ECG Classification based on Inter-lead Correlation Features

Marina Ilie-Ablachim * **Andreea Udrea** *

* *National University of Science and Technology POLITEHNICA Bucharest*
email: marina.ciocirlan@stud.acs.upb.ro, andreea.udrea@upb.ro

Abstract: Detection of cardiac abnormalities in 12-lead electrocardiograms (ECGs) requires models that are both accurate and interpretable. In this work, we investigate the diagnostic value of handcrafted inter-lead correlation features computed with Pearson, Spearman, and Kendall coefficients. Using the PTB-XL dataset, we evaluate statistical methods and classical machine learning methods on both binary (normal vs. abnormal) and multiclass (five diagnostic categories) tasks. Results show that simple statistical descriptors achieve strong performance in the binary setting and moderate performance in the multiclass setting. To interpret model decisions, we apply SHAP and permutation feature importance, which consistently highlight correlations between anatomically related leads as key predictors. These findings confirm that inter-lead dependencies carry physiologically meaningful information and can be effectively captured with simple, transparent models. While limited in their ability to represent nonlinear or patient-specific dynamics, correlation-based features provide a strong, interpretable baseline and motivate further exploration of adaptive, data-driven approaches.

Keywords: Electrocardiogram, Cardiac abnormality detection, Inter-lead correlation, Statistical learning, Model interpretability

1. INTRODUCTION

Cardiovascular diseases remain the leading cause of mortality worldwide, underscoring the need for timely and reliable electrocardiogram (ECG) interpretation. Deep learning models have achieved impressive accuracy in ECG classification, but their limited interpretability remains a barrier to clinical adoption (Esteva et al., 2019). Many modern approaches treat ECG channels independently or rely on abstract features, which can obscure physiologically meaningful relationships across the 12 standard leads. This challenge has motivated research into interpretable ECG classification, including the use of explanation methods such as SHapley Additive exPlanations (SHAP) within deep learning pipelines (Zhang et al., 2021; Anand et al., 2022; Lundberg and Lee, 2017).

Clinical studies further emphasize the diagnostic importance of individual leads, sometimes overlooked in practice—for example, lead aVR in acute coronary syndromes (Gorgels et al., 2001). More broadly, neglecting inter-lead dependencies may hinder the detection of subtle or spatially diffuse problems such as conduction delays or non-ST-elevation myocardial infarction. Foundational work in ECG monitoring has long shown that electrode placement and inter-lead dependencies are crucial for correct interpretation (Dash, 2002; Khunti, 2014). Early studies also demonstrated that reduced-lead systems, when mathematically corrected, can approximate the diagnostic information of the full 12-lead ECG (Pipberger et al., 1961; Trobec and Tomašić, 2011; Tomašić and Trobec, 2013), reinforcing the value of cross-lead coherence.

We expect that normal ECGs exhibit stable inter-lead relationships, while pathological signals disrupt these patterns in ways consistent with their anatomical origin. To evaluate this hypothesis, we compute statistical correlations—Pearson, Spearman, and Kendall—across all pairs of the 12 leads and use them as features for classification.

Our work makes three main contributions:

(1) **Feature design and modeling.** We propose a feature set based exclusively on inter-lead correlations and evaluate its predictive value using classical machine learning models: Logistic Regression (LR), Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest (RF). These methods remain competitive baselines for ECG analysis (Jambukia et al., 2015), and we show that correlation-based features alone yield consistent performance in both binary and multiclass tasks.

(2) **Benchmarking on PTB-XL.** We assess our approach on PTB-XL, one of the largest and most widely used ECG datasets (Wagner et al., 2020). While deep learning approaches on PTB-XL report state-of-the-art performance (macro AUC $> 93\%$) (Anand et al., 2022), they often require millions of parameters and offer limited transparency. In contrast, our correlation features enable lightweight classical models that are both efficient and interpretable.

(3) **Interpretability.** We apply SHAP to highlight which lead-pair correlations drive classification decisions. Unlike deep models, where SHAP must be applied post hoc to complex representations, our features are intrinsically interpretable because each

directly encodes a physiologically meaningful relationship between leads.

In the binary classification task (Normal vs. Abnormal), our best models (SVM with RBF kernel and Random Forest) achieve accuracies around 77% and ROC-AUC up to 0.85, demonstrating that correlations capture discriminative information even without waveform-level features. In the multiclass setting, performance declines, with accuracies between 55–65% and macro ROC-AUC values around 0.80–0.82. While below deep learning results, our framework highlights a useful trade-off: correlation-based features with classical models remain computationally efficient, competitive, and highly interpretable.

We also find that correlation patterns differ clearly between normal and abnormal ECGs in ways consistent with known cardiac anatomy. This aligns with prior evidence that inter-lead relationships enhance both accuracy and interpretability in ECG analysis (Ramli and Ahmad, 2003), and extends earlier benchmarks on arrhythmia detection (Clifford et al., 2017) by showing that correlation features offer not only predictive value but also clinically meaningful insights.

In this work, we model inter-lead dependencies using simple, well-established statistical techniques to provide a transparent and effective baseline for ECG classification. Although deep neural networks achieve higher accuracy, our results show that inter-lead correlation structures already capture clinically meaningful information and can guide future methods that learn channel-wise relationships directly from raw ECG signals.

2. RELATED WORK

Deep learning has become the dominant paradigm for automatic ECG interpretation, with numerous models reporting high performance on arrhythmia detection and related tasks. For example, Zhang et al. (2021) proposed an interpretable framework for 12-lead ECGs, achieving an average F1-score of 0.81 across nine arrhythmia classes and outperforming traditional feature-based baselines. Similarly, Anand et al. (2022) introduced a convolutional model evaluated on PTB-XL, obtaining an AUC of 93.4% for multi-class cardiac abnormality detection and 95.8% accuracy on a benchmark arrhythmia dataset. Both studies incorporated SHAP analysis to explain predictions, showing that deep learning models can align decision-making with clinically meaningful ECG features.

