gradient-based method attributes a prediction back to its input features, highlighting the most influential signal segments [Sundararajan et al. 2017]. IG identifies these segments by aggregating gradients along a linear path from a neutral baseline (a zero-tensor) to the actual input, X. The attribution for each point i in the signal is formally defined as:

$$\text{IG}_i(X) = (X_i - X'_{\text{baseline},i}) \times \int_{\alpha=0}^{1} \frac{\partial S(X'_{\text{baseline}} + \alpha(X - X'))}{\partial x_i} d\alpha$$

where S is the scoring function (e.g., negative loss). To derive a single importance score for an entire lead l, we sum the absolute attribution values over its time dimension, T:

$$\text{Importance}(l) = \sum_{t=1}^{T} |\text{IG}(X)_{l,t}|$$

A higher aggregate score indicates that the morphology within that lead has a greater direct influence on the model's final decision.

4.2.3. Formulating Lead Subsets from a Unified Importance Ranking

The final step in our selection strategy is to synthesize the results from our multi-perspective interpretability analyses into a single, robust ranking of lead importance. The procedure follows two main steps:

Metric Normalization. To enable a fair comparison between our interpretability methods, which produce scores on varying scales, we first normalize them. Each lead's raw importance score is converted into a relative contribution by dividing it by the sum of all scores from that specific analysis, transforming the raw values into a probability distribution.

$$\text{Score}_{\text{norm}}(l, m) = \frac{\text{Score}_{\text{raw}}(l, m)}{\sum_{i=1}^{12} \text{Score}_{\text{raw}}(i, m)}$$

where $\text{Score}_{\text{norm}}(l, m)$ is the normalized score for lead l in analysis method m.

Aggregation and Ranking. We then compute a final, unified importance score for each lead by averaging its normalized scores across all M analysis methods. This aggregated score represents the consensus on each lead's diagnostic importance.

$$\text{FinalScore}(l) = \frac{1}{M} \sum_{m=1}^{M} \text{Score}_{\text{norm}}(l, m)$$

Based on this unified ranking, we select three subsets corresponding to the Top-1, Top-2, and Top-3 most informative leads to systematically evaluate the trade-off between input reduction and performance. This systematic approach allows for a direct and objective evaluation of how diagnostic accuracy scales with the number of available leads.

4.3. Model Adaptation Techniques

Once the reduced-lead subsets have been identified, the central question becomes how to best adapt the model to these new reduced input formats. The choice of strategy involves a fundamental trade-off between leveraging pre-trained knowledge from the 12-lead model and training a new, specialized, and more efficient architecture. To determine the most effective approach, we conduct a comparative evaluation of four distinct strategies:

Inference with Masking. This strategy serves as a performance baseline and represents a "zero-effort" adaptation scenario. The original LGA-ECG model, fully pre-trained on 12-lead data, is used directly for inference without any modification or fine-tuning. The input ECGs from the reduced-lead subset are prepared by setting the channels of the discarded leads to zero (zero-masking) to match the model's 12-channel input requirement. While it has no training cost, its performance may be suboptimal as the model was never explicitly taught to handle such inputs.

Fine-Tuning with Masking. This approach tests the hypothesis that pre-trained knowledge can be effectively transferred to the reduced-lead context. Here, we start with the fully trained 12-lead model and fine-tune its weights using the same zero-masked data. By using a low learning rate (e.g., 10^{-7}), this method adapts the learned representations to the new data distribution without erasing the powerful features learned from the full ECG. It aims for a balance, seeking high performance at a lower computational cost than training from scratch.

Training from Scratch with Masking. In this strategy, we retain the original 12-channel model architecture but train it from a random initialization using only the reduced-lead data. As before, unused lead channels are zero-masked. This forces the model to learn representations exclusively from the selected leads while explicitly ignoring the zeroed channels. While this adapts the model's weights specifically for the task, it retains the same number of parameters and computational overhead during inference as the full 12-lead model.

Training a Reduced-Channel Architecture. This final strategy represents the most efficient solution, designed for resource-constrained deployment. We permanently modify the LGA-ECG architecture by changing the 'in_channels' parameter of its first convolutional layer from 12 to the number of leads in the selected subset (e.g., 1, 2, or 3). This new, lighter model is then trained from scratch. This approach yields a model with fewer parameters and lower inference latency, but at the cost of discarding all knowledge from the pre-trained 12-lead model and requiring a full training cycle.

5. Results

Building on the lead selection and model adaptation frameworks detailed previously, this section presents the validation of our pipeline. First, we present the lead importance ranking derived from ourinterpretability analysis, which forms the basis for our reduced-lead subsets. Second, we evaluate the diagnostic performance of models adapted to these subsets, benchmarking them against the 12-lead baseline and concluding with an ablation study that validates our data-driven selection process.

