Research on Engineering Structures & Materials
www.jresm.org
Research Group

Research Article

Deep learning-based prediction of myocardial ischemia from ECG signals using a fully convolutional network and explainable AI

Fidha Ezzaldin Aslan [1,a], Muthana R. Alsultan [*,2,b], Shivan A. Mohammed [3,c]

[1]*Department of Administrative and Financial Affairs, University of Telafer, Nineveh, Iraq*
[2]*Department of Reconstruction and Projects, University of Baghdad, Baghdad, Iraq*
[3]*Department of AI Engineering, University of Zakho, Duhok, Iraq*

Article Info

Article History:

Received 21 Jul 2026

Accepted 21 Sep 2026

Keywords:

Myocardial ischemia;
Deep learning (DL);
Electrocardiogram (ECG);
Fully Convolutional Network (FCN);
Local Interpretable Model-Agnostic Explanations (LIME);
Explainable artificial intelligence (XAI);
PTB-XL;
External validation

Abstract

Myocardial ischemia is a major cause of illness and death worldwide, and its early detection remains difficult. The electrocardiogram (ECG) is the most widely used test for ischemia, but manual interpretation is time-consuming and subject to inter-observer variability. This study develops a deep learning framework based on a Fully Convolutional Network (FCN) that classifies myocardial ischemia from 12-lead ECGs and explains its decisions with Local Interpretable Model-Agnostic Explanations (LIME). The public PTB-XL database was used, comprising 11,018 ECGs from 10,135 patients (2,441 ischemia and 8,577 normal). Signals were cleaned, segmented into beats and summarized as lead-wise median beats. Training, validation and independent test sets were separated at the patient level, performance was estimated over ten patient-level bootstrap iterations, and the model was externally validated on 2,033 ECGs from the Georgia 12-lead ECG Challenge database. On the independent test set, the FCN achieved an accuracy of 97.82% ± 0.19%, precision of 95.98%, recall (sensitivity) of 94.23%, specificity of 98.86%, F1-score of 95.10% and an area under the receiver operating characteristic curve (AUC) of 0.995 (95% CI 0.995–0.996). In external validation, the AUC was 0.959 and the accuracy was 90.31%. The FCN was compared with an attention-based encoder, a simple convolutional neural network, a Transformer encoder, random forest, support vector machine and XGBoost under the same protocol. LIME explanations placed 64.7% of the positive importance of ischemic ECGs in the ST–T region (chance level 48%) and 57.3% in the cardiologist-annotated ischemia territory (chance level 47.9%), and removing the most important segments reduced the predicted ischemia probability significantly more than removing random segments. The framework therefore combines competitive performance with explanations that are quantitatively consistent with clinical knowledge. Prospective, multicenter validation is required before clinical use.

© 2026 MIM Research Group. All rights reserved.

1. Introduction

Cardiovascular disease (CVD) remains the leading cause of death worldwide [1], and the early detection of conditions such as myocardial ischemia can reduce morbidity and mortality [2]. Myocardial ischemia occurs when the heart muscle does not receive enough blood, usually because of blockage of the coronary arteries. The first diagnostic tool used by physicians to identify myocardial ischemia non-invasively is Electrocardiogram (ECG) because it detects heart electrical activity [3]. Determining ischemia from ECG requires complex interpretation skills and extensive experience, minor variations in ST-segment and T-wave alterations could be missed leading to

*Corresponding author: muth85cnet@uobaghdad.edu.iq
[a]orcid.org/0009-0002-3796-7760; [b]orcid.org/0009-0009-8801-1209; [c]orcid.org/0000-0001-5713-8519
DOI: http://dx.doi.org/10.17515/resm2026-1990ic0721rs

delayed or misclassification [4]. Availability of large-scale ECG datasets is becoming less of an issue with advancing technology. Automated diagnosis with ML is now possible [5]. DL models are especially promising for identifying disease-specific patterns. They can take multivariate time series data such as a 12-lead ECG as input, and learn subtle nonlinear features that may be imperceptible to human readers. However, ECG signals can be noisy and differ between patients and devices, making robust pre-processing necessary for accurate classification [6]. Also, while accuracy is important for a clinically useful model, prediction results also need to be interpretable by clinicians such that they can verify them.

Computer-aided diagnostic systems may aid timely recognition of myocardial ischemia and lighten clinicians' load, especially where there are limited resources. By integrating deep learning's discrimination abilities with the explainability needed in healthcare settings, automated diagnosis could enhance recognition and treatment of ischemic heart disease. The research problem solved by the proposed algorithm/system/tool/model is automatic detection of myocardial ischemia from 12-lead ECG data with high robustness to noise and generalizability across patients, while remaining easily interpretable [5]. Current methods lack in one or more of these aspects, making their use in clinical practice slow to be realized since robustness and interpretability are key. Another problem with current models is that they take in the raw ECG as a whole and fail to account for localized features within leads that may be useful for detecting ischemia. The approaches proposed so far to solve this problem have been based on convolutional neural networks (CNNs), recurrent neural networks (RNNs) and recently Transformer-based architectures [7, 8]. These methods perform very well; however, most of them lack explainability [9, 10]. Models that cannot explain their predictions are hard for clinicians to trust and therefore difficult to implement into clinical practice. Approaches like saliency maps, gradient-weighted class activation mapping (Grad-CAM), SHapley Additive exPlanations (SHAP) and LIME have been suggested to solve this problem [11, 12, 13]. However, a recent review found that XAI method evaluation for ECG is still sparse and seldomly quantified [10]. Furthermore, most research works use small datasets, record-level splits rather than patient-level and do not perform external validation. This leads to concerns about the reliability of the presented results [14]. This study presents a framework that combines a Fully Convolutional Network (FCN) [15] applied to lead-wise median beats with Local Interpretable Model-Agnostic Explanations (LIME) [16]. The FCN and LIME are established techniques, and their combination alone is not claimed as a methodological novelty; FCN-based ECG classifiers with attribution maps have been reported previously for myocardial infarction [11]. The contribution of this work lies in how the framework is designed, explained and validated for myocardial ischemia, as summarized below.

Beat-level, lead-by-segment explanations. LIME is applied to 300 interpretable features (12 leads × 25 temporal segments of the median beat), so each explanation directly indicates which lead and which part of the cardiac cycle (P-wave, QRS complex, ST-segment or T-wave) supports the prediction, unlike sample-level saliency maps that are difficult to read clinically. Quantitative validation of the explanations. The explanations are compared with the cardiologist-assigned ischemia territories of PTB-XL (anterior, anteroseptal, lateral, anterolateral, inferior and inferolateral), and their temporal localization, reproducibility across random seeds and faithfulness (deletion test) are measured, which addresses the lack of quantitative XAI evaluation identified in [10]. Rigorous and reproducible evaluation. All experiments use public data [17, 18, 19] with patient-level separation of training, validation and independent test sets, ten patient-level bootstrap iterations with confidence intervals, statistical significance tests and external validation on a dataset from a different country and hospital system.

Controlled benchmark. Under an identical protocol, the FCN is compared with an attention-based encoder, a simple CNN, a Transformer encoder, random forest, support vector machine (SVM) and extreme gradient boosting (XGBoost), including computational cost, so that the performance of the FCN relative to alternative models is established by statistical evidence rather than assumed.

The research objectives of this study are as follows:

- Pre-process 12-lead ECG signals by removing noise and extracting lead-wise median beats to provide a reliable input for deep learning models.

- Design and evaluate deep learning models, including an FCN, an attention-based encoder, a simple CNN and a Transformer, together with classical ML baselines, using patient-level internal testing and external validation.
- Integrate LIME into the framework and quantitatively verify that its explanations agree with clinically relevant ECG regions and cardiologist-annotated ischemia territories.

The proposed solution fills voids in state-of-the-art automated diagnosis systems by not only providing accuracy but also interpretable explanations that can be verified, a necessity for clinical settings. A deployable system capable of intelligently prioritizing patients is especially useful in low-resource environments where cardiologists may be scarce. The rest of this paper is organized as follows: Section 2 reviews related literature. Section 3 outlines our methodology. This includes pre-processing, model architecture, evaluation, and explainability. Section 4 details the datasets used and the experimental setup. Section 5 discusses the results. We then discuss the clinical significance of our findings and limitations in Section 6. We conclude our paper and propose future work in Section 7.

2. Literature Review

Singh et al. summarized all ECG parameters that have been reported to predict cardiac mortality [20]. Parameters used for the diagnosis and prognosis of CVDs and their relationship to cardiac mortality are presented in Table 3. Prolongation of QT interval, QRS duration, T-wave inversions and bundle branch blocks (BBB) are all commonly associated with increased cardiac mortality risk. An increase in QRS duration by 10 ms was associated with an 18% increase in cardiovascular risk. Deep terminal negativity of the P wave in V1 correlated with cardiac mortality risk from atrial fibrillation/flutter and ischemic heart disease. ST-segment depression is a predictor of myocardial infarction. The spatial QRS-T angle was a predictor of sudden cardiac death. Several studies have shown promise in ECG parameters used simultaneously to stratify sudden cardiac death risk, however there is no gold standard ECG measure yet. Advancements in computational approaches for ECG analysis may provide more robust measures for clinical use. Sarıçam et al. [14] proposed an algorithm based on ECG findings for predicting the culprit vessel territory of acute MI including lateral ventricle. ST depression in lead V2 had a PPV and NPV of 100% and 90%, respectively for prediction of obtuse marginal artery occlusion in a retrospective multicenter study of 74 patients with STEMI and isolated lateral MI. The presence of hyperacute T wave (10mm) in lead V2 and 2mm ST depression in lead III accurately predicted occlusion of large diagonal branch (D1) with PPV and NPV of 98% and 100%, respectively.

Al-Zaiti et al. [21] trained an ML model to diagnose and risk-stratify occlusion myocardial infarction (OMI). This tool can help clinicians reach diagnoses when patients present with chest pain but lack ST-elevation, a typical ECG biomarker of OMI. The researchers leveraged a large, multicenter cohort of 7313 patients to test their model externally. The random forest model surpassed clinicians and FDA-approved ECG interpretation systems' AUROC of 0.91, increasing precision and sensitivity. Additionally, the model's derived OMI risk score improved triage and appropriately reclassified one-third of patients and reduced false negatives. Erol et al. [22] looked at the role of single-photon emission computed tomography (SPECT) myocardial perfusion imaging (MPI) in diagnosing myocardial ischemia among patients with long COVID. They compared 105 long COVID patients to 184 controls matched by age. There were increased rates of ischemia and abnormal summed stress score (SSS > 3) in the long COVID cohort ($p < 0.05$). Long COVID was independently associated with ischemia on multivariate analysis ($p < 0.001$) along with high serum C-reactive protein (CRP) levels and SPECT lung-to-heart ratio (LHR). Zhong et al. [23] proposed an SVM-based automatic ECG diagnosis framework for myocardial ischemia using the Long-Term ST Database (LTSTDB). Through advanced filtering, precise localization of relevant waveforms (98.36% R-wave accuracy), and automated feature extraction (e.g., ST-segment shift, slope and wave amplitude), SVM classification yielded 97.98% overall accuracy utilizing 24 features and 95.93% accuracy with 10 features. Zaichenko et al. [24] developed Ultra-High-Resolution Electrocardiography (UHR ECG). This new method allowed for earlier detection of transmural and subendocardial myocardial ischemia when compared with standard ECG (C ECG). Measurements with expanded frequency (up to 2 000 Hz) and sensitivity bandwidths revealed UHR ECG to be more sensitive. Using this method, researchers were able to detect ischemia induced power spectral density decreases in the high-

frequency channel before any ST-segment deviation was detected on C ECG. These findings were shown using rat and porcine models. Proposed by Panchal et al. [25] was DL based, where the researchers used CNN along with the pre-trained models like VGG model (VGG16 and VGG19). They developed an automated system using Deep Learning techniques to find out the cardiac severity from ECG images. In their study, a total of 929 ECG images were taken. Out of 929 images, 743 images were used for training purpose whereas 186 images were used for testing purpose. Different datasets were taken into consideration including datasets of heart attack, abnormal pulse, history of myocardial and normal ECG pattern. Experimental results showed that DL approaches based on the pre-trained model VGG-16 gives better results (accuracy of 70.71%) than CNN (50.69%) and VGG-19 (61.75%). Zhang et al. [9] assessed the predictive performance of adding high-frequency QRS complex analysis (HFQRS) to exercise ECG and ML models for predicting angiography validated CAD. Six ML models using clinical variables including age, sex, hypertension, diabetes, and HFQRS results were compared, and Logistic Regression demonstrated the strongest predictive ability (AUC 0.88).