Despite these advances, a recurring limitation is that reported accuracies are often condition-specific rather than general across all pathologies. For instance, lightweight CNN–LSTM architectures such as that of Alamatsaz et al. (2024) achieve 98.2% accuracy on the MIT-BIH arrhythmia dataset and atrial fibrillation recordings, but this performance does not generalize to broader multi-condition classification tasks. At a larger scale, Chen et al. (2019) demonstrated that CNN–LSTMs could classify six arrhythmia types from 7,000 12-lead ECGs, while the landmark study of Hannun et al. (2019) reported cardiologist-level arrhythmia detection (AUC $\approx$ 0.97, F1 = 0.837) on over 90,000 ambulatory ECGs. These results confirm the strong potential of deep neural networks but also underline their reliance on large, well-curated datasets and the difficulty of achieving uniformly high accuracy across multiple cardiac conditions.

The importance of interpretability in medical AI has been stressed across multiple clinical domains (Esteva et al., 2019). SHAP (Lundberg and Lee, 2017), in particular, has become a widely used explanation technique and has been applied to ECG models (Zhang et al., 2021; Anand et al., 2022). Broader surveys (Carvalho et al., 2019; Saranya and Subhashini, 2023; Ayano et al., 2022; Saraswat et al., 2022) highlight persistent issues, including the lack of standardized evaluation metrics, the computational cost of model explanations, and the gap between algorithmic interpretability and clinical usability. These limitations motivate the exploration of lightweight, domain-specific approaches to enhance both practicality and clinical relevance in ECG analysis.

Most existing studies treat the 12 ECG leads as parallel input channels to neural networks, without explicitly modeling correlations between them. While deep models can extract abstract features, these representations are constrained by the network architecture and may overlook physiologically meaningful inter-lead relationships. Recent work has suggested that explicitly incorporating cross-lead dependencies could yield additional diagnostic value, indicating a gap in approaches that rely solely on end-to-end feature learning. Beyond classification accuracy, several studies have also underscored the clinical importance of specific leads and their interactions. For example, Gao et al. (2020) analyzed ECG morphology during left bundle branch pacing and demonstrated how conduction mechanisms manifest across multiple leads. Other research has focused on reconstructing 12-lead ECGs from reduced-lead systems, with modern methods achieving accurate reconstructions using only three differential leads (Trobec and Tomašić, 2011; Tomašić and Trobec, 2013). Such findings reinforce the importance of preserving lead integrity and leveraging inter-lead dependencies to improve both classification performance and clinical interpretability.

Before the widespread adoption of deep learning, a variety of feature-based methods were developed for ECG analysis. Surveys such as Jambukia et al. (2015) reviewed feature extraction strategies (e.g., wavelets, morphological descriptors) and traditional classifiers, identifying artificial neural networks as among the most effective at the time. Earlier decision support systems also investigated neural models with confidence estimation; for instance, Holst et al. (1999) reported an AUC of 0.887 for anterior myocardial infarction detection, and AUC 0.995 for the most confident subset of predictions. These works illustrate the longstanding tension between accuracy, interpretability, and clinical reliability—a balance that continues to shape current research.

Large-scale benchmarks and clinical studies further highlight the importance of lead selection. Gorgels et al. (2001) showed that lead aVR, often overlooked in practice, provides critical diagnostic information in acute coronary syndromes. Similarly, the PhysioNet Computing in Cardiology Challenge (Clifford et al., 2017) demonstrated that carefully designed single-lead models for atrial fibrillation detection can reach cardiologist-level performance,

with top systems achieving F1-scores around 0.83. These findings collectively emphasize that exploiting inter-lead dependencies and optimizing lead usage are central to advancing both accuracy and interpretability in ECG classification.

3. METHODOLOGY

3.1 PTB-XL Dataset Overview

We base our study on the PTB-XL dataset (Wagner et al., 2020), one of the largest publicly accessible corpora of clinical 12-lead ECGs. It contains 21,799 recordings collected from 18,869 patients. Each trace is roughly 10 seconds in duration and is provided at two sampling frequencies (100 Hz and 500 Hz). In this work we employ the 100 Hz version, which offers an effective compromise between preserving signal quality and ensuring computational efficiency, making it suitable for lightweight model training.

Every ECG record is supplemented with structured metadata and diagnostic labels following the SCP-ECG standard. The labels are grouped into five high-level diagnostic categories: *NORM* (Normal ECG), *MI* (Myocardial Infarction), *STTC* (ST/T Change), *CD* (Conduction Disturbances), and *HYP* (Hypertrophy). This labeling scheme enables both binary (normal vs. abnormal) and multi-class learning tasks.

The dataset is not evenly distributed across categories: there are 9,514 normal ECGs and 18,251 abnormal ECGs covering one or more conditions. Within the abnormal group, 5,469 cases are labeled as MI, 5,235 as STTC, 4,898 as CD, and 2,649 as HYP. Importantly, some recordings carry multiple labels, which reflects the frequent coexistence of several cardiac pathologies in real clinical practice.

3.2 Label Mapping and Classification Setup

To enable evaluation from different clinical perspectives, we constructed two label formats. In the *binary classification* setup, all non-normal recordings were grouped under a single *abnormal* class. Samples labeled exclusively with the *NORM* superclass were assigned label 0 (normal), while all others were assigned label 1 (abnormal). In the *multi-label 5-class* setup, each sample was encoded as a 5-dimensional multi-hot vector indicating the presence or absence of each diagnostic superclass. This design reflects the fact that a single ECG recording may exhibit multiple co-occurring pathologies. In the PTB-XL dataset, 16,244 recordings (74.51%) have a single label, 4,068 (18.66%) have exactly two labels, and the remainder contain three or more concurrent labels. Such overlap illustrates the clinical complexity of ECG interpretation and motivates the need for multi-label learning frameworks.

The standard 12-lead electrocardiogram is derived from ten electrodes: six limb electrodes and six precordial (chest) electrodes. These generate six limb leads (I, II, III, aVR, aVL, aVF) and six precordial leads (V1–V6), providing a comprehensive spatial representation of cardiac electrical activity in both the frontal and horizontal planes (Dash, 2002). Figure 1 shows the conventional electrode positions and their anatomical references. Proper electrode placement is critical for reproducibility and diagnostic accuracy, as deviations can significantly alter wave morphology and lead to misinterpretation (Khunti, 2014).

Clinically, leads are often grouped into anatomical regions that project cardiac surfaces onto the body: *Inferior leads* (II, III, aVF) capture signals from the inferior wall of the left ventricle, primarily supplied by the right coronary artery. *Lateral leads* (I, aVL, V5, V6) reflect the lateral wall of the left ventricle, often linked to the circumflex artery. *Anterior leads* (V1–V4) correspond to the anterior wall and interventricular septum, supplied mainly by the left anterior descending artery (LAD).