5.1. Lead Importance Analysis

The cross-method comparison of the four importance heatmaps presented in Figure 2 reveals a consistent hierarchy of lead contributions. The analysis shows that lead V1 provides high diagnostic value across numerous pathologies, with its influence being particularly pronounced for classifying right and left bundle branch blocks (RBBB and LBBB), an observation consistent with clinical knowledge. For other conditions, the heatmaps indicate that the diagnostic contribution is more distributed among a small group of top-tier leads, primarily V1, I, and II. This shared reliance on the same highest-ranked leads across different diagnoses further validates their selection for constructing a robust, general-purpose reduced-lead set.

The resulting unified importance hierarchy, displayed in Figure 3, establishes a definitive, data-driven consensus on each lead's overall contribution. This fused ranking confirms the prominence of lead V1 as the single most influential lead, with leads I and II emerging as the next most critical. This ranking forms the empirical foundation for the next stage of our research, where we systematically evaluate model adaptation techniques using a series of reduced-lead subsets constructed from our findings: Top-1 (V1), Top-2 (V1, I), and Top-3 (V1, I, II).

5.2. Performance of Reduced-Lead Adaptation Strategies

Our results highlight a clear hierarchy in adaptation strategies, with the Train with Reduced Channels method consistently proving superior. This validates our pipeline's core hypothesis: a model architecturally tailored to a few, highly informative leads can maintain robust performance. Therefore, our analysis centers on this optimal strategy, with all results referenced from Table 1.

The findings reveal a compelling narrative of performance versus complexity. A model using only the top-ranked lead (V1) is remarkably effective, achieving a macro

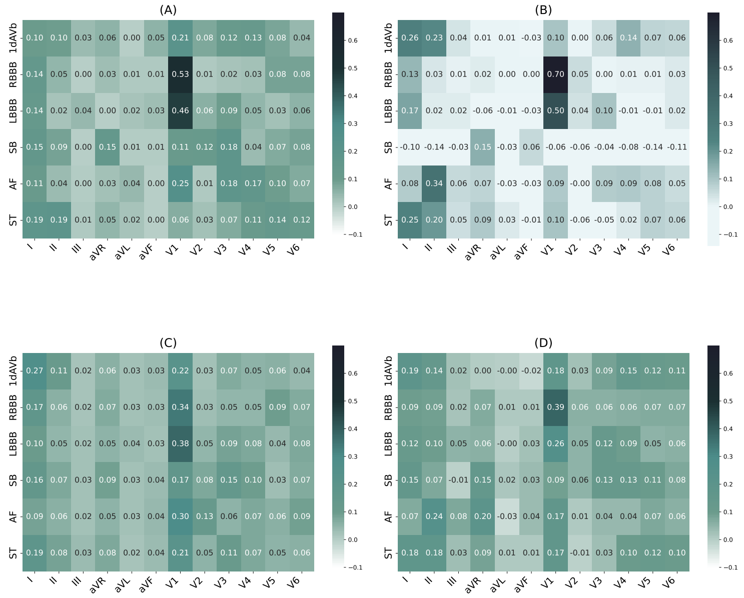

Figure 2. Lead importance heatmaps by method: (A) Output Perturbation, (B) Loss Impact, (C) Attention Perturbation, and (D) Integrated Gradients.

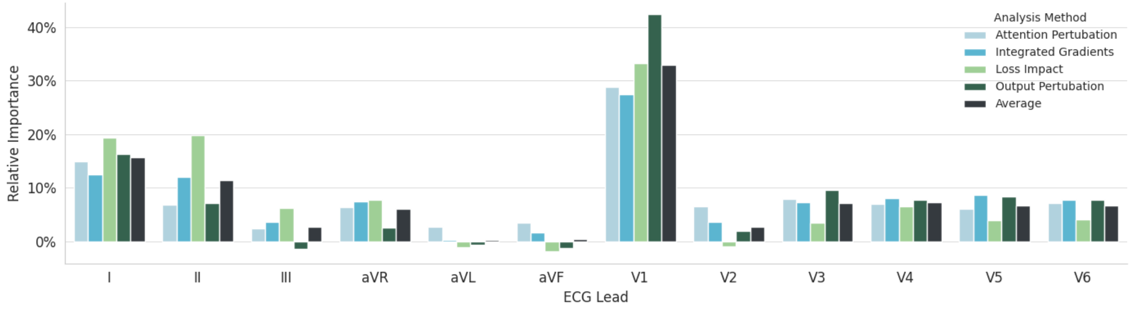

Figure 3. Relative importance of each ECG lead obtained from four complementary interpretability analyses.