In another novel study, Alahdab et al. [26] trained an ML model to predict myocardial flow reserve (MFR) from standard ECGs that was externally validated on ECG- PET and ECG- SPECT datasets. They developed a swarm intelligence-based support vector regression model that had AUC of 0.83 and showed strong prognostic performance for prediction of major adverse cardiovascular events (MACE) with a C-statistic of 0.76 upon external validation. By using such an inexpensive and widely available diagnostic tool, namely ECGs, cardiac patients could be risk stratified and benefit from improved access to cardiac care. Chen et al. [27] conducted a novel study which created an ML-based model that assessed the risk for myocardial ischemia (MI) in patients exposed to high-altitude regions. Based on their study population of 2855 soldiers who were participating in high-altitude training, they discovered hematologic markers including eosinophil percentage, globulin, calcium, glucose and aspartate aminotransferase (AST) levels to be significant predictors of risk. Their random forest model was able to classify with high accuracy and demonstrated an AUC of 0.86, indicating the potential for early identification of MI risk. Jorda et al. [28] conducted a secondary analysis of the ISCHEMIA and ISCHEMIA CKD trials, including 5876 patients with chronic coronary syndrome, to evaluate the prognostic value of the baseline ECG findings. They reported that patients with abnormal ECG tracings had a higher incidence of cardiovascular death (8.6% vs. 3.3%; adjusted HR 2.01, 95% CI 1.58–2.55) over a median follow-up of 3.1 years. Specific abnormalities such as ST-segment depression (0.5mm; adjusted HR 2.0, 95% CI 1.52–2.63), T-wave inversion (1mm; adjusted HR 1.89, 95% CI1.40–2.54), LBBB (adjusted HR 1.74, 95% CI1.05–2.90), and RBBB (adjusted HR 1.52, 95% CI 1.04–2.22) were significantly associated with increased risk. Al Younis et al. [29] evaluated the use of ML models to classify patients with heart failure (HF) into three categories, HFpEF (preserved), HFmrEF (mid-range) and HFrEF (reduced), based on circadian ECG features. Using a dataset of 24-hour ECG recordings from 303 patients, features such as QTc, QRS, and ST-T intervals were extracted and analyzed. Among the ML models tested, decision trees achieved the highest accuracy (91.2%) and AUROC (0.98), particularly between 10–11 pm, outperforming k-nearest neighbors (KNN, 90.9%) and SVM (87.8%). QTc and QRS were identified as the most common features of the classification. Sun et al. [30] developed a heterogeneous ensemble learning model for detecting myocardial ischemia in patients with non-diagnostic 12-lead ECGs by integrating deterministic learning-derived dynamic features, such as the Lempel-Ziv complexity and Lyapunov exponent. Using a dataset of 499 patients, the model achieved an average accuracy of 89.10%, sensitivity of 91.72%, and specificity of 82.69%. The model demonstrated 91.11% accuracy, on the independent public PTB dataset outperforming the traditional ML classifiers. Zhao et al. [31] developed an ML-based model that utilized ECG and vector cardiogram (VCG) features to accurately identify myocardial ischemia. Sample entropy as well as spatial heterogeneity index and temporal heterogeneity index from ECG and VCG measurements were used in their model. Performance on the training cohort was 90.3% accuracy, 90.3% sensitivity, and 90.5% specificity. When applied to an independent testing set, the model exhibited an AUC of 0.814, indicating good generalizability. Wang et al. [32] found ML models, particularly artificial neural network (ANN), to have the best predictive ability for tachyarrhythmia after acute myocardial infarction (AMI). Their ANN model utilized lesion location, abnormal wall motion, bundle branch block, age and heart rate as features. The ANN had 66.8% accuracy and an AUC of .654. This suggests including myocardial ischemia-related variables such as lesion location and wall

motion abnormalities lead to better arrhythmia prediction than conventional models like GRACE score. Sun et al. [33] proposed a novel approach that employs self-adaptive variational mode decomposition (VMD) deterministic learning approach was employed to detect myocardial ischemia and identify the cause of myocardial ischemia from ECG signals. ECG signals were decomposed into depolarization and repolarization components before dynamic modeling was applied to perform feature extraction. Hybrid feature reduction scheme was used to identify informative features and select them. The performance of the method was evaluated using 297 ECG records and the method achieved an AUROC of 0.9 126 in detecting myocardial ischemia and an AUROC of 0.9 238 in identifying the cause of myocardial ischemia.

L. Wang et al. [34] studied how emotional fluctuation features from heart rate variability (HRV) can be combined with ECG analysis to predict heart diseases including myocardial ischemia. This paper presents a Heart Disease Prediction (HDP) model that uses machine learning, specifically an SVM algorithm for accurate heart disease classification. The model used feature fusion to combine emotional fluctuation patterns and stress detection. It achieved 91.2% accuracy for arrhythmia, 90.9% for atrial fibrillation and 95.7% for myocardial ischemia. Seoni et al. [35] developed a Spatial Uncertainty Estimator (SUE) to evaluate the reliability of AI predictions for detecting myocardial ischemia using ECG signals. Their proposed CNN-BiLSTM model showed outstanding performance. It achieved 99.6% accuracy, 99.8% sensitivity and 98.2% specificity. Gómez et al. [36] introduced a time-warping-based morphology index called dPCAw Tpe to evaluate ischemia-induced ventricular repolarization dispersion. This index was used to predict ventricular fibrillation (VF) in a porcine myocardial infarction model. In this study the dPCAw Tpe index performed better than traditional Tpe- based measures (TPCApe). It showed higher predictive accuracy with an AUC of 0.85, sensitivity of 90% and specificity of 75% for thresholds of 20 ms in detecting delayed VF episodes during left anterior descending artery occlusion. Strodthoff and Strodthoff [11] trained an ensemble of FCNs directly on raw ECGs from the PTB database for myocardial infarction detection and applied attribution methods to inspect its decision criteria. The classifier reached 93.3% sensitivity and 89.7% specificity, and the attribution maps highlighted physiologically plausible ECG segments, although the explanations were assessed qualitatively. Jahmunah et al. [12] developed DenseNet and CNN models for the classification of healthy subjects and ten myocardial infarction classes defined by infarct location using the PTB database. Both models exceeded 95% accuracy, and Grad-CAM showed that lead V4 was the most activated lead and that different leads and wave segments were activated for each class. Anand et al. [12] proposed a lead-wise CNN with global average pooling (ST-CNN-GAP-5) for the diagnosis of cardiac disorders on PTB-XL and used SHAP to identify the leads and time points that contributed to each diagnostic class. Taleban et al. [10] systematically reviewed 45 studies on XAI in electrocardiography and grouped the methods into perturbation-based, gradient-based, intrinsically interpretable, sequence-aware and counterfactual approaches. They reported that gradient and perturbation methods dominate because they are simple to implement, but that the evaluation of explanations is fragmented, with inconsistent metrics and no standard benchmarks. Moody et al. [7] pretrained a vision Transformer with self-supervised learning on 800,035 unlabeled ECGs from MIMIC-IV-ECG and fine-tuned it for 12 tasks labeled by positron emission tomography (PET), SPECT and magnetic resonance imaging (MRI). The model achieved AUROC values from 0.763 for impaired myocardial flow reserve to 0.955 for reduced left ventricular ejection fraction, and self-supervised pretraining improved performance in 11 of the 12 tasks. Apostol and Nutu [8] compared one- and two-dimensional CNNs, ResNet and vision Transformers for arrhythmia classification on five public 12-lead ECG databases, including PTB-XL and the Georgia 12-lead ECG Challenge database. A one-dimensional CNN with gated recurrent units (GRU) trained on raw signals achieved the highest accuracy (93.4%). Abdullah et al. [37] evaluated a fine-tuned CNN and 17 other models for five-class arrhythmia classification on the MIT-BIH database (98.51% accuracy) and quantitatively assessed Grad-CAM and Integrated Gradients using QRS attribution (61.4%), deletion/insertion scores (0.89) and parameter-randomization sanity checks. Overall, the literature shows that deep learning achieves high performance for ECG-based detection of ischemic heart disease, but few ischemia-specific studies report patient-level splits, external validation and a quantitative assessment of whether the explanations agree with expert annotations. The present study targets these gaps.

Table 1. Comparison of explainable AI approaches for ischemic heart disease and related ECG classification tasks

Study	Task (dataset)	Model	XAI method	Quantitative XAI evaluation	Patient-level or external validation
Strodthoff and Strodthoff [11]	MI vs. healthy (PTB)	FCN ensemble	Attribution maps	No	No
Jahmunah et al. [12]	11-class MI location (PTB)	DenseNet, CNN	Grad-CAM	No	No
Anand et al. [13]	Cardiac disorders (PTB-XL)	Lead-wise CNN-GAP	SHAP	No	Not reported
Al-Zaiti et al. [21]	Occlusion MI (multicenter)	Random forest	Feature importance	No	Yes
Seoni et al. [35]	CAD vs. normal	CNN-BiLSTM	Spatial uncertainty estimator	Uncertainty only	No
Abdullah et al. [37]	Arrhythmia (MIT-BIH)	Fine-tuned CNN	Grad-CAM, Integrated Gradients	Yes (deletion/insertion, sanity checks)	No
Present study	Ischemia vs. normal (PTB-XL, G12EC)	FCN	Segment-wise LIME	Yes (expert territories, ST-T mass, reproducibility, deletion)	Yes (patient-level test and external set)

3. Methodology

The methodology for predicting myocardial ischemia from ECG data comprises signal pre-processing, median-beat extraction, model design, training, evaluation and explainability, as illustrated in Figure 1.

3.1 ECG Signal Representation and Pre-processing

Each ECG recording was represented as a multivariate time series. The dataset is denoted $X \in \mathbb{R}^{N\times T\times L}$, where N is the number of ECG recordings, T is the number of time samples (1,000 samples, 10 s at 100 Hz) and $L = 12$ is the number of leads. Each instance $x^{(i)} \in \mathbb{R}^{T\times L}$ is the raw ECG of recording i.

To mitigate baseline wander, muscle noise and power-line interference, a cleaning function was applied to every lead:

$$x_{clean}^{(i)} = f_{clean}\left(x^{(i)}\right) \tag{1}$$

where $f_{\text{clean}}(\cdot)$ is the NeuroKit2 ECG cleaning function [38] (method "neurokit"), which applies a fifth-order Butterworth high-pass filter with a 0.5 Hz cut-off followed by 50 Hz power-line filtering independently to each lead, and $x_{\text{clean}}^{(i)}$ is the cleaned recording.

R-peaks were detected on the cleaned lead II, which usually shows the most prominent QRS complexes:

$$R = \{t_1, t_2, \ldots, t_K\} = f_{peak}\left(x_{clean,II}^{(i)}\right) \tag{2}$$

where $f_{\text{peak}}(\cdot)$ is the NeuroKit2 QRS detector, t_k is the sample index of the k-th R-peak and K is the number of detected beats in the recording. Recordings with fewer than three valid beats were excluded.