This regional grouping is clinically relevant for localizing myocardial infarctions, conduction abnormalities, and structural changes. For example, ST-elevation confined to inferior leads typically indicates an inferior myocardial infarction, while anterior lead abnormalities are commonly associated with LAD occlusion. In the context of our correlation-based analysis, grouping leads by anatomical region enhances interpretability: disruptions in inter-lead coherence within specific regions may reflect localized pathological alterations in conduction pathways or myocardial tissue structure.

We followed the official PTB-XL split protocol (Wagner et al., 2020), using folds 1–9 (90% of the data) for model development and fold 10 (10%) as a dedicated test set. The folds are patient-wise and preserve class distributions, ensuring no patient overlap between training and test data.

We first conducted an exploratory study to assess whether statistical correlations between ECG leads alone can serve as reliable features for distinguishing normal and abnormal signals. We hypothesized that inter-lead relationships encode characteristic patterns across diagnostic categories, providing empirical motivation to emphasize channel-level dependencies in downstream modeling.

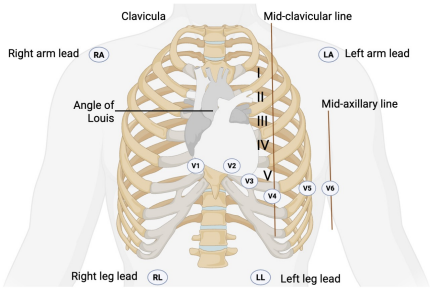


Fig. 1. Standard 12-lead ECG placement with anatomical regions highlighted (inferior, lateral, anterior).

3.3 Signal Preprocessing and Normalization

All signals were parsed from WFDB-formatted files using the official PTB-XL metadata. Each ECG was represented as a matrix of shape $(12, 1000)$, corresponding to 12 leads and 1000 time points sampled at 100 Hz.

To ensure numerical stability, we applied z-score normalization independently to each lead. The mean and standard deviation were computed across the training set and used to normalize both train and test data, preventing information leakage and ensuring consistent scaling.

3.4 Extracted Features

Formally, consider an ECG sample represented as $X \in \mathbb{R}^{C\times T}$, where $C = 12$ denotes the number of leads and $T = 1000$ the number of time steps. For any pair of leads, let

$$x = (x_1, \ldots, x_T), \quad y = (y_1, \ldots, y_T)$$

denote the corresponding time series. The correlation coefficients between two leads were computed as follows:

Pearson Coefficients. The Pearson coefficients coefficient measures the linear relationship between two signals:

$$\rho_{\mathrm{P}}(x, y) = \frac{\sum_{i=1}^{T}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{T}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{T}(y_i - \bar{y})^2}},$$

where $\bar{x}$ and $\bar{y}$ are the mean values of x and y, respectively.

Spearman Coefficients. The Spearman coefficients is a rank-based version of Pearson's correlation, which captures monotonic rather than strictly linear relationships:

$$\rho_{\mathrm{S}}(x, y) = \rho_{\mathrm{P}}(\mathrm{rank}(x), \mathrm{rank}(y)).$$

Here, $\mathrm{rank}(x)$ replaces each value x_i with its rank among $(x_1, \ldots, x_T)$, and analogously for y. Ties are handled by averaging ranks.

Kendall's Tau. Kendall's τ quantifies the ordinal association between two variables by counting concordant and discordant pairs:

$$\tau_{\mathrm{K}}(x, y) = \frac{2}{T(T-1)}\sum_{i<j}\mathrm{sign}(x_i - x_j)\cdot\mathrm{sign}(y_i - y_j).$$

Two pairs (x_i, y_i) and (x_j, y_j) are concordant if the signs of their differences agree, and discordant otherwise.

Since there are $\binom{12}{2} = 66$ possible lead pairs, each metric produces a 66-dimensional feature vector. Concatenating Pearson, Spearman, and Kendall yields a 198-dimensional descriptor per recording, which serves as the feature representation for downstream analyses.

Compared to the original feature space of size 12×1000, we now work with only 198 features and show that the intrinsic information they capture leads to promising results in a classification task.

3.5 Classification Models

Having established that inter-lead correlations differ systematically across diagnostic categories, we assessed their predictive value in a supervised setting. We trained classical statistical and machine learning classifiers on the handcrafted 198-dimensional correlation descriptors: logistic regression, SVM (with linear and radial basis function kernels), Gradient Boosting, and Random Forest. Experiments were conducted on PTB-XL under two scenarios: (i) binary classification of normal vs. abnormal ECGs, and (ii) multiclass classification into the five diagnostic superclasses. For each experiment, we evaluated model performance using accuracy, precision, recall, F1 score, and ROC AUC.

3.6 Explainable AI Methods

To interpret the behavior of machine learning models, we employed two widely used explainable AI techniques: SHAP values and permutation feature importance. These methods provide complementary perspectives on feature relevance, helping to understand which aspects of the input data drive model predictions.

- **SHAP values** quantify the contribution of each feature to the model's prediction for individual samples. By aggregating these contributions across the dataset, we obtain global rankings of the most influential features.
- **Permutation feature importance** assesses the effect of each feature on model performance by randomly shuffling its values and measuring the resulting decrease in predictive accuracy. This provides a model-agnostic estimate of global feature relevance.

Using both approaches enables a more robust and interpretable understanding of model decision-making, combining fine-grained, sample-level explanations (SHAP) with overall, model-level importance estimates (permutation importance).

4. EXPERIMENTS

4.1 Binary Case: Normal vs Abnormal

Before training classifiers, we examined whether correlation patterns differ systematically between normal and pathological ECGs. Figure 2 shows representative heatmaps of Pearson, Spearman, and Kendall coefficients for normal (left) and abnormal (right) recordings. Normal ECGs display strong and consistent inter-lead coherence, whereas abnormal signals show disrupted or weakened dependencies. These structural differences suggest that even simple correlation measures capture meaningful diagnostic variation.

As detailed in Section 3, each ECG is represented by 198 correlation features capturing linear, monotonic, and ordinal dependencies across all lead pairs.