F1-score of 0.875, nearing the 12-lead baseline. The key breakthrough comes with the addition of the second-ranked lead (I), as the Top-2 model (V1, I) reaches an F1-score of 0.885, precisely matching the 12-lead standard. This confirms that an 83% reduction in input channels can be achieved without sacrificing overall diagnostic accuracy. Interestingly, adding a third lead (II) introduces a critical trade-off: while it does not improve the overall macro F1-score, it is necessary to achieve peak performance for specific complex diagnoses like LBBB (F1-score of 0.951). This suggests the optimal number of leads is application-dependent, balancing general screening needs against high-precision diagnostics.

5.3. Ablation Study: The Impact of Lead Ranking

To isolate and validate the impact of our ranking methodology, we conducted a final ablation study. The goal was to demonstrate that the high performance of our reduced-lead models stems directly from our selection process, not just from the model adaptation itself. To this end, we employed our best-performing adaptation strategy, the reduced channels technique, to train a model using only the three lowest-ranked leads identified

Table 1. Performance Breakdown for each Lead Subset and Evaluation Method. The best-performing result for each condition across all experiments is highlighted.

		F1 Performance by Condition						Overall Performance		
Lead Subset	**Evaluation Method**	**1dAVb**	**RBBB**	**LBBB**	**SB**	**AF**	**ST**	**F1**	**Accuracy**	**Recall**
(V1)	Inference with Masking	0.207	0.699	0.810	0.363	0.301	0.758	0.524	0.690	0.489
	Finetune with Masking	0.457	0.891	0.914	0.825	0.868	0.902	0.809	0.947	0.736
	Train with Masking	0.792	0.844	0.901	0.873	0.878	0.902	0.866	0.913	0.811
	Train with Reduced Channels	0.739	0.891	0.894	**0.901**	**0.911**	**0.914**	0.875	0.920	0.831
(V1, I)	Inference with Masking	0.439	0.800	0.913	0.753	0.650	0.827	0.730	0.840	0.684
	Finetune with Masking	0.683	0.902	0.925	0.844	0.899	0.904	0.859	0.942	0.799
	Train with Masking	0.727	0.889	0.911	0.831	0.886	0.912	0.860	0.916	0.797
	Train with Reduced Channels	**0.800**	0.909	0.925	0.869	0.888	**0.914**	**0.885**	0.925	**0.849**
(V1, I, II)	Inference with Masking	0.680	0.863	0.889	0.733	0.710	0.846	0.787	0.867	0.721
	Finetune with Masking	0.683	**0.920**	0.938	0.825	0.868	0.902	0.856	**0.963**	0.781
	Train with Masking	0.683	0.902	0.934	0.793	0.878	0.902	0.850	0.910	0.780
	Train with Reduced Channels	0.782	0.901	**0.951**	0.823	0.860	0.880	0.866	0.909	0.815
(12-lead)	Baseline LGA-ECG Model	0.783	0.909	0.938	0.870	0.901	0.907	**0.885**	0.927	0.848

by our pipeline (Bottom-3: III, aVL, aVF) and compared its performance against our established high-performing Top-2 subset and the 12-lead baseline.

The results, presented in Table 2, are definitive. As previously established, the Top-2 model's performance mirrors the 12-lead baseline, confirming its efficacy. The critical finding here, however, is the catastrophic failure of the Bottom-3 model. Its macro F1-Score plummets to 0.710 and its recall to 0.576. This model's misleadingly high accuracy is a classic artifact of class imbalance, indicating it defaulted to predicting the majority class in the absence of useful diagnostic signals. This sharp contrast offers empirical support for the ability of our interpretability-guided ranking to distinguish clinically informative leads from those with limited relevance.

Table 2. Ablation study contrasting the full 12-lead baseline with models trained on the Top-2 and Bottom-3 lead subsets.

Lead Subset	F1-Score	Accuracy	Recall
12-Lead (Baseline)	**0.885**	0.927	0.848
Top-2 (V1, I)	**0.885**	0.925	**0.849**
Bottom-3 (III, aVL, aVF)	0.710	**0.968**	0.576

6. Discussion

Our study highlights a critical insight for the future of remote cardiac monitoring: the optimal number of ECG leads is not a fixed choice, but an application-specific hyperparameter. We established a quantifiable trade-off between the hardware simplicity of using fewer leads and the diagnostic precision required for complex conditions. While a two-lead configuration (V1, I) matched the overall performance of the 12-lead standard, the inclusion of a third lead was necessary to maximize accuracy for challenging diagnoses like LBBB. This finding suggests a practical path for developing tiered diagnostic systems: highly accessible devices for broad screening and slightly more complex ones for specialized, high-fidelity monitoring.