Table 2. Summary of the literature review: objectives, results and limitations

Study	Objective	Result	Limitation
Singh et al. [20]	Review of ECG parameters for predicting cardiac mortality	QT interval, QRS duration, T-wave inversion and bundle branch block associated with higher mortality	No single ECG parameter reliably stratifies sudden cardiac death risk
Sarıçam et al. [14]	Algorithm for predicting the culprit vessel in lateral MI	100% PPV for obtuse marginal and 98% PPV for large diagonal branch occlusion	Small sample (74 STEMI patients)
Al-Zaiti et al. [21]	ML model for diagnosing occlusion MI	Random forest AUROC 0.91, outperforming clinicians and FDA-approved ECG systems	Limited focus on interpretability
Erol et al. [22]	Ischemia in long COVID patients using SPECT	Long COVID was an independent predictor of ischemia	No prospective validation
Zhong et al. [23]	SVM-based ischemia diagnosis on LTSTDB	97.98% accuracy (24 features); 95.93% (10 features)	Manual feature engineering; ECG only
Zaichenko et al. [24]	Ultra-high-resolution ECG for early ischemia detection	Detected ischemia-induced changes earlier than conventional ECG	Validated only in animal models
Panchal et al. [25]	DL on ECG images for cardiac severity	VGG-16 accuracy 70.71%	Small dataset (929 images)
Zhang et al. [9]	HFQRS and ML for CAD prediction	Logistic regression AUC 0.88	Requires high-quality HFQRS data
Alahdab et al. [26]	Predicting myocardial flow reserve from ECG	AUC 0.83; prognostic for MACE	Needs validation in diverse populations
Chen et al. [27]	Ischemia risk at high altitude using ML	Random forest AUC 0.86	Specific cohort (soldiers)
Jorda et al. [28]	Prognostic value of baseline ECG in chronic coronary syndrome	Abnormal ECG associated with cardiovascular death (adjusted HR 2.01)	Secondary analysis; causality not inferred
Al Younis et al. [29]	Heart failure type classification using circadian ECG features	Decision tree accuracy 91.2%, AUROC 0.98	303 patients; no external validation
Sun et al. [30]	Ensemble learning for ischemia with non-diagnostic ECGs	Accuracy 89.10%; 91.11% on PTB	Limited interpretability
Zhao et al. [31]	ECG and VCG features for ischemia detection	Accuracy 90.3%; external AUC 0.814	Needs larger, diverse populations
Wang et al. [32]	Tachyarrhythmia after AMI using ANN	Accuracy 66.8%, AUC 0.654	Low accuracy
Sun et al. [33]	Ischemia and etiology detection using VMD and deterministic learning	AUROC 0.9126 (ischemia), 0.9238 (etiology)	297 records
L. Wang et al. [34]	HRV emotional features and ECG for heart disease prediction	SVM accuracy 95.7% for ischemia	Limited long-term analysis
Seoni et al. [35]	Spatial uncertainty estimation for CNN-BiLSTM on CAD ECGs	Accuracy 99.6%, sensitivity 99.8%, specificity 98.2%	Uncertainty not explored in diverse conditions
Gómez et al. [36]	VF prediction using a time-warping morphology index	AUC 0.85, sensitivity 90%, specificity 75%	Animal model only
Strodthoff and Strodthoff [11]	FCN ensemble with attribution maps for MI (PTB)	Sensitivity 93.3%, specificity 89.7%	Qualitative evaluation of explanations
Jahmunah et al. [12]	DenseNet/CNN with Grad-CAM for 10 MI locations (PTB)	Accuracy above 95%; V4 most activated lead	Qualitative evaluation of explanations
Anand et al. [13]	Lead-wise CNN with SHAP (PTB-XL)	SHAP identified contributing leads and time points	Explanations not validated against expert labels
Taleban et al. [10]	Systematic review of XAI in ECG (45 studies)	Gradient and perturbation methods dominate	Evaluation of XAI fragmented, no benchmarks
Moody et al. [7]	Self-supervised vision Transformer (MIMIC-IV-ECG)	AUROC 0.763–0.955 across 12 tasks	Requires very large pretraining data
Apostol and Nutu [8]	CNNs and vision Transformers for arrhythmia (5 databases)	Best accuracy 93.4% (GRU-based 1D CNN)	Arrhythmia, not ischemia
Abdullah et al. [37]	CNN with Grad-CAM and Integrated Gradients (MIT-BIH)	Accuracy 98.51%; deletion/insertion score 0.89	Arrhythmia, not ischemia

Heartbeats were segmented with fixed pre- and post-R-peak windows applied synchronously to all 12 leads:

$$x_{beat}^{(k)} = x_{clean}^{(i)}[t_k - w_1 : t_k + w_2] \quad (3)$$

where w_1 = 0.25 s and w_2 = 0.45 s are the pre- and post-R-peak window durations (25 and 45 samples at 100 Hz), so that each beat contains the P-wave, QRS complex, ST-segment and T-wave, and $x_{\text{beat}}^{(k)} \in \mathbb{R}^{70\times12}$ is the k-th beat. Beats whose window exceeded the recording boundaries were discarded.

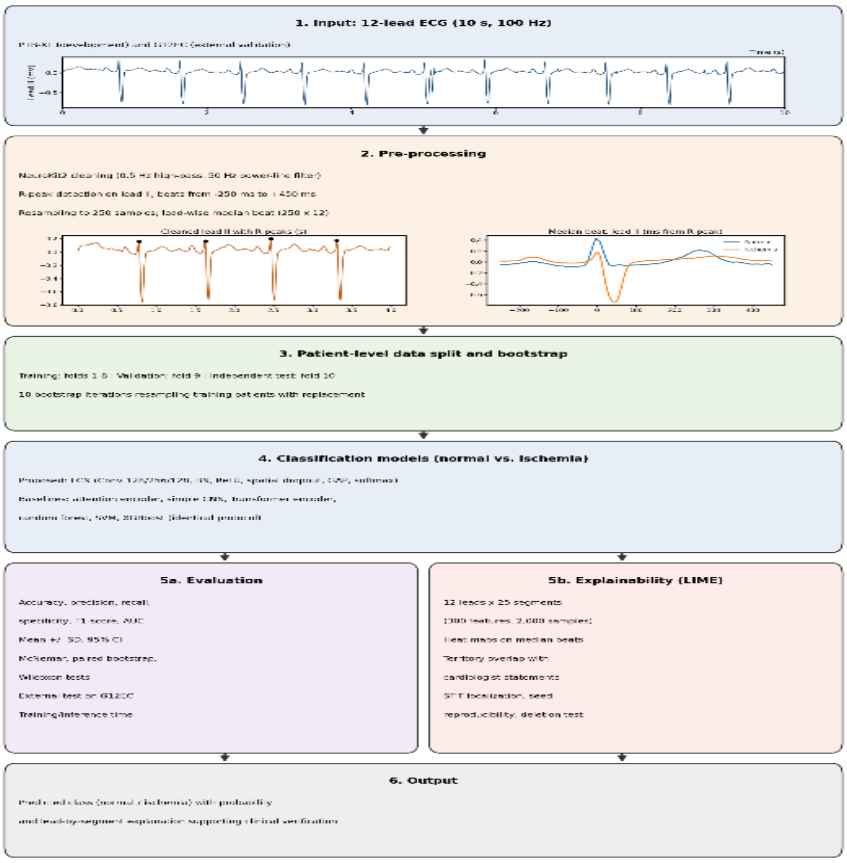


Fig. 1. Overview of the proposed methodology

Each heartbeat was resampled to a fixed length T_b = 250 samples:

$$x_{res}^{(k)} = f_{resample}\left(x_{beat}^{(k)}, T_b\right) \tag{4}$$

where $f_{\text{resample}}(\cdot)$ is Fourier-domain resampling and $x_{\text{res}}^{(k)} \in \mathbb{R}^{250\times12}$ is the resampled beat. After resampling, one sample corresponds to 2.8 ms and the R-peak is located at sample 89.

The median beat was computed sample-wise across all beats of a recording to obtain a noise-robust representative morphology:

$$x_{median}^{(i)} = \underset{k=1,\ldots,K}{median}\left(x_{res}^{(k)}\right) \tag{5}$$

where the median is taken independently at every time sample and lead, K denotes the number of beats in the recording, and the processed ECG of recording i is $x_{\text{median}}^{(i)} \in \mathbb{R}^{T_b\times L}$ with T_b = 250 and L = 12.

3.2 Deep Learning Models

3.2.1 Fully Convolutional Network (FCN)

The FCN [15] extracts temporal features within each lead and spatial features across leads. The input $Z^{(0)} = X \in \mathbb{R}^{T_b\times L}$ is processed by three convolutional layers:

$$Z^{(l)} = \sigma\left(BN\left(W^{(l)} * Z^{(l-1)} + b^{(l)}\right)\right) \tag{6}$$

where l = 1, 2, 3 is the layer index, $*$ denotes one-dimensional convolution, $W^{(l)}$ and $b^{(l)}$ are the kernel weights and biases of the l-th layer, BN(·) is batch normalization, σ(·) is the rectified linear unit (ReLU) and $Z^{(l)} \in \mathbb{R}^{T'\times C_l}$ is the resulting feature map with C_l channels.

Global average pooling (GAP) summarizes each channel of the last feature map over time:

$$z_{gap} = \frac{1}{T'}\sum_{t=1}^{T'} Z_t^{(3)} \tag{7}$$

where $T' = 250$ is the temporal length of the final feature map ("same" padding), $Z_t^{(3)} \in \mathbb{R}^{128}$ is the feature vector of the third layer at time t and $z_{gap} \in \mathbb{R}^{128}$ is the pooled feature vector.

A fully connected layer with softmax activation produces the class probabilities:

$$\hat{y}_k = \frac{exp\left(w_k^T z_{gap} + c_k\right)}{\sum_{j=1}^{2} exp\left(w_j^T z_{gap} + c_j\right)} \tag{8}$$

where $k \in \{1, 2\}$ indexes the classes (normal and ischemia), w_k and c_k are the weights and bias of the output unit for class k, and $\hat{y}_k$ is the predicted probability of class k.

3.2.2 Encoder Model

The encoder [39] transforms the ECG into a compressed latent representation using three convolutional blocks:

$$Z^{(l)} = PReLU\left(IN\left(W^{(l)} * Z^{(l-1)} + b^{(l)}\right)\right) \tag{9}$$

where IN(·) denotes instance normalization, PReLU is the parametric rectified linear unit, and $W^{(l)}$, $b^{(l)}$ and $Z^{(l)}$ are defined as in Eq. (6). Max-pooling with a factor of 2 follows the first two blocks.

The attention mechanism splits the output of the third block channel-wise into a data part Z_d and an attention part Z_a of equal size:

$$Z_{att} = Softmax_t(Z_a),\ Z_w = Z_{att} \odot Z_d \tag{10}$$

where $Softmax_t$ normalizes the attention scores along the time axis, $\odot$ denotes element-wise multiplication and Z_w is the attention-weighted feature map.

The weighted features are passed through a dense layer with sigmoid activation, flattened and mapped to class probabilities:

$$\hat{y}_k = Softmax_k\left(W_o\ Flatten\left(\sigma(W_s Z_w)\right)\right) \tag{11}$$

where W_s is the weight matrix of the sigmoid dense layer applied at each time step, σ(·) is the sigmoid function and W_o is the weight matrix of the output layer.

3.2.3 Simple CNN

A lightweight CNN extracts features through two convolution and max-pooling stages:

$$z_{flat} = Flatten\left(MaxPool\left(\sigma\left(W_2 * MaxPool\left(\sigma(W_1 * X + b_1)\right) + b_2\right)\right)\right) \tag{12}$$

where W_1, W_2 and b_1, b_2 are the kernels and biases of the two convolutional layers, σ(·) is ReLU and z_{flat} is the flattened feature vector that is passed to two fully connected layers for classification.

3.2.4 Transformer Encoder

Because Transformer models are increasingly used for ECG classification [7, 8], a Transformer encoder [40] was included as a baseline. The median beat is divided into 25 non-overlapping patches of 10 samples, which are linearly embedded and combined with a learnable classification token and positional embeddings:

$$e_j = W_e x_{[j]} + p_j,\ j = 0, 1, \ldots, 25 \tag{13}$$

where $x_{[j]} \in \mathbb{R}^{120}$ is the j-th flattened patch (10 samples × 12 leads), W_e is the embedding matrix, p_j is the positional embedding, and e_0 is the classification token. Each of the four encoder layers applies multi-head self-attention:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_h}}\right)V \quad (14)$$

where Q, K and V are the query, key and value projections of the token sequence and $d_h = 16$ is the dimension of each of the four attention heads (model dimension 64, feed-forward dimension 128). The output of the classification token is mapped to the class probabilities by a linear layer with softmax.

3.2.5 Classical Machine Learning Baselines

Random forest [41], an SVM with a radial basis function kernel [42] and XGBoost [43] were trained on the median beats downsampled by a factor of 5 (50 samples × 12 leads = 600 features) and standardized to zero mean and unit variance using statistics of the training data only.

3.3 Training and Optimization

All deep models minimize the class-weighted categorical cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{2}\alpha_k y_k^{(i)} \log \hat{y}_k^{(i)} \quad (15)$$

where N is the number of training ECGs, $y_k^{(i)} \in \{0, 1\}$ is the one-hot true label, $\hat{y}_k^{(i)}$ is the predicted probability for class k and $\alpha_k = N/(2N_k)$ is the weight of class k, with N_k training ECGs, which compensates for class imbalance.

Optimization is performed with the Adam optimizer:

$$\theta_{t+1} = \theta_t - \eta\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \quad (16)$$

where θ_t are the model parameters at step t, $\eta = 10^{-4}$ is the learning rate, $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates of the gradient of $\mathcal{L}$ ($\beta_1 = 0.9$, $\beta_2 = 0.999$) and $\varepsilon = 10^{-8}$.