4.2 Multiclass Extension

While binary separation provides an initial validation, clinical diagnosis often requires distinguishing between multiple cardiac conditions. We therefore extended the analysis to the five diagnostic superclasses in PTB-XL: Normal (Class 0), Myocardial Infarction (Class 1), ST/T Changes (Class 2), Conduction Disturbance (Class 3), and Hypertrophy (Class 4).

Figure 3 presents the average pairwise correlation matrices for each class, computed separately using Pearson, Spearman, and Kendall coefficients. Distinctive class-specific structures are visible: Class 1 (Myocardial Infarction) shows attenuated correlations in inferior leads, consistent with regional ischemic effects; Class 2 (ST/T Changes) maintains moderate coherence but with subtler perturbations; Class 3 (Conduction Disturbance) displays asymmetrical patterns that reflect altered conduction pathways; and Class 4 (Hypertrophy) exhibits diffuse changes likely tied to chamber enlargement. These findings highlight the diagnostic value of correlation structures beyond a simple normal/abnormal dichotomy.

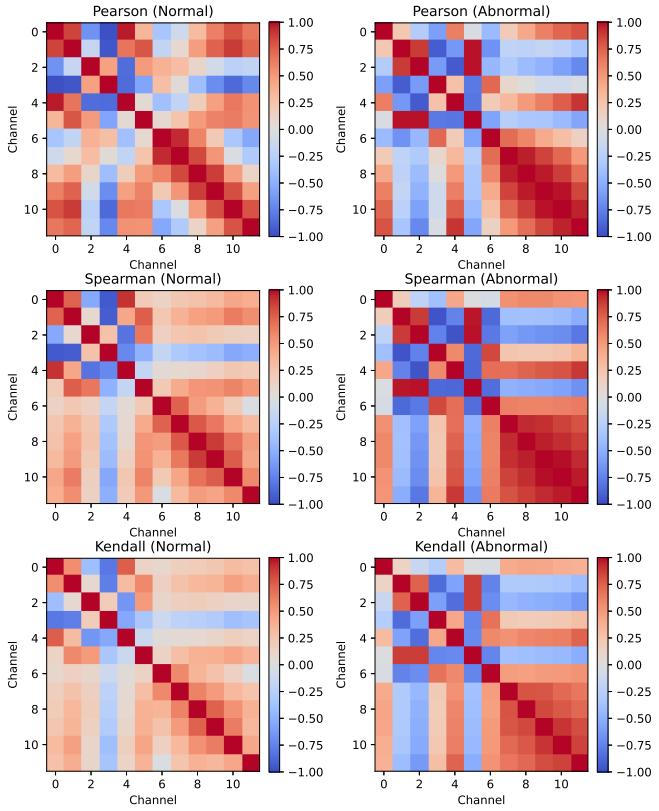


Fig. 2. Pairwise correlation matrices computed using Pearson (top), Spearman (middle), and Kendall (bottom) coefficients for normal (left) and abnormal (right) ECG samples. Normal signals show consistent inter-lead dependencies, while abnormal signals exhibit disrupted correlation structures.

A comparison of Figures 2 and 3 reveals both consistent and pathology-specific trends. In the binary case, the contrast between normal and abnormal is visually stark: coherence is high in normal ECGs but weakened in abnormal ones. In the multiclass setting, the disruptions vary in location and structure, aligning with known anatomical and pathological mechanisms. These results reinforce the clinical plausibility of inter-lead correlation features and motivate their use in interpretable ECG classification.

4.3 Performance and Ablations

Tables 1 and 2 summarize the performance of classical models on binary and multiclass ECG classification tasks under full, ablation, and single-correlation settings. Since the dataset is imbalanced, we report weighted performance metrics, where each class is weighted according to its relative frequency.

In the **binary scenario**, all classifiers achieved solid performance, with F1-scores typically in the 77–80 range and ROC-AUC values consistently above 0.84. The strongest overall results were obtained by the random forest, which reached an F1-score of **80.28** and a ROC-AUC of **0.8430**, closely followed by gradient boosting with an F1-score of **79.41**. Logistic regression and SVM (RBF) remained highly competitive: their F1-scores (**78.25** and **78.43**) and ROC-AUC values (**0.8514** and **0.8568**) place them among the best-performing models. Accuracy values also clustered in a narrow band across models (typically **75–77%**), reinforcing the consistency of the feature space. These results confirm that correlation-based representations provide a strong, nearly linearly separable feature space for binary ECG classification.

The **ablation analysis** showed that removing Pearson correlation consistently led to the largest performance drops across all models, with decreases of up to 3–4 points in both F1-score and ROC-AUC. In contrast, excluding Spearman or Kendall correlations produced only minor degradations, suggesting that Pearson captures the most critical information for discriminating between normal and abnormal ECGs. Interestingly, when using a *single correlation measure* in isolation, performance remained reasonable but was noticeably lower than the full-feature configuration. Among the single-measure settings, Spearman tended to yield slightly better performance than Pearson or Kendall, though none matched the effectiveness of combining all three.

As shown in Table 1, the proposed lead-correlation based models achieve an accuracy of up to **77.02%** for binary ECG classification on PTB-XL, specifically with the random forest classifier. Logistic regression, SVM (RBF), and gradient boosting all achieve accuracies between **76.3%** and **76.9%**, further demonstrating the stability of the approach across different learning algorithms. Remarkably, this performance is comparable to that of recent deep learning models trained end-to-end on raw ECG signals, which typically achieve accuracies in the 78–85% range Strodthoff et al. (2020); Wagner et al. (2020). This demonstrates that inter-lead correlation features capture a substantial portion of the diagnostic signal, enabling competitive results with significantly lower model complexity.

In the more challenging **multiclass setting**, overall performance declined across all models, as expected given the finer distinctions between related diagnostic categories (e.g., myocardial infarction vs. ST/T changes). Logistic regression remained the strongest model, achieving an accuracy of **65.27%**, a weighted F1-score of **60.88**, and a ROC-AUC of **0.8093**. Gradient boosting achieved similar results, with an accuracy of **64.73%** and F1-score of **60.40**. SVM (RBF) also performed competitively, achieving a ROC-AUC of **0.8182**, the highest among the models. In contrast, linear SVM and random forest produced lower F1-scores (**56.69** and **58.32**), indicating that the multiclass task benefits from either nonlinear decision boundaries or more flexible ensemble models.