The effectiveness of our selected leads is rooted in fundamental electrophysiological principles, empirically validated by our data-driven pipeline. The model's reliance on V1, with its direct view of the ventricular septum, and lead I, providing a complementary lateral perspective, demonstrates that our interpretability-based ranking successfully identified leads that capture essential, non-redundant axes of the heart's electrical activity. The catastrophic failure of the model trained on the lowest-ranked leads further confirms that our methodology effectively distinguishes clinically informative signals from noise, a crucial step for building trust in automated diagnostic systems.

Our approach also represents a methodological shift from traditional lead-selection techniques. Unlike clinical heuristics or signal reconstruction methods, our pipeline interrogates the internal reasoning of the deep learning model itself. Classical feature selection methods often fail to capture the complex, non-linear relationships that deep models learn; in contrast, our model-centric approach provides a more faithful representation of the actual decision-making process. This ensures the selected leads are precisely those the model uses to make its decisions, creating a synergistic relationship between the input data and the architecture. We acknowledge, however, that these findings are based on a single model architecture and dataset. The generalization of our specific lead subset should be validated across more diverse populations and models. Furthermore, our current ranking is static and universal, while a dynamic, patient-specific approach could present a valuable, albeit more complex, avenue of research.

7. Conclusion

In this work, we introduced and validated a comprehensive, data-driven pipeline for designing and adapting deep learning models for reduced-lead ECG diagnosis. We demonstrated that a model architecturally adapted for only two strategically chosen leads (V1 and I) can achieve the same diagnostic accuracy as the full 12-lead baseline, confirming that significant input reduction is possible without sacrificing overall performance. Our analysis established that training a new, architecturally leaner model from scratch is the most effective adaptation strategy.

The primary contribution of this research is a validated and generalizable template that provides a clear, evidence-based guide for developing resource-efficient solutions in cardiovascular diagnostics. This is particularly relevant for applications in wearable health devices and low-resource clinical settings. Building on this foundation, future work could explore two promising avenues: first, the development of dynamic selection protocols that choose optimal leads based on a patient's initial data; and second, a hybrid generative approach where a minimal set of leads is used to algorithmically reconstruct the full 12-lead signal. Such advances will continue to drive innovation towards cardiovascular care systems that are more democratic, effective, and safe.

References

Buzelin, A., Dutenhefner, P. R., Rezende, T., Porfirio, L. G., Bento, P., Aquino, Y., Fernandes, J., Santana, C., Miana, G., Pappa, G. L., Ribeiro, A., and Jr, W. M. (2025). A cnn-based local-global self-attention via averaged window embeddings for hierarchical ecg analysis. *arXiv preprint*, arXiv:2504.16097.

Gradowski, T. and Buchner, T. (2025). Deep learning model for ecg reconstruction reveals the information content of ecg leads. *arXiv preprint*, arXiv:2502.00559.

Grand View Research (2025). Smart wearable ecg monitors market size, share & trends analysis report, 2030. Accessed 23 June 2025.

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

Oh, J., Chung, H., myoung Kwon, J., gyun Hong, D., and Choi, E. (2022). Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. *arXiv preprint*, arXiv:2203.06889.

Ramirez, E., Ruiperez-Campillo, S., Casado-Arroyo, R., and Merino, J. L. (2024). The art of selecting the ECG input in neural networks to classify heart diseases: maximizing information and reducing redundancy. *Frontiers in Physiology*, 15:1452829.

Ribeiro, A. H., Paixao, G. M., Lima, E. M., Horta Ribeiro, M., Pinto Filho, M. M., Gomes, P. R., Oliveira, D. M., Meira Jr, W., Schon, T. B., and Ribeiro, A. L. P. (2021). CODE-15%: a large scale annotated dataset of 12-lead ECGs.

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., Schön, T. B., and Ribeiro, A. L. P. (2020). Code-test: An annotated 12-lead ecg dataset.

Rigueira, P. B., Evangelista, G. H. G., Porfírio, L. G., Grossi, C. S., Buzelin, A., Pappa, G. L., Paixão, G. M. M., Ribeiro, A., and Jr., W. M. (2024). Optimizing ecg audits: Clustering-based identification of ambiguous exams. *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 21:61–72.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*, 70:3319–3328.

Tsouri, G. R. and Ostertag, M. H. (2014). Patient-specific 12-lead ecg reconstruction from sparse electrodes using independent component analysis. *IEEE Journal of Biomedical and Health Informatics*, 18(2):476–481.

Vijayarangan, S., Murugesan, B., R, V., P., P. S., Joseph, J., and Sivaprakasam, M. (2020). Interpreting deep neural networks for single-lead ecg arrhythmia classification. *arXiv preprint*, arXiv:2004.05399.

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

World Health Organization (2024). Cardiovascular diseases (cvds) – fact sheet. Accessed 23 June 2025.