3.4 Validation and Evaluation

Patient-level separation was enforced throughout. PTB-XL provides ten stratified folds in which all recordings of a patient belong to the same fold [17]; folds 1–8 were used for training, fold 9 for validation (early stopping and learning-rate scheduling) and fold 10 as the independent test set, which was never used for model selection. No patient appears in more than one of these subsets, which was verified programmatically, so data leakage between training and testing was avoided. Robustness was assessed with ten bootstrap iterations: in each iteration, patients (not recordings) of folds 1–8 were sampled with replacement, all recordings of the sampled patients formed the training set, and the trained model was evaluated on the unchanged independent test fold and on the external dataset. For a test set with TP true positives, TN true negatives, FP false positives and FN false negatives (ischemia as the positive class, threshold 0.5), the metrics were computed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (17)$$

recall (sensitivity):

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN} \quad (18)$$

specificity:

$$Specificity = \frac{TN}{TN + FP} \quad (19)$$

precision:

$$Precision = \frac{TP}{TP + FP} \quad (20)$$

and the F1-score:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (21)$$

In addition, the area under the receiver operating characteristic (ROC) curve (AUC) was computed as a threshold-independent measure of discrimination:

$$AUC = \int_0^1 TPR(FPR)\, d(FPR) \quad (22)$$

where the true positive rate (TPR) and false positive rate (FPR) are obtained by varying the decision threshold.

For each metric, the mean and standard deviation (SD) across the B = 10 bootstrap iterations were reported with a 95% confidence interval (CI):

$$\mu_m = \frac{1}{B}\sum_{b=1}^{B} m_b \ , \ 95\%\ CI = \mu_m \pm t_{0.975,B-1}\frac{s_m}{\sqrt{B}} \quad (23)$$

where m_b is the value of metric m in iteration b, μ_m and s_m are its mean and SD, and $t_{0.975,\,B-1}$ is the 97.5th percentile of Student's t distribution with B − 1 degrees of freedom. The final model of each architecture was the bootstrap-aggregated (bagged) ensemble that averages the predicted probabilities of the ten models. For the final models, 95% CIs were also estimated by 1,000 bootstrap resamples of the test ECGs [44]. Paired comparisons between the FCN and every other model on the same test ECGs used McNemar's test with continuity correction for the accuracy [45], a paired bootstrap test (2,000 resamples) for the difference in AUC, and the Wilcoxon signed-rank test for the per-iteration AUC values. A p-value below 0.05 was considered statistically significant. Training time, inference time per ECG, number of parameters and peak GPU memory were recorded for every model.

3.5 Model Explainability Using LIME

LIME [16] was used to explain the predictions of the final FCN. LIME approximates the model locally around one ECG with a simple, interpretable linear model and thereby identifies the ECG regions that contribute most to the prediction. This supports the interpretation and clinical verification of the model's decisions.

3.5.1 Setup of LIME

Explanations were computed for ECGs of the independent test fold. Each median beat $x^{(i)} \in \mathbb{R}^{T_b \times L}$ (T_b = 250, L = 12) was divided into interpretable features, and the model prediction $\hat{y}^{(i)}$ and the true label $y^{(i)}$ were recorded.

The LIME explainer was configured as follows:

- Features: 300 interpretable features formed by 25 non-overlapping temporal segments of 10 samples (28 ms) in each of the 12 leads (Lead I, II, III, aVR, aVL, aVF, V1–V6).
- Perturbation: 2,000 random binary masks per ECG; a masked segment is replaced by the corresponding segment of the mean training beat, so that perturbed signals remain physiologically plausible. The class labels are normal and ischemia.
- Surrogate: weighted ridge regression trained on the predicted ischemia probability of the perturbed samples, using an exponential kernel of width 0.25 on the cosine distance. Features were not pruned.

For each ECG, the explainer provides an importance matrix $E^{(i)} \in \mathbb{R}^{25\times12}$, where $E_{s,l}^{(i)}$ quantifies the contribution of segment s of lead l to the predicted probability of ischemia.

3.5.2 LIME Explanation Workflow

For a given ECG, the predicted class probabilities of the trained model are:

$$\hat{y}^{(i)} = f\left(x^{(i)};\, \Theta\right) \quad (24)$$

where f is the trained model with parameters Θ, and the prediction is compared with the true label $y^{(i)}$.

The explanation is the interpretable model that best approximates the classifier in the neighborhood of the ECG:

$$g^{(i)*} = \underset{g \in G}{arg\,min} \sum_{z \in Z_i} \pi_i(z) \left(f\left(h(z)\right) - g(z) \right)^2 + \Omega(g), \quad \pi_i(z) = exp\left(-\frac{D(1,z)^2}{\omega^2} \right) \tag{25}$$

where $z \in \{0, 1\}^{300}$ is a binary mask indicating which segments are kept, Z_i is the set of 2,000 sampled masks, $h(z)$ maps a mask to the perturbed ECG, G is the class of linear models, $\Omega(g)$ is the ridge penalty, $\pi_i(z)$ is the proximity kernel with cosine distance D and width $\omega = 0.25$, and 1 is the unmasked vector. The coefficients of $g^{(i)*}$ form the importance matrix $E^{(i)}$.

3.5.3 Construction of Heat Maps

The positive importance values, which support the ischemia class, were normalized to the interval [0, 1] for visualization:

$$H_{s,l} = \frac{max\left(0, E_{s,l}^{(i)}\right)}{\underset{s',l'}{max}\, max\left(0, E_{s',l'}^{(i)}\right)} \tag{26}$$

where $H_{s,l}$ is the heat map intensity of segment s in lead l. Each segment value is assigned to the ECG samples it covers:

$$H_l(t) = H_{[t/k],l}, \quad t = 1, \ldots, T_b, \quad k = 10 \tag{27}$$

where $H_l(t)$ is the intensity at sample t of lead l and k is the segment length. The heat maps were overlaid on the median beats, providing a direct comparison between the ECG waveform and its contribution to the classification.

3.5.4 Quantitative Evaluation of the Explanations

Because visual inspection alone cannot demonstrate that the explanations are clinically meaningful, four quantitative analyses were performed. Expert reference: PTB-XL provides cardiologist statements that localize ischemia to anterior (V3–V4), anteroseptal (V1–V4), lateral (I, aVL, V5–V6), anterolateral (I, aVL, V3–V6), inferior (II, III, aVF) or inferolateral (II, III, aVF, I, aVL, V5–V6) leads [17]. For ischemic ECGs with such a statement, the territory mass is the share of positive importance located in the annotated leads:

$$TM^{(i)} = \frac{\sum_{l \in \mathcal{T}_i} \sum_{s=1}^{25} E^+_{s,l}}{\sum_{l=1}^{12} \sum_{s=1}^{25} E^+_{s,l}} \tag{28}$$

where $E^+_{s,l} = \max(0, E_{s,l}^{(i)})$ and $\mathcal{T}_i$ is the set of leads of the annotated territory. TM was compared with the share expected by chance (number of territory leads divided by 12), and top-1 and top-3 lead localization accuracy were defined as the proportion of ECGs in which the most important lead, or at least one of the three most important leads, belonged to the annotated territory. Temporal localization was measured as the share of positive importance within the ST–T window (60–400 ms after the R-peak, segments 12–23), in the QRS complex and in the isoelectric baseline before the P-wave and after the T-wave; attribution to baseline noise would place importance in the isoelectric segments. Reproducibility was measured by repeating LIME with three random seeds and computing the mean Spearman rank correlation of the importance matrices and the Jaccard overlap of the 30 most important features. Faithfulness was assessed with a deletion test:

$$\Delta_k = f\left(x^{(i)}\right) - f\left(h\left(x^{(i)}, S_k\right)\right) \tag{29}$$

where S_k is the set of the k most important features (k = 15 or 30) and $h(x^{(i)}, S_k)$ replaces them by the reference beat. Δ_k was compared with the drop obtained by removing k randomly chosen features using the one-sided Wilcoxon signed-rank test; faithful explanations produce a significantly larger drop. Results of these analyses are reported in Section 5.4.

4. Experiments

This section describes the datasets, the experimental setup and the implementation of the models.

4.1 Dataset Description

The development dataset was PTB-XL (version 1.0.3), a public database of 21,799 clinical 10-second 12-lead ECGs from 18,869 patients recorded with Schiller AG devices between October 1989 and June 1996 [17, 18, 46]. The signals are available at 500 Hz and as a 100 Hz version; the 100 Hz signals (1,000 samples per recording, 1 μV resolution) were used, as in the PTB-XL benchmark study [47]. Each ECG was annotated by up to two cardiologists with SCP-ECG statements and likelihoods, and 78.7% of the included ECGs were validated by a second cardiologist. The data are fully de-identified and released under the Creative Commons Attribution 4.0 license.

The inclusion and exclusion criteria were defined from the diagnostic statements. The ischemia class included all ECGs with at least one ischemic statement (non-specific ischemic, or ischemic in anterior, anteroseptal, lateral, anterolateral, inferior or inferolateral leads) with a likelihood of at least 50%. The normal class included ECGs whose only diagnostic statement was “normal ECG” with a likelihood of at least 80%. ECGs with any other diagnostic pattern without an ischemic statement (for example, isolated myocardial infarction, conduction disturbances or hypertrophy) were excluded, as were recordings in which fewer than three valid beats were detected. The final cohort comprised 11,018 ECGs from 10,135 patients: 2,441 ischemia ECGs from 2,109 patients and 8,577 normal ECGs from 8,058 patients, corresponding to a class ratio of approximately 1:3.5. The mean age was 68.1 ± 12.7 years in the ischemia class and 51.7 ± 17.4 years in the normal class, and 54.1% and 45.2% of the patients were male, respectively. A median of 11 beats per ECG contributed to each median beat. Table 3 summarizes the characteristics of the training, validation and test sets.

External validation used the Georgia 12-Lead ECG Challenge (G12EC) database from Emory University, Atlanta, USA, which is part of the PhysioNet/Computing in Cardiology Challenge 2020 [19, 46]. It contains 10,344 10-second 12-lead ECGs sampled at 500 Hz and labeled with SNOMED CT codes. G12EC does not use the general myocardial ischemia code; its ischemia labels are anterior ischemia (code 426434006) and acute myocardial ischemia (code 413444003). ECGs with either label formed the ischemia class and ECGs whose only label was sinus rhythm (code 426783006) formed the normal class, giving 2,033 ECGs (281 ischemia and 1,752 normal; mean age 56.1 ± 15.7 years; 45.6% male). The signals were resampled to 100 Hz and processed with the same pipeline. This dataset was never used for training or model selection and differs from PTB-XL in country, hospital system, acquisition period, devices and labeling procedure.

Table 3. Characteristics of the development (PTB-XL) and external (G12EC) datasets

Subset	ECGs	Patients	Ischemia / normal	Age (years), mean ± SD	Male
PTB-XL training (folds 1–8)	8,819	8,103	1,947 / 6,872	55.4 ± 17.8	46.1%
PTB-XL validation (fold 9)	1,093	1,024	246 / 847	55.0 ± 17.9	50.9%
PTB-XL test (fold 10)	1,106	1,008	248 / 858	55.8 ± 17.8	51.8%
G12EC external test	2,033	Not available	281 / 1,752	56.1 ± 15.7	45.6%

Patients are disjoint between the PTB-XL training, validation and test sets. G12EC does not provide patient identifiers. Ages above 89 years are coded as 90 in PTB-XL. The data were split at the patient level using the predefined PTB-XL folds (Section 3.4): folds 1–8 for training, fold 9 for validation and fold 10 as the independent test set. Ten bootstrap iterations resampled the training patients with replacement, and every model was evaluated on the same independent test set and on the external dataset. Compared with the initial submission, which used 400 ECG signals, the revised experiments use a substantially larger cohort, patient-level separation and external validation, which reduces the risk of overfitting and of optimistic performance estimates.

4.2 Model Architectures

Seven models were implemented and evaluated under the same protocol: an FCN, an encoder with an attention mechanism, a simple CNN, a Transformer encoder, random forest, SVM and XGBoost. All deep models receive the 250 × 12 median beat as input. The FCN consists of three convolutional layers with 128, 256 and 128 filters and kernel sizes of 15, 10 and 1. Each layer is followed by batch normalization, ReLU activation and spatial dropout (rates of 10%, 20% and 20%) to stabilize training and reduce overfitting.

Global average pooling reduces the final feature map to a vector, and a dense layer with softmax activation produces the class probabilities. The FCN has 385,282 trainable parameters (Figure 5).

The encoder model consists of three convolutional blocks with kernel sizes of 5, 11 and 21 and 128, 256 and 512 filters, each with instance normalization, parametric rectified linear unit (PReLU) activation and dropout, with max-pooling after the first two blocks. An attention mechanism splits the feature map into data and attention components, which are multiplied to emphasize informative time steps. The output passes through a dense layer with sigmoid activation, is flattened and is mapped to class probabilities (3,221,762 parameters).