The **multiclass ablation results** reinforced the dominant role of Pearson correlation: removing it caused the largest degradation in performance for all models (e.g., logistic regression F1 dropping from **60.88** to **54.88**, and its ROC-AUC decreasing from **0.8093** to **0.7733**). Spearman and Kendall again had a smaller impact, indicating some redundancy between these rank-based measures. When each correlation type was used alone, performance declined substantially across all metrics. For instance, logistic regression using only Kendall correlation yielded an F1-score of just **30.60**, and only Spearman correlation increased it slightly to **31.40**. F1-scores dropped by more than 20 points compared to the full-feature setup, highlighting that

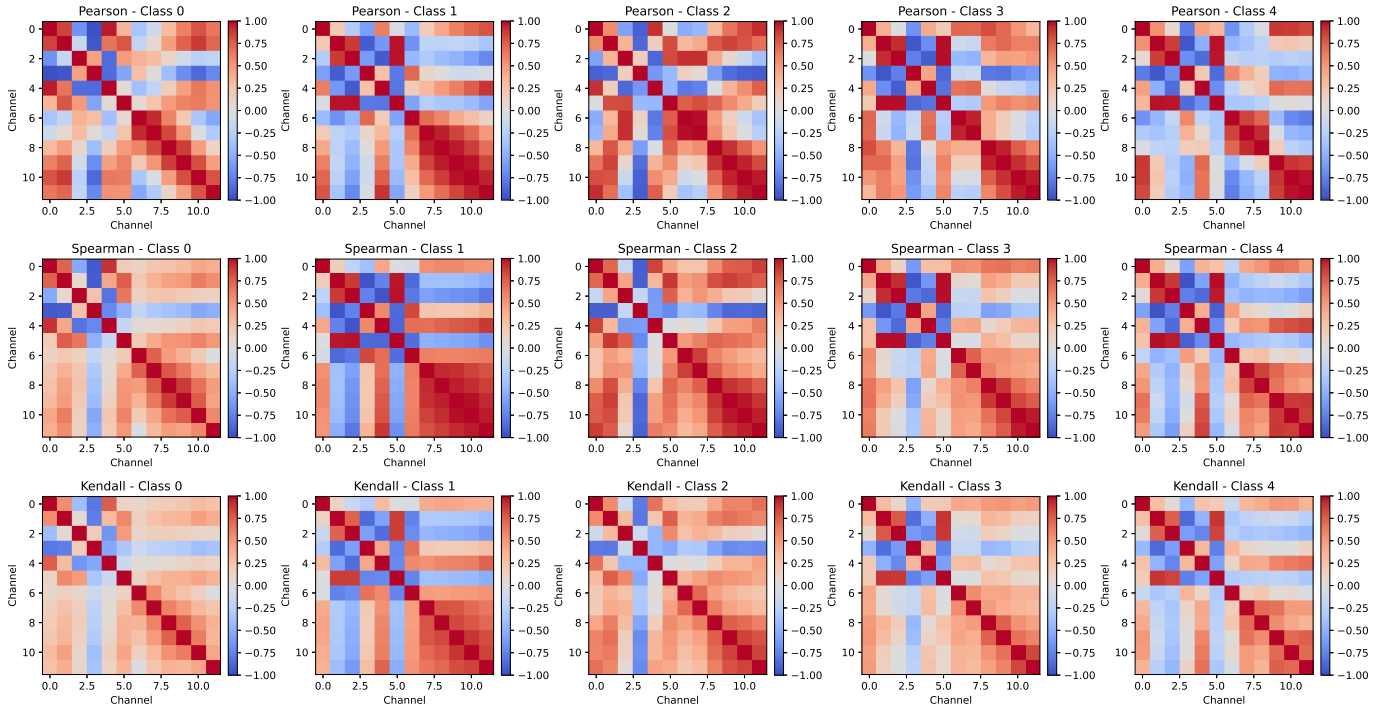


Fig. 3. Average pairwise correlation matrices for the five diagnostic superclasses in PTB-XL, computed with Pearson (top), Spearman (middle), and Kendall (bottom). Each column corresponds to one class: Normal (0), Myocardial Infarction (1), ST/T Changes (2), Conduction Disturbance (3), and Hypertrophy (4). Class-specific patterns are visible, particularly in anterior, inferior, and lateral leads.

no single correlation measure alone is sufficient to capture the complexity of multiclass ECG patterns, though Spearman again showed a slight advantage among the three.

Overall, these results demonstrate that correlation-based features are effective for both binary and multiclass ECG classification tasks, with Pearson correlation emerging as the single most informative component. At the same time, combining multiple correlation measures provides complementary information that leads to consistently better performance than using any one measure alone.

4.4 Confusion Matrix Analyses

To better understand the sources of error, we examined representative confusion matrices for a subset of classifiers, as shown in Figures 4 and 5.

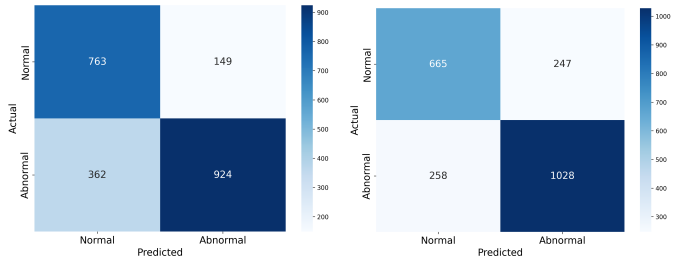
In the **binary setting**, the separation between normal and abnormal ECGs is generally strong across models. Logistic regression and SVM (RBF) provide a good balance between true positives and true negatives, consistent with their high F1-scores. For example, logistic regression and SVM (RBF) correctly identify around 920–925 abnormal cases but still miss roughly 360, whereas gradient boosting and random forest recover more abnormal cases (around 1000 true positives) at the cost of higher false positive rates. Linear SVM shows lower sensitivity, misclassifying more abnormal recordings. These results illustrate a trade-off: linear models provide more balanced decision boundaries, while ensemble methods emphasize sensitivity at the cost of specificity.