The simple CNN is a lightweight baseline with two convolutional layers of 64 filters and a kernel size of 3, each followed by spatial dropout and max-pooling, and two dense layers with 100 and 2 neurons (405,422 parameters). The Transformer encoder has four layers with a model dimension of 64 and four attention heads (143,618 parameters). Random forest used 500 trees, the SVM used a radial basis function kernel with C = 1, and XGBoost used 400 trees with a maximum depth of 6 and a learning rate of 0.05; balanced class weights were used for random forest and SVM.

4.3 Experimental Setup

All deep models were trained with the Adam optimizer (learning rate 10^{-4}) and a batch size of 30 for a maximum of 50 epochs. Early stopping with a patience of 15 epochs monitored the validation accuracy, the weights of the best validation epoch were restored, and the learning rate was halved when the validation accuracy did not improve for one epoch.

The class-weighted loss of Eq. (15) compensated for class imbalance. All random seeds were fixed for reproducibility.

The experiments were implemented in Python 3.12.0 with PyTorch 2.11.0 [48] (CUDA 12.6, cuDNN 9.10.2), scikit-learn, XGBoost, NeuroKit2 [38], WFDB and LIME [16]. Training and inference were performed on a laptop with an NVIDIA Quadro M620 GPU (2 GB memory), a quad-core Intel Core i5-7440HQ CPU and 16 GB RAM running Windows 10.

This modest hardware was chosen deliberately to show that the framework does not require high-end computing resources.

5. Results and Analysis

The models were evaluated on the patient-level independent test set of PTB-XL and on the external G12EC dataset.

Section 5.1 reports the performance of the FCN, Section 5.2 compares all models, Section 5.3 presents the external validation, Section 5.4 quantifies the explainability results and Section 5.5 reports the computational cost.

5.1 Performance of the FCN

Across the ten bootstrap iterations, the FCN achieved an accuracy of 97.82% ± 0.19% (95% CI 97.68%–97.96%), a precision of 95.98% ± 0.58%, a recall (sensitivity) of 94.23% ± 0.78%, a specificity of 98.86% ± 0.17%, an F1-score of 95.10% ± 0.44% and an AUC of 0.995 ± 0.001 (95% CI 0.995–0.996) on the independent test set (Table 4). The small standard deviations show that the performance was stable across resampled training sets. The bagged FCN ensemble reached an accuracy of 97.92% (95% CI 97.11%–98.64%) and an AUC of 0.996 (95% CI 0.991–0.999).

Table 4. Performance of the FCN on the independent test set and the external dataset

Metric	PTB-XL test, 10 bootstrap iterations: mean ± SD [95% CI]	PTB-XL test, bagged ensemble [95% CI]	External G12EC, bagged ensemble [95% CI]
Accuracy	0.978 ± 0.002 [0.977, 0.980]	0.979 [0.971, 0.986]	0.903 [0.889, 0.916]
Precision	0.960 ± 0.006 [0.956, 0.964]	0.963 [0.938, 0.985]	0.597 [0.547, 0.644]
Recall (sensitivity)	0.942 ± 0.008 [0.937, 0.948]	0.944 [0.912, 0.970]	0.922 [0.887, 0.952]
Specificity	0.989 ± 0.002 [0.987, 0.990]	0.990 [0.982, 0.995]	0.900 [0.886, 0.914]
F1-score	0.951 ± 0.004 [0.948, 0.954]	0.953 [0.934, 0.970]	0.724 [0.686, 0.762]
AUC	0.995 ± 0.001 [0.995, 0.996]	0.996 [0.991, 0.999]	0.959 [0.947, 0.967]

Precision, recall, specificity and F1-score use ischemia as the positive class and a threshold of 0.5.

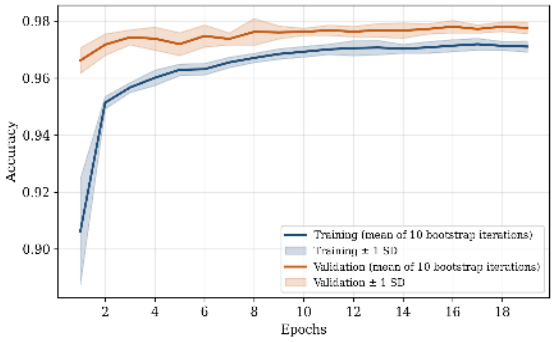


Fig. 2. Training and validation accuracy of the FCN across epochs (mean ± SD over ten patient-level bootstrap iterations)

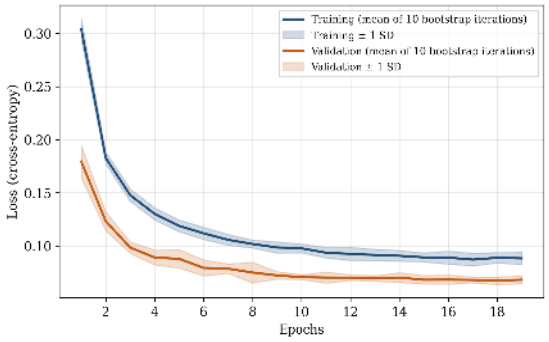


Fig. 3. Training and validation loss of the FCN across epochs (mean ± SD over ten patient-level bootstrap iterations)

Figures 2 and 3 show the training and validation accuracy and loss across epochs (mean ± SD over the ten iterations). Training stopped after 19–35 epochs. The best training accuracy averaged 97.36% and the best validation accuracy 97.99%; the small gap and the parallel loss curves indicate that early stopping, dropout and the larger cohort limited overfitting. The validation accuracy was slightly higher than the training accuracy because dropout is active only during training and the training accuracy is computed on bootstrap-resampled data. All performance values in the abstract, results, discussion and conclusion are taken from this revised experiment and are reported consistently, with the evaluation setting (bootstrap mean on the independent test set, bagged ensemble or external dataset) stated for every value.

Figure 4 shows the ROC curves with 95% confidence bands. The confusion matrices of the bagged FCN are shown in Figure 6: on the PTB-XL test set, 234 of 248 ischemia ECGs and 849 of 858 normal ECGs were classified correctly, with 14 false negatives and 9 false positives. The architecture of the FCN is depicted in Figure 5.

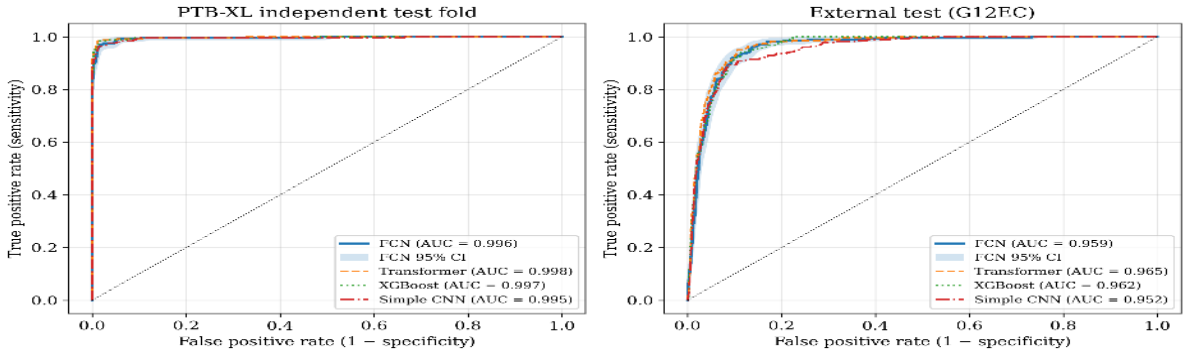


Fig. 4. ROC curves on the PTB-XL independent test set (left) and the external G12EC dataset (right). The shaded band is the bootstrap 95% CI of the FCN

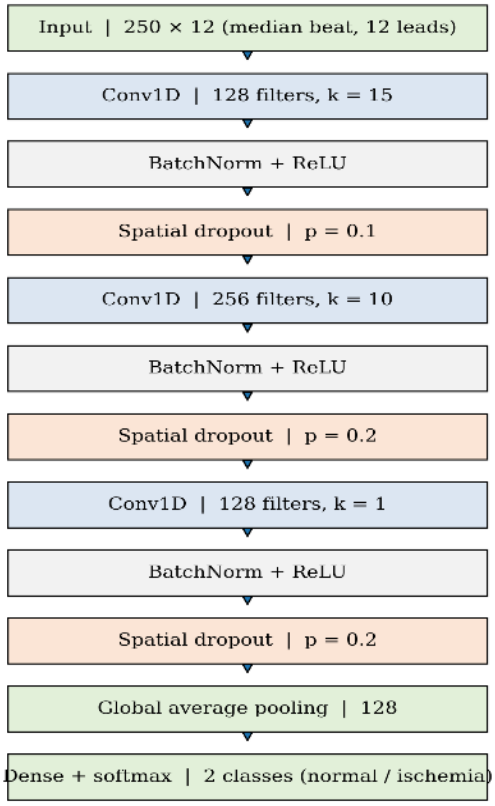


Fig. 5. Architecture of the FCN. Convolutional layers (blue), regularization (orange) and input/pooling/output layers (green)

5.2 Comparison with Deep Learning and Machine Learning Baselines

Table 5 compares all models on the independent test set. The highest mean AUC was obtained by XGBoost (0.9963 ± 0.0004), and the FCN reached 0.9954 ± 0.0006. The AUC differences between the FCN and the encoder, the simple CNN, random forest, the SVM and XGBoost were not statistically significant. The Transformer achieved a significantly higher AUC than the FCN, although the difference was small. McNemar's test, which compares the correctness of the paired predictions at the 0.5 threshold, and the Wilcoxon signed-rank test on the per-iteration AUC values are reported in Table 6. These results show which performance differences are statistically supported and which are within the variability of the data.

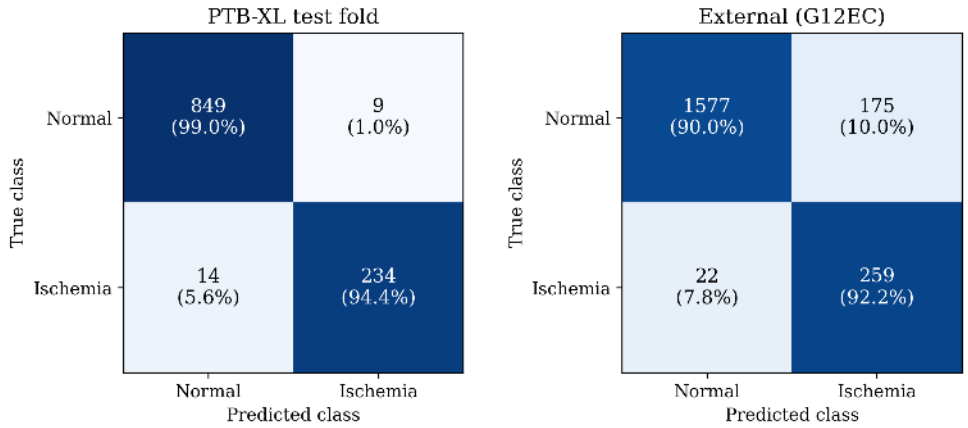


Fig. 6. Confusion matrices of the bagged FCN on the PTB-XL test set (left) and the external G12EC dataset (right); row percentages in parentheses

5.3 External Validation

On the external G12EC dataset, the bagged FCN achieved an AUC of 0.959 (95% CI 0.947–0.967), an accuracy of 90.31%, a recall of 92.17%, a specificity of 90.01% and an F1-score of 72.45% (Table 4 and Table 7). Compared with the internal test set, the AUC decreased by 0.037, which reflects differences in population, devices, recording era and labeling procedure between the two databases. The external ischemia class consists almost entirely of anterior ischemia, whereas the PTB-XL ischemia class covers all territories, so the external test evaluates generalization to a narrower but clinically important subgroup. The precision (59.68%) was lower than the recall (92.17%) because ischemia accounted for only 13.8% of the external ECGs, compared with 22.2% in PTB-XL: at a similar false positive rate, a lower prevalence produces proportionally more false positives (175 false positives versus 259 true positives). In clinical use, the decision threshold should therefore be calibrated to the prevalence and to the intended role of the model, for example a high-sensitivity threshold for triage. Among all models, the highest external AUC was obtained by the Transformer (0.965). Because no ECG from G12EC was used for training or threshold selection, these results provide an unbiased estimate of generalization to a different institution, although further multicenter validation is still needed.

Table 5. Performance of all models on the PTB-XL independent test set (mean ± SD over ten bootstrap iterations) and computational time

Model	Accuracy	Precision	Recall	Specificity	F1-score	AUC	Training time (s)	Inference (ms/ECG)
FCN	0.978 ± 0.002	0.960 ± 0.006	0.942 ± 0.008	0.989 ± 0.002	0.951 ± 0.004	0.9954 ± 0.0006	310	0.29
Encoder	0.978 ± 0.003	0.960 ± 0.012	0.940 ± 0.016	0.989 ± 0.004	0.950 ± 0.007	0.9937 ± 0.0017	741	0.70
Simple CNN	0.977 ± 0.002	0.962 ± 0.011	0.936 ± 0.010	0.989 ± 0.003	0.949 ± 0.004	0.9942 ± 0.0007	162	0.08
Transformer	0.982 ± 0.003	0.967 ± 0.010	0.953 ± 0.013	0.991 ± 0.003	0.960 ± 0.007	0.9960 ± 0.0008	168	0.11
Random Forest	0.973 ± 0.002	0.982 ± 0.004	0.896 ± 0.006	0.995 ± 0.001	0.937 ± 0.004	0.9960 ± 0.0005	32	0.15
SVM	0.978 ± 0.002	0.973 ± 0.003	0.928 ± 0.007	0.993 ± 0.001	0.950 ± 0.004	0.9956 ± 0.0005	22	0.50
XGBoost	0.980 ± 0.002	0.983 ± 0.004	0.925 ± 0.007	0.995 ± 0.001	0.953 ± 0.005	0.9963 ± 0.0004	13	0.01

Training time is per bootstrap iteration; inference time is per ECG in batch mode on the hardware described in Section 4.3. McNemar's test and the bootstrap ΔAUC use the bagged ensembles; the Wilcoxon test uses the ten per-iteration AUC values.

Table 6. Statistical comparison of the FCN with the other models on the PTB-XL independent test set

FCN vs.	Discordant pairs (FCN only / other only correct)	McNemar χ^2	McNemar p	ΔAUC (FCN – other) [95% CI]	Bootstrap p	Wilcoxon p
Encoder	10 / 11	0.00	1.000	-0.000 [-0.002, +0.001]	0.623	0.014
Simple CNN	10 / 8	0.06	0.814	+0.001 [-0.001, +0.003]	0.412	0.004
Transformer	4 / 9	1.23	0.267	-0.002 [-0.004, -0.000]	0.011	0.131
Random Forest	17 / 10	1.33	0.248	-0.001 [-0.002, +0.001]	0.347	0.084
SVM	12 / 11	0.00	1.000	-0.000 [-0.002, +0.001]	0.633	0.492
XGBoost	10 / 10	0.05	0.823	-0.001 [-0.002, +0.000]	0.192	0.006

Table 7. External validation of all models (bagged ensembles) on the G12EC dataset

Model	Accuracy	Precision	Recall	Specificity	F1-score	AUC
FCN	0.903	0.597	0.922	0.900	0.724	0.959
Encoder	0.921	0.655	0.900	0.924	0.759	0.960
Simple CNN	0.920	0.684	0.786	0.942	0.732	0.952
Transformer	0.927	0.688	0.861	0.937	0.765	0.965
Random Forest	0.904	0.665	0.609	0.951	0.636	0.952
SVM	0.921	0.689	0.779	0.943	0.731	0.961
XGBoost	0.922	0.706	0.751	0.950	0.728	0.962

5.4 Explainability Results

LIME explanations were computed for 100 correctly classified ischemic test ECGs with a cardiologist-annotated territory and 30 correctly classified normal test ECGs. Figure 7 shows a normal ECG and Figures 8–11 show ischemic ECGs from different territories, with the annotated leads marked in the panel titles. For normal ECGs, importance was low and scattered across leads, whereas for ischemic ECGs it was concentrated after the QRS complex, in the ST-segment and T-wave, predominantly in leads belonging to the annotated territory.

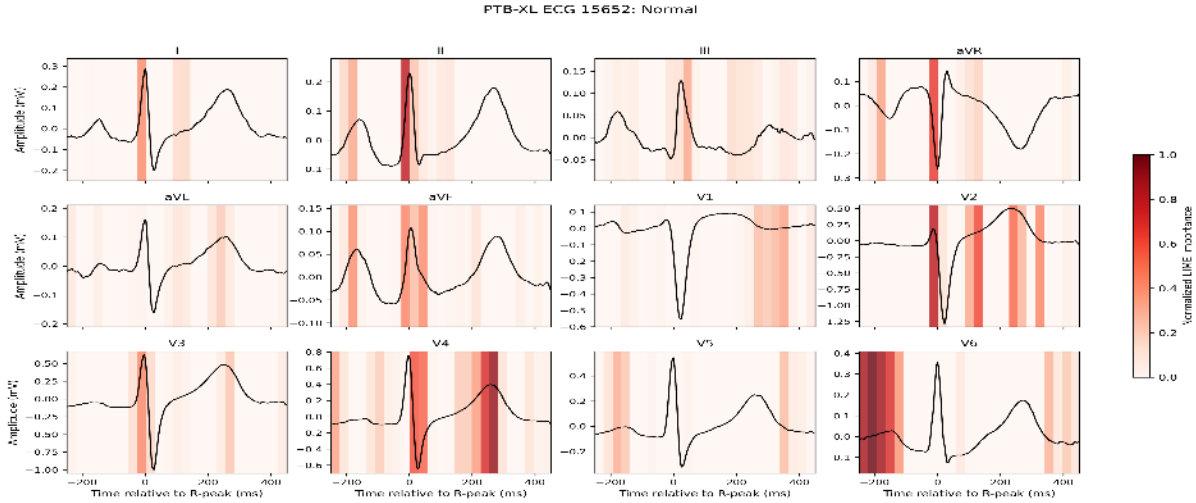


Fig. 7. LIME heat map of a normal ECG from the independent test set. Importance values are low and spread without a consistent pattern

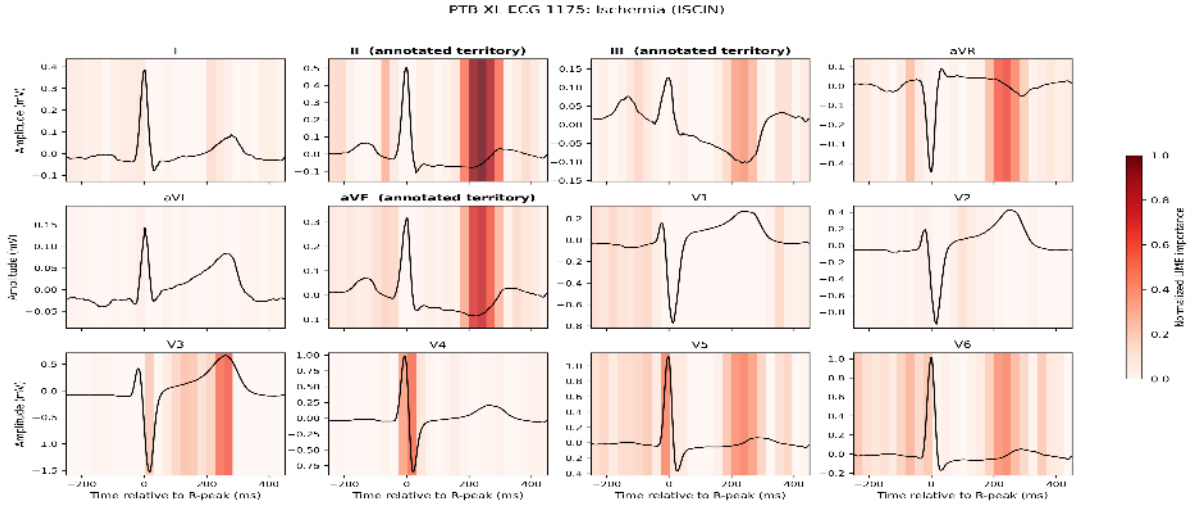


Fig. 8. LIME heat map of an ischemic ECG (PTB-XL ECG 1175, cardiologist statement ISCIN). Bold panel titles mark the annotated territory; darker red indicates higher contribution to the ischemia prediction

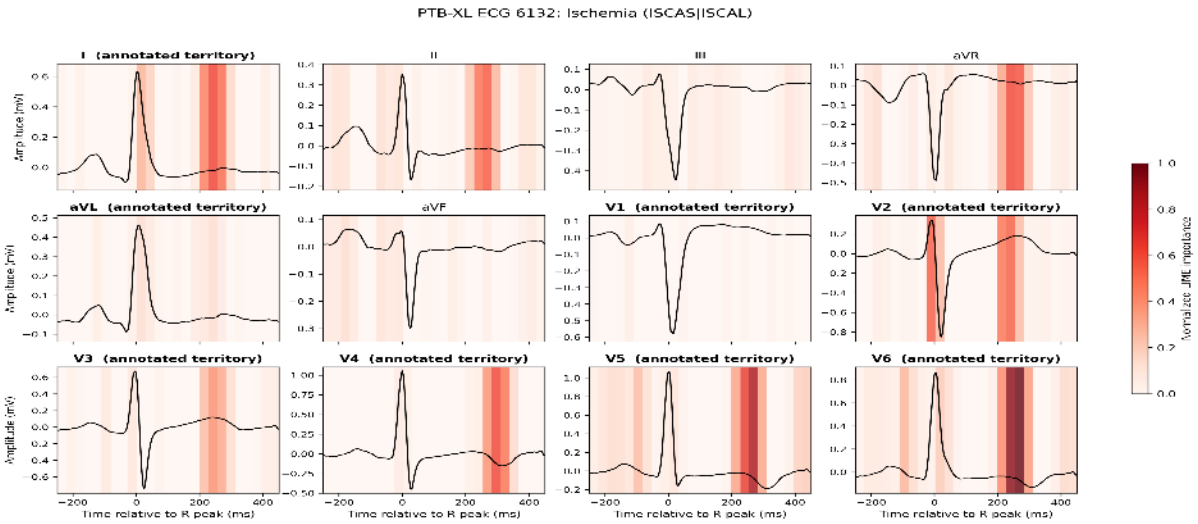


Fig. 9. LIME heat map of an ischemic ECG (PTB-XL ECG 6132, cardiologist statement ISCAS|ISCAL). Bold panel titles mark the annotated territory; darker red indicates higher contribution to the ischemia prediction

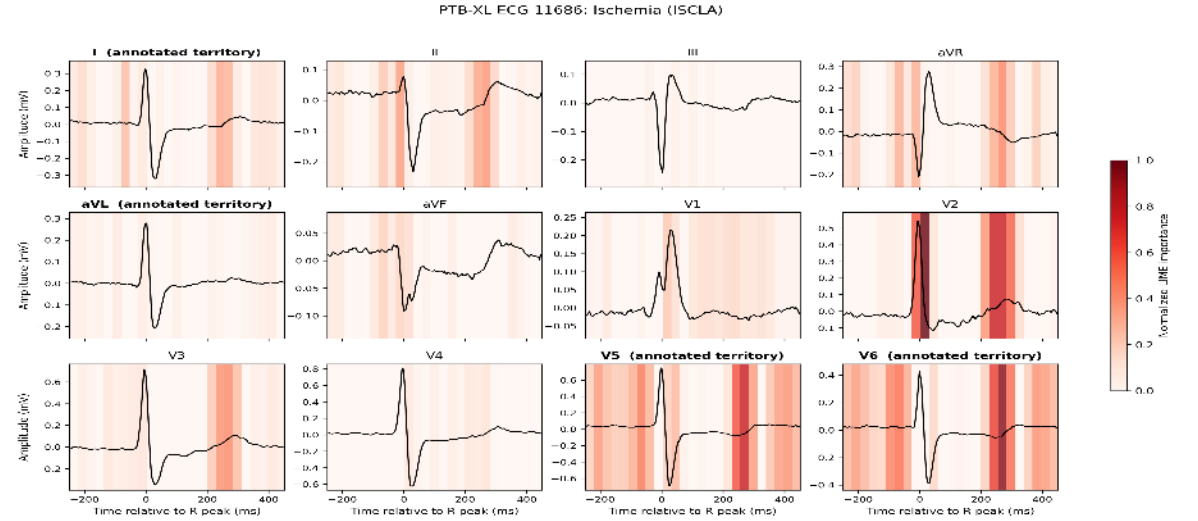


Fig. 10. LIME heat map of an ischemic ECG (PTB-XL ECG 11686, cardiologist statement ISCLA). Bold panel titles mark the annotated territory; darker red indicates higher contribution to the ischemia prediction