Across all models, the dominant normal class (Class 0) is recognized reliably, with logistic regression and gradient boosting exceeding 840 correct predictions. However, minority classes show significant confusion. Logistic regression provides relatively balanced recall: for example, it captures 80 cases of Class 2 (hypertrophy) and 68 of Class 3 (myocardial infarction), outperforming random forest, which collapses most minority predictions into the normal class. SVMs (linear and RBF) demonstrate stronger off-diagonal patterns, with frequent misclassification of pathological categories into one another; for example, Class 1 (ST/T changes) is often confused with Classes 2 and 3, and hypertrophy is frequently misattributed to infarction. Gradient boosting offers performance similar to logistic regression, with more stable recall across Classes 1–3, though it still struggles with the smallest class (Class 4).

Overall, these confusion matrices reinforce two key findings. First, in the binary setting, correlation-based features are nearly linearly separable, enabling simple models to achieve robust performance with well-calibrated trade-offs between sensitivity and specificity. Second, in the multiclass task, class imbalance and overlapping morphologies create substantial challenges, leading to systematic confusion among pathological categories. While handcrafted correlation features capture meaningful diagnostic structure, their limitations in fine-grained discrimination suggest the need for richer representations or hybrid feature learning approaches.

Table 1. Binary ECG classification — performance of classical models using correlation-based features under full, ablation, and separated settings.

Model	Task	Setting	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	Binary	All Features	76.66	86.02	71.77	78.25	0.8514
		No Kendall	76.48	85.63	71.85	78.14	0.8452
		No Pearson	74.98	84.46	70.14	76.64	0.8303
		No Spearman	75.75	84.83	71.31	77.48	0.8432
		Only Kendall	72.79	82.58	67.81	74.47	0.8019
		Only Pearson	72.02	82.29	66.49	73.55	0.8017
		Only Spearman	73.61	84.21	67.57	74.98	0.8147
SVM (Linear)	Binary	All Features	75.84	86.69	69.36	77.06	0.8473
		No Kendall	75.43	85.80	69.52	76.80	0.8436
		No Pearson	74.48	85.57	67.81	75.66	0.8283
		No Spearman	75.07	85.83	68.74	76.34	0.8408
		Only Kendall	72.38	84.12	65.09	73.39	0.7985
		Only Pearson	71.52	83.47	64.00	72.45	0.7976
		Only Spearman	72.88	85.20	64.93	73.70	0.8127
SVM (RBF)	Binary	All Features	76.93	86.57	71.70	78.43	0.8568
		No Kendall	76.93	86.50	71.77	78.45	0.8568
		No Pearson	74.07	85.03	67.57	75.30	0.8290
		No Spearman	76.75	86.52	71.38	78.23	0.8566
		Only Kendall	72.93	84.11	66.25	74.12	0.8219
		Only Pearson	71.70	83.54	64.31	72.67	0.8079
		Only Spearman	73.52	84.71	66.80	74.70	0.8282
Gradient Boosting	Binary	All Features	76.34	80.89	77.99	79.41	0.8472
		No Kendall	76.48	81.34	77.60	79.43	0.8465
		No Pearson	72.84	77.80	74.96	76.36	0.8107
		No Spearman	75.80	80.45	77.45	78.92	0.8428
		Only Kendall	72.52	77.72	74.34	75.99	0.8014
		Only Pearson	73.07	78.07	75.04	76.53	0.8131
		Only Spearman	73.11	78.41	74.57	76.44	0.8102
Random Forest	Binary	All Features	77.02	80.63	79.94	80.28	0.8430
		No Kendall	76.02	80.00	78.69	79.34	0.8440
		No Pearson	72.84	77.76	75.04	76.38	0.8141
		No Spearman	76.02	79.95	78.77	79.36	0.8423
		Only Kendall	72.25	77.35	74.34	75.81	0.8086
		Only Pearson	73.34	78.32	75.27	76.76	0.8121
		Only Spearman	73.75	78.61	75.74	77.15	0.8159



(a) Binary — Logistic Regression (b) Binary — Random Forest

Fig. 4. Confusion matrices for the binary classification task using correlation-based features. Logistic Regression provides a balanced trade-off between sensitivity and specificity, while Random Forest emphasizes sensitivity with more false positives.

5. EXPLAINABILITY

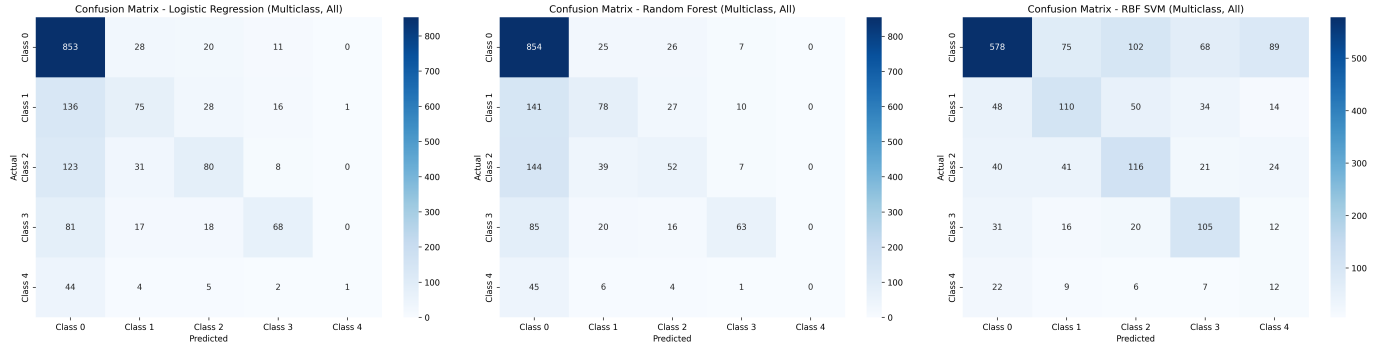
5.1 Feature Importance and Interpretability

Beyond evaluating predictive performance, an important goal of this study was to understand *which inter-lead correlations most strongly influenced model decisions*, and whether these patterns were physiologically meaningful. To this end, we applied SHAP values and permutation feature importance analyses to the trained classifiers, focusing on both the binary and multiclass classification scenarios.