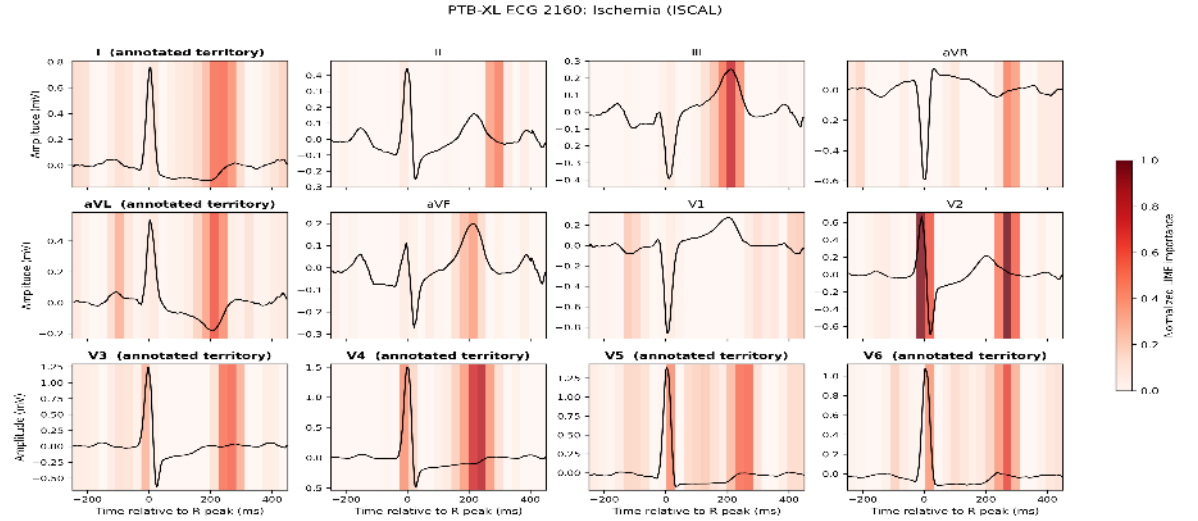


Fig. 11. LIME heat map of an ischemic ECG (PTB-XL ECG 2160, cardiologist statement ISCAL). Bold panel titles mark the annotated territory; darker red indicates higher contribution to the ischemia prediction

Table 8 summarizes the quantitative evaluation. On average, 57.3% ± 15.8% of the positive importance was located in the annotated territory leads, compared with 47.9% expected by chance (one-sided Wilcoxon $p < 0.001$). The most important lead belonged to the annotated territory in 68.0% of the ECGs (chance 47.9%; binomial $p < 0.001$), and at least one of the three most important leads did so in 93.0% (chance 85.6%; $p = 0.019$). Temporally, 64.7% ± 7.2% of the importance fell in the ST–T window, which covers 48% of the beat ($p < 0.001$), whereas only 12.4% fell in the isoelectric baseline segments that cover 20% of the beat ($p < 0.001$). The explanations were reproducible across random seeds (mean Spearman correlation 0.87 ± 0.08; Jaccard overlap of the top 30 features 0.80 ± 0.11), and the local surrogate models fitted the classifier well (mean R^2 0.87). In the deletion test, removing the 15 most important segments reduced the ischemia probability by 0.266 on average, compared with 0.013 for 15 random segments ($p < 0.001$); for 30 segments, the reductions were 0.503 and 0.021 ($p < 0.001$).

5.5 Computational Cost

Training one FCN required 310 ± 62 s (27 epochs on average) on the Quadra M620 GPU with a peak GPU memory of 382 MB. Inference took 0.29 ms per ECG in batch mode on the GPU and 3.2 ms for a single ECG on the CPU; with the ten-model ensemble, the CPU time remains below 0.1 s per ECG. Pre-processing (cleaning, R-peak detection and median-beat extraction) and one LIME explanation (2,000 perturbations) add computation mainly on the CPU, and explanations can be generated on demand only for ECGs that a clinician wants to review. The model file of the FCN is about 1.5 MB, so the framework can run on standard hospital workstations without specialized hardware. In comparison, the encoder required 741 s per training run and has 3,221,762 parameters (Table 5).

Table 8. Quantitative evaluation of the LIME explanations on the PTB-XL independent test set

Measure	LIME result	Chance level	p-value
Territory mass (share of importance in annotated leads)	57.3% ± 15.8%	47.9%	<0.001
Top-1 lead in annotated territory	68.0%	47.9%	<0.001
At least one top-3 lead in annotated territory	93.0%	85.6%	0.019
Importance in ST–T window (ischemic ECGs)	64.7% ± 7.2%	48%	<0.001
Importance in isoelectric baseline (ischemic ECGs)	12.4%	20%	<0.001
Importance in QRS complex (ischemic ECGs)	13.2%	16%	–
Probability drop, top-15 vs. random-15 deletion	0.266 vs. 0.013	–	<0.001
Probability drop, top-30 vs. random-30 deletion	0.503 vs. 0.021	–	<0.001
Reproducibility: Spearman correlation (3 seeds)	0.87 ± 0.08	–	–
Reproducibility: Jaccard overlap of top 30 features	0.80 ± 0.11	–	–

Expert reference: cardiologist ischemia-territory statements of PTB-XL. One-sided Wilcoxon signed-rank tests, except binomial tests for the top-1 and top-3 localization rates.

6. Comparative Analysis and Discussion

Table 9 places the results in the context of previous studies. Such comparisons must be interpreted with caution: the studies differ in task (ischemia, occlusion myocardial infarction, culprit vessel or general cardiac severity), in data size and source, in whether the data were split at the patient level and in whether external validation was performed. A model trained on a few hundred signals cannot be compared fairly with a model trained on thousands of clinically diverse patients, as in Al-Zaiti et al. [10], and higher numbers on small or record-level splits do not necessarily indicate a better model. For this reason, the revised study uses a large public cohort, a patient-level independent test set and an external dataset, and it compares the FCN with baselines under an identical protocol (Tables 5–7), which is more informative than the cross-study comparison in Table 9.

Table 9. Comparison of the proposed model with previous studies (results as reported in each study)

Study	Data	Task	Method	Reported performance	External validation
Sarıçam et al. [14]	74 patients	Lateral MI culprit vessel	ECG criteria	PPV 100% (OM), 98% (D1)	No
Al-Zaiti et al. [21]	7,313 patients	Occlusion MI	Random forest	AUROC 0.91	Yes
Zhong et al. [23]	LTSTDB	Ischemia	SVM	Accuracy 97.98%	No
Panchal et al. [25]	929 images	Cardiac severity	VGG-16	Accuracy 70.71%	No
Sun et al. [30]	499 patients	Ischemia (non-diagnostic ECG)	Ensemble learning	Accuracy 89.10%; 91.11% on PTB	Yes
Zhao et al. [31]	Training + independent cohort	Ischemia	ECG/VCG features + ML	Accuracy 90.3%; external AUC 0.814	Yes
Present study (FCN)	11,018 ECGs / 10,135 patients	Ischemia vs. normal	FCN + LIME	Accuracy 97.82%, AUC 0.995; external AUC 0.959	Yes

Within the controlled benchmark, all seven models achieved high and similar performance: the mean AUC on the independent test set varied by only 0.003 across models. The AUC of the FCN did not differ significantly from that of the encoder, the simple CNN, random forest, the SVM and XGBoost. The Transformer achieved a small but statistically significant AUC advantage (ΔAUC 0.0018). This indicates that, once ECGs are cleaned and summarized as median beats, most of the diagnostic information is accessible to both deep and classical models, and that the differences between architectures are small compared with the effect of data quality, cohort size and evaluation design. The FCN is therefore not presented as the most accurate classifier, but as an accurate, stable and computationally light model whose predictions can be explained and quantitatively checked. Aside from predictive accuracy, the primary benefit of the proposed framework is that its explanations can be validated. Since LIME explanations are derived from local perturbations, it is necessary to validate that regions selected by LIME correspond to diagnostic ST-segment changes and are not driven by noisy fluctuations of the baseline. Quantitative evaluation presented in Section 5.4 aim to answer this question: importance did not concentrate uniformly throughout the recording, it was focused on the ST–T window and concentrated in the cardiologist-annotated leads significantly better than chance, the baseline received less importance than would be expected by chance, the explanations were stable across random seeds, and deletion of highlighted segments changed the prediction much more than deletion of random segments. Additionally, by using a physiologically plausible reference beat as our zero-perturbation, as opposed to zeros, and aggregating our inputs into segments of 28 ms we further limit the impact of noisy individual samples.

Clinically, we see this framework as augmenting, but not replacing, cardiologists. Several implementation options are possible. For triage, the model could bump ECGs with a high predicted probability of ischemia to the top of the reading worklist so they get read first. Alternatively, in facilities lacking on-site cardiology coverage, it could serve as a second reader flagging ECGs for remote cardiologists. The heat maps could then direct the cardiologist's attention to the leads and ST–T regions that factored most into the prediction. Focusing a remote reader on only suspicious ECGs and providing intelligently cropped images would allow rapid validation of predictions, which may decrease decision latency. Because inference time is on the order of milliseconds, the deep learning model introduces no perceived latency to ECG acquisition and could be seamlessly integrated into ECG management systems. A decrease in total diagnostic time has not been quantified in this work and would need to be validated in a reader study or prospective trial. In emergency settings, several limitations apply. The model distinguishes ischemic from normal ECGs and was not trained to separate ischemia from other conditions that alter the ST-segment, such as left ventricular hypertrophy, bundle branch block, pericarditis or electrolyte disturbances, nor to identify ST-elevation myocardial infarction requiring immediate reperfusion. A negative prediction

must therefore never delay guideline-based care, and ECG findings must be interpreted together with symptoms, troponin and serial ECGs [49]. The median-beat representation discards rhythm information and beat-to-beat variation, and the model analyzes a single 10-second recording rather than dynamic changes over time. This study has further limitations. First, the classes are imbalanced (approximately one ischemic ECG for every 3.5 normal ECGs); class weighting was used, and precision, recall, F1-score and AUC are reported in addition to accuracy, but the positive predictive value will differ in populations with a different prevalence. Second, the study is retrospective, and no prospective clinical validation has been performed. Third, although an external dataset was used, the data come from only two sources, and PTB-XL was recorded between 1989 and 1996 with devices from one manufacturer; broader multicenter validation with contemporary devices is needed. Fourth, ischemia labels are based on ECG statements rather than on angiographic or perfusion imaging reference standards, the G12EC labels were not adjudicated in the same manner as PTB-XL, and the external ischemia class is limited almost entirely to anterior ischemia, so external performance for other territories remains to be established. Fifth, deep models may overfit, particularly when the number of training examples is limited; the patient-level test set, bootstrap confidence intervals and external validation reduce but do not eliminate this risk. Sixth, the analysis used the 100 Hz version of the signals, which may attenuate high-frequency QRS information. Finally, the explanations were evaluated against lead-level territory statements rather than against sample-level annotations drawn by cardiologists on the heat maps, and a formal reader study of the clinical usefulness of the explanations remains future work.

7. Conclusions

This study presented a deep learning framework for the automated detection of myocardial ischemia from 12-lead ECG signals that combines a Fully Convolutional Network with Local Interpretable Model-Agnostic Explanations. ECGs were cleaned, segmented into beats and summarized as lead-wise median beats, which provide a compact and noise-robust input that preserves the morphology of the P-wave, QRS complex, ST-segment and T-wave. Using 11,018 ECGs from 10,135 patients of the public PTB-XL database with strict patient-level separation, the FCN achieved an accuracy of 97.82% ± 0.19%, a recall of 94.23%, a specificity of 98.86%, an F1-score of 95.10% and an AUC of 0.995 on the independent test set, and an AUC of 0.959 on the external G12EC dataset. The revised evaluation provides a more reliable picture than the initial submission. All performance values are reported with standard deviations across ten bootstrap iterations and 95% confidence intervals, precision, recall, F1-score and confusion matrices complement accuracy and AUC, and statistical tests show which differences between the FCN and the encoder, simple CNN, Transformer, random forest, SVM and XGBoost baselines are significant. The AUC of the FCN did not differ significantly from that of the encoder, the simple CNN, random forest, the SVM and XGBoost. The Transformer achieved a small but statistically significant AUC advantage (ΔAUC 0.0018). The FCN trains in minutes on a laptop GPU and classifies an ECG in milliseconds, so its value lies less in a higher accuracy than in the combination of stable performance with explanations that can be verified. The main contribution of the study is the quantitative validation of the explanations. Segment-wise LIME explanations concentrated on the ST–T region and on the leads of the cardiologist-annotated ischemia territory significantly more than expected by chance, gave little weight to the isoelectric baseline, were reproducible across random seeds and were faithful to the model according to the deletion test. These findings indicate that the model relies on clinically meaningful ECG features rather than on noise, which is an important requirement for the trust and adoption of artificial intelligence in cardiology. The framework could support clinicians by prioritizing ECGs for review, acting as a second reader where cardiologists are scarce and highlighting the leads and waveform segments that justify each prediction. The results should be interpreted in light of the retrospective design, the class imbalance, the limited number of data sources, the ECG-based reference labels and the absence of prospective clinical validation. Future research will extend the model to multi-class differentiation of ischemia from other causes of ST–T abnormalities and to ischemia localization, use contemporary multicenter datasets and higher sampling rates, compare LIME with other explanation methods such as SHAP and Integrated Gradients, evaluate the explanations in a reader study with cardiologists, and assess the effect of the framework on diagnostic accuracy and decision time in a prospective clinical study.