For the binary classification task, we concentrated on *logistic regression* and *gradient boosting* models, which exhibited the best overall performance and represent two distinct modeling paradigms (linear vs. ensemble-based). For each model, we ranked all 198 handcrafted correlation features and visualized the top 20 using SHAP summary plots (Figures 6a–6b). The most influential features consistently involved correlations between anatomically adjacent or physiologically related leads—such as V3–V4, I–aVL, and II–aVF. This matches established clinical knowledge: under normal conditions, neighboring leads show synchro-

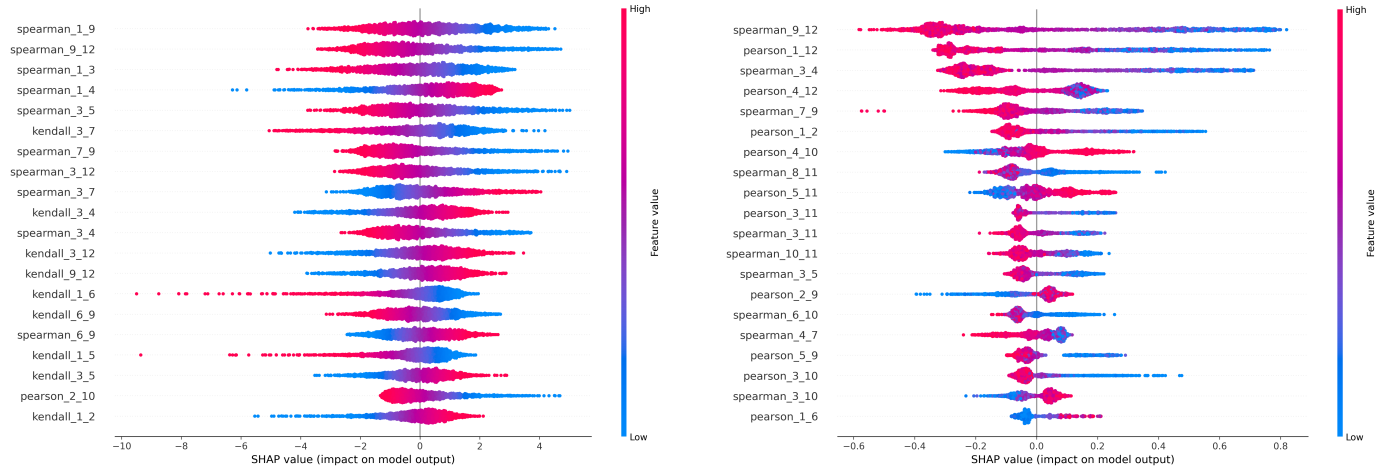
Table 2. Multiclass ECG classification — performance of classical models using correlation-based features under full, ablation, and separated settings. Precision/Recall/F1 use macro-averages; ROC-AUC is one-vs-rest.

Model	Task	Setting	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	Multiclass	All Features	65.27	62.31	65.27	60.88	0.8093
		No Kendall	63.76	58.45	63.76	58.81	0.8002
		No Pearson	60.91	54.75	60.91	54.88	0.7733
		No Spearman	63.82	58.67	63.82	58.74	0.7977
		Only Kendall	59.82	39.91	29.82	30.60	0.7414
		Only Pearson	59.15	34.83	29.54	28.83	0.7423
		Only Spearman	60.18	39.52	30.50	31.40	0.7519
SVM (Linear)	Multiclass	All Features	53.76	61.80	53.76	56.69	0.8114
		No Kendall	52.73	61.88	52.73	56.01	0.8026
		No Pearson	50.12	59.65	50.12	53.52	0.7740
		No Spearman	51.33	60.77	51.33	54.69	0.7994
		Only Kendall	47.27	35.56	36.01	34.50	0.7435
		Only Pearson	45.94	35.90	40.00	35.51	0.7411
		Only Spearman	48.91	37.68	39.57	36.94	0.7547
SVM (RBF)	Multiclass	All Features	55.82	62.27	55.82	58.24	0.8182
		No Kendall	56.24	62.74	56.24	58.68	0.8206
		No Pearson	49.52	57.92	49.52	52.49	0.7687
		No Spearman	56.42	62.80	56.42	58.81	0.8189
		Only Kendall	47.88	37.21	38.84	36.10	0.7613
		Only Pearson	46.55	37.31	41.04	36.91	0.7602
		Only Spearman	49.52	38.11	40.37	37.43	0.7686
Gradient Boosting	Multiclass	All Features	64.73	60.50	64.73	60.40	0.8062
		No Kendall	63.94	59.30	63.94	59.53	0.8051
		No Pearson	61.45	56.45	61.45	55.90	0.7526
		No Spearman	64.12	59.42	64.12	59.60	0.8043
		Only Kendall	60.67	40.88	32.47	33.76	0.7446
		Only Pearson	61.03	40.02	32.70	33.14	0.7581
		Only Spearman	60.79	40.49	32.28	33.47	0.7508
Random Forest	Multiclass	All Features	63.45	58.49	63.45	58.32	0.7914
		No Kendall	63.33	58.28	63.33	57.88	0.7932
		No Pearson	61.39	56.05	61.39	54.97	0.7413
		No Spearman	63.27	58.23	63.27	58.06	0.7930
		Only Kendall	60.91	42.57	32.36	33.84	0.7408
		Only Pearson	60.91	41.11	32.61	33.20	0.7487
		Only Spearman	61.33	42.80	32.61	34.14	0.7495



(a) Multiclass — Logistic Regression (b) Multiclass — Random Forest (c) Multiclass — SVM (RBF)

Fig. 5. Confusion matrices for the multiclass classification task. Logistic Regression yields comparatively balanced recall across minority classes; Random Forest tends to collapse minority classes into the majority; SVM (RBF) shows stronger off-diagonal confusion among pathological categories.



(a) Binary — Logistic Regression

(b) Binary — Gradient Boosting

Fig. 6. Comparison of SHAP summary plots for binary classification: (a) logistic regression and (b) gradient boosting. Each dot is a sample's SHAP value for a feature (color encodes feature value, blue=low to red=high), with features ordered by mean |SHAP|. Positive SHAP values push predictions toward the positive class.

nized waveforms, while pathological changes typically disrupt these dependencies.

Permutation feature importance analyses provided convergent evidence. Shuffling individual features resulted in the largest performance drops for the same set of lead pairs identified by SHAP, further validating their predictive role. Interestingly, a number of less conventional correlations, such as Spearman(3–5) or Kendall(2–4), also appeared among the top-ranked features. This suggests that diagnostically relevant information may also arise from lead combinations that are not commonly emphasized in routine clinical interpretation.

In the multiclass classification task, feature importance patterns became more distributed. Logistic regression tended to highlight limb–chest interactions (e.g., I with V3–V5), which are frequently affected in conditions like ischemia and hypertrophy. Ensemble methods such as Gradient Boosting and Random Forest identified similar patterns but with greater diversity among the top features, including lead pairs such as Spearman(V9, V12), Pearson(V1, V12), and Spearman(V7, V9). These results indicate that more complex models can exploit a broader range of correlations, possibly reflecting the increased heterogeneity across multiple diagnostic categories.

Taken together, these analyses reveal a clear and clinically interpretable structure in the feature space. Correlations between physiologically related leads consistently emerge as key predictors, lending support to the medical relevance of the proposed feature representation. At the same time, the presence of less conventional lead interactions points to the potential for machine learning models to uncover patterns that may be diagnostically useful but underexplored in traditional ECG interpretation.

6. CONCLUSION

Our findings confirm that inter-lead coherence is systematically disrupted by pathology in anatomically consistent ways. Despite their simplicity, inter-lead correlations capture clinically meaningful dependencies and, when combined with classical models and SHAP analysis, provide a transparent and effective baseline for explainable ECG classification on PTB-XL. The results suggest that correlation-based features can serve both as lightweight clinical decision aids and as interpretable modules complementing deep learning models.

Beyond this work, we are already extending these ideas in two directions: (i) a frequency-domain analysis using saliency maps to highlight abnormal spectral patterns, and (ii) a deep learning framework designed to learn enhanced inter-lead correlation representations directly from raw ECG signals. These ongoing studies further explore temporal–spectral dependencies and learned correlation structures, and will help clarify how best to leverage inter-lead information in ECG analysis

REFERENCES

Alamatsaz, N., Tabatabaei, L., Yazdchi, M., Payan, H., Alamatsaz, N., and Nasimi, F. (2024). A lightweight hybrid cnn-lstm explainable model for ecg-based arrhythmia detection. *Biomedical Signal Processing and Control*, 90, 105884.

Anand, A., Kadian, T., Shetty, M.K., and Gupta, A. (2022). Explainable ai decision model for ecg data of cardiac disorders. *Biomedical Signal Processing and Control*, 75, 103584.

Ayano, Y.M., Schwenker, F., Dufera, B.D., and Debelee, T.G. (2022). Interpretable machine learning techniques in ecg-based heart disease classification: a systematic review. *Diagnostics*, 13(1), 111.

Carvalho, D.V., Pereira, E.M., and Cardoso, J.S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 832.

Chen, Y.J., Liu, C.L., Tseng, V.S., Hu, Y.F., and Chen, S.A. (2019). Large-scale classification of 12-lead ecg with deep learning. In *2019 IEEE EMBS international conference on biomedical & health informatics (BHI)*, 1–4. IEEE.

Clifford, G.D., Liu, C., Moody, B., Lehman, L.w.H., Silva, I., Li, Q., Johnson, A.E., and Mark, R.G. (2017). Af

classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017. In *2017 computing in cardiology (CinC)*, 1–4. IEEE.

Dash, P. (2002). Electrocardiogram monitoring. *Indian Journal of Anaesthesia*, 46(4), 251–260.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine*, 25(1), 24–29.

Gao, M.Y., Tian, Y., Shi, L., Wang, Y.J., Xie, B.Q., Qi, J., Zeng, L.J., Li, X.X., Yang, X.C., and Liu, X.P. (2020). Electrocardiographic morphology during left bundle branch area pacing: characteristics, underlying mechanisms, and clinical implications. *Pacing and Clinical Electrophysiology*, 43(3), 297–307.

Gorgels, A.P., Engelen, D., and Wellens, H.J. (2001). Lead avr, a mostly ignored but very valuable lead in clinical electrocardiography.

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.

Holst, H., Ohlsson, M., Peterson, C., and Edenbrandt, L. (1999). A confident decision support system for interpreting electrocardiograms. *Clinical Physiology*, 19(5), 410–418.

Jambukia, S.H., Dabhi, V.K., and Prajapati, H.B. (2015). Classification of ecg signals using machine learning techniques: A survey. In *2015 international conference on advances in computer engineering and applications*, 714–721. IEEE.

Khunti, K. (2014). Accurate interpretation of the 12-lead ecg electrode placement: A systematic review. *Health Education Journal*, 73(5), 610–623.

Lundberg, S.M. and Lee, S.I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Pipberger, H.V., Bialek, S.M., Perloff, J.K., and Schnaper, H.W. (1961). Correlation of clinical information in the standard 12-lead ecg and in a corrected orthogonal 3-lead ecg. *American Heart Journal*, 61(1), 34–43.

Ramli, A.B. and Ahmad, P.A. (2003). Correlation analysis for abnormal ecg signal features extraction. In *4th National Conference of Telecommunication Technology, 2003. NCTT 2003 Proceedings.*, 232–237. IEEE.

Saranya, A. and Subhashini, R. (2023). A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision analytics journal*, 7, 100230.

Saraswat, D., Bhattacharya, P., Verma, A., Prasad, V.K., Tanwar, S., Sharma, G., Bokoro, P.N., and Sharma, R. (2022). Explainable ai for healthcare 5.0: opportunities and challenges. *IEEe Access*, 10, 84486–84517.

Strodthoff, N., Wagner, P., Schaeffter, T., and Samek, W. (2020). Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE journal of biomedical and health informatics*, 25(5), 1519–1528.

Tomašić, I. and Trobec, R. (2013). Electrocardiographic systems with reduced numbers of leads—synthesis of the 12-lead ecg. *IEEE reviews in biomedical engineering*, 7, 126–142.

Trobec, R. and Tomašić, I. (2011). Synthesis of the 12-lead electrocardiogram from differential leads. *IEEE transactions on information technology in biomedicine*, 15(4), 615–621.

Wagner, P., Strodthoff, N., Bousseljot, R.D., Kreiseler, D., Lunze, F.I., Samek, W., and Schaeffter, T. (2020). Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1), 1–15.

Zhang, D., Yang, S., Yuan, X., and Zhang, P. (2021). Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram. *Iscience*, 24(4).