Data Access Statement

The ECG data used in this study are publicly available. PTB-XL (version 1.0.3) is available from PhysioNet at https://doi.org/10.13026/kfzx-aw45, and the Georgia 12-lead ECG Challenge database is available from PhysioNet as part of the PhysioNet/Computing in Cardiology Challenge 2020 (https://physionet.org/content/challenge-2020/1.0.2/). The analysis code is available from the corresponding author upon reasonable request.

Acknowledgement

The authors acknowledge that this study is supported by 1) Department of Administrative and Financial Affairs, University of Telafer, Nineveh, Iraq, 2) Department of Reconstruction and Projects, University of Baghdad, Baghdad, Iraq, and 3) Department of AI, College of Engineering, University of Zakho, Zakho, KRG-Iraq.

References

[1] Jagannathan R, Patel SA, Ali MK, Narayan KV. Global updates on cardiovascular disease mortality trends and attribution of traditional risk factors. Curr Diab Rep. 2019;19(7):44. https://doi.org/10.1007/s11892-019-1161-2

[2] Almansouri NE, Awe M, Rajavelu S, Jahnavi K, Shastry R, Hasan A, et al. Early diagnosis of cardiovascular diseases in the era of artificial intelligence: An in-depth review. Cureus. 2024;16(3):e56333. https://doi.org/10.7759/cureus.55869

[3] Birnbaum Y, Rankinen J, Jneid H, Atar D, Nikus K. The role of ECG in the diagnosis and risk stratification of acute coronary syndromes: an old but indispensable tool. Curr Cardiol Rep. 2022;24(2):109-118. https://doi.org/10.1007/s11886-021-01628-7

[4] Mohialdin SH, Abdulrahman LQ, Al-Yoonus MH, Abduljabbar ZA, Honi DG, Nyangaresi VO, et al. Utilizing machine learning for the early detection of coronary heart disease. Eng Technol Appl Sci Res. 2024;14(5):17363-17375. https://doi.org/10.48084/etasr.8171

[5] Moreno-Sánchez PA, García-Isla G, Corino VD, Vehkaoja A, Brukamp K, van Gils M, et al. ECG-based data-driven solutions for diagnosis and prognosis of cardiovascular diseases: A systematic review. Comput Biol Med. 2024;172:108235. https://doi.org/10.1016/j.compbiomed.2024.108235

[6] Uwaechia AN, Ramli DA. A comprehensive survey on ECG signals as new biometric modality for human authentication: Recent advances and future challenges. IEEE Access. 2021;9:97760-97802. https://doi.org/10.1109/ACCESS.2021.3095248

[7] Moody JB, Poitrasson-Rivière A, Renaud JM, et al. A foundation transformer model with self-supervised learning for ECG-based assessment of cardiac and coronary function. NEJM AI. 2025;2(12):AIoa2500164. https://doi.org/10.1056/AIoa2500164

[8] Apostol A, Nutu M. Arrhythmia classification from 12-lead ECG signals using convolutional and transformer-based deep learning models. arXiv preprint arXiv:2502.17887. 2025.

[9] Zhang J, Zhang H, Wei T, Kang P, Tang B, Wang H. Predicting angiographic coronary artery disease using machine learning and high-frequency QRS. BMC Med Inform Decis Mak. 2024;24(1):217. https://doi.org/10.1186/s12911-024-02620-1

[10] Taleban A, Sparapani R, Noffke P, Zlochiver S, Lu Q, Widlansky ME, et al. Explainable artificial intelligence in electrocardiography: A systematic review. Biomed Signal Process Control. 2025;114:109325. https://doi.org/10.1016/j.bspc.2025.109325

[11] Strodthoff N, Strodthoff C. Detecting and interpreting myocardial infarction using fully convolutional neural networks. Physiol Meas. 2019;40(1):015001. https://doi.org/10.1088/1361-6579/aaf34d

[12] Jahmunah V, Ng EYK, Tan RS, Oh SL, Acharya UR. Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals. Comput Biol Med. 2022;146:105550. https://doi.org/10.1016/j.compbiomed.2022.105550

[13] Anand A, Kadian T, Shetty MK, Gupta A. Explainable AI decision model for ECG data of cardiac disorders. Biomed Signal Process Control. 2022;75:103584. https://doi.org/10.1016/j.bspc.2022.103584

[14] Sarıçam E, Erdol MA, Bozkurt E, Ilkay E, Cantekin OF. New ECG algorithm for the prediction of culprit vessel in acute myocardial infarction involving lateral part of the ventricle: Ilkay classification. Int J Gen Med. 2023;16:2643-2651. https://doi.org/10.2147/IJGM.S416376

[15] Wang Z, Yan W, Oates T. Time series classification from scratch with deep neural networks: A strong baseline. In: 2017 International Joint Conference on Neural Networks (IJCNN). IEEE; 2017. p. 1578-1585. https://doi.org/10.1109/IJCNN.2017.7966039

[16] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 1135-1144. https://doi.org/10.1145/2939672.2939778

[17] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data. 2020;7(1):154. https://doi.org/10.1038/s41597-020-0495-6
[18] Wagner P, Strodthoff N, Bousseljot R, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3). PhysioNet. 2022.
[19] Perez Alday EA, Gu A, Shah AJ, Robichaux C, Wong AKI, Liu C, et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiol Meas. 2020;41(12):124003. https://doi.org/10.1088/1361-6579/abc960
[20] Singh PK, Akhtar S, Gupta A, Singh S. The clinical relevance of ECG parameters in the prediction of cardiac mortality: A comprehensive review. Open Bioinform J. 2024;17(1):e18750362287847. https://doi.org/10.2174/0118750362295563240620111209
[21] Al-Zaiti SS, Martin-Gill C, Zègre-Hemsey JK, Bouzid Z, Faramand Z, Alrawashdeh MO, et al. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. Nat Med. 2023;29(7):1804-1813. https://doi.org/10.1038/s41591-023-02396-3
[22] Erol M, Tezcan H, Duran M, Uygun T, Kurtipek E, Onner H. The role of myocardial perfusion imaging in predicting myocardial ischemia in patients diagnosed with long COVID. Int J Cardiovasc Imaging. 2023;39(11):2279-2284. https://doi.org/10.1007/s10554-023-02928-5
[23] Zhong D, Huang L, Jin S, An Y, Zhu S, Li J. Automatic electrocardiograph diagnosis of myocardial ischemia with support vector machine. Digital Medicine. 2023;9(2):e00009. https://doi.org/10.1097/DM-2023-00009
[24] Zaichenko KV, Kordyukova AA, Sonin DL, Galagudza MM. Ultra-high-resolution electrocardiography enables earlier detection of transmural and subendocardial myocardial ischemia compared to conventional electrocardiography. Diagnostics. 2023;13(17):2795. https://doi.org/10.3390/diagnostics13172795
[25] Panchal N, Raikar MM, Baligar VP. Prediction of cardiac severity based on ECG images using deep learning models. In: 2024 Second International Conference on Advances in Information Technology (ICAIT). IEEE; 2024. p. 1-5. https://doi.org/10.1109/ICAIT61638.2024.10690399
[26] Alahdab F, Saad MB, Ahmed AI, Al Tashi Q, Aminu M, Han Y, et al. Development and validation of a machine learning model to predict myocardial blood flow and clinical outcomes from patients' electrocardiograms. Cell Rep Med. 2024;5(10):101763. https://doi.org/10.1016/j.xcrm.2024.101746
[27] Chen Y, Zhang X, Ye Q, Zhang X, Cao N, Li SY, et al. Machine learning-based prediction model for myocardial ischemia under high altitude exposure: a cohort study. Sci Rep. 2024;14(1):686. https://doi.org/10.1038/s41598-024-51202-8
[28] Jorda A, Pecho T, Horvath LC, Nishani E, Bull LE, Bergmann F, et al. Association of electrocardiogram findings with clinical outcomes in patients with chronic coronary syndrome: an analysis of the ischemia trials. Am J Med. 2024;137(10):974-982.
[29] Al Younis SM, Hadjileontiadis LJ, Khandoker AH, Stefanini C, Soulaidopoulos S, Arsenos P, et al. Prediction of heart failure patients with distinct left ventricular ejection fraction levels using circadian ECG features and machine learning. PLoS One. 2024;19(5):e0302639. https://doi.org/10.1371/journal.pone.0302639
[30] Sun Q, Liang C, Chen T, Ji B, Liu R, Wang L, et al. Early detection of myocardial ischemia in 12-lead ECG using deterministic learning and ensemble learning. Comput Methods Programs Biomed. 2022;226:107124. https://doi.org/10.1016/j.cmpb.2022.107124
[31] Zhao X, Zhang J, Gong Y, Xu L, Liu H, Wei S, et al. Reliable detection of myocardial ischemia using machine learning based on temporal-spatial characteristics of electrocardiogram and vector cardiogram. Front Physiol. 2022;13:854191. https://doi.org/10.3389/fphys.2022.854191
[32] Wang S, Li J, Sun L, Cai J, Wang S, Zeng L, et al. Application of machine learning to predict the occurrence of arrhythmia after acute myocardial infarction. BMC Med Inform Decis Mak. 2021;21(1):122. https://doi.org/10.1186/s12911-021-01667-8
[33] Sun Q, Wang L, Li J, Liang C, Yang J, Chen Y, et al. Multi-phase ECG dynamic features for detecting myocardial ischemia and identifying its etiology using deterministic learning. Biomed Signal Process Control. 2024;88:105498. https://doi.org/10.1016/j.bspc.2023.105498
[34] Wang L, Hao J, Zhou TH, Bi T. Heart disease prediction model based on HRV emotional features fusion. In: Proceedings of the 2024 8th International Conference on High Performance Compilation, Computing and Communications. 2024. p. 71-75. https://doi.org/10.1145/3675018.3675773
[35] Seoni S, Molinari F, Acharya UR, Lih OS, Barua PD, García S, et al. Application of spatial uncertainty predictor in CNN-BiLSTM model using coronary artery disease ECG signals. Inf Sci. 2024;665:120383. https://doi.org/10.1016/j.ins.2024.120383
[36] Gómez N, Ramírez J, Martín-Yebra A, Demidova MM, Platonov P, Martínez JP, et al. T-wave peak-to-end changes quantified by time-warping predicts ventricular fibrillation in a porcine myocardial infarction model. IEEE Trans Biomed Eng. 2024;71(10):2991-2999. https://doi.org/10.1109/TBME.2024.3404254

[37] Abdullah ZF, Abadian A, Sánchez Mejorada CG, Torres Ruiz MJ, Quintero Téllez R. Integration of machine learning techniques in ECG-based multiclass arrhythmia classification with explainability analysis. Biosensors. 2026;16(6):326. https://doi.org/10.3390/bios16060326
[38] Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, et al. NeuroKit2: A Python toolbox for neurophysiological signal processing. Behav Res Methods. 2021;53(4):1689-1696. https://doi.org/10.3758/s13428-020-01516-y
[39] Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Deep learning for time series classification: a review. Data Min Knowl Discov. 2019;33(4):917-963. https://doi.org/10.1007/s10618-019-00619-1
[40] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. 2017. p. 5998-6008.
[41] Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. https://doi.org/10.1023/A:1010933404324
[42] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-297. https://doi.org/10.1023/A:1022627411411
[43] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785-794. https://doi.org/10.1145/2939672.2939785
[44] Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman & Hall/CRC; 1993. https://doi.org/10.1007/978-1-4899-4541-9
[45] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153-157. https://doi.org/10.1007/BF02295996
[46] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):e215-e220. https://doi.org/10.1161/01.CIR.101.23.e215
[47] Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. IEEE J Biomed Health Inform. 2021;25(5):1519-1528. https://doi.org/10.1109/JBHI.2020.3022989
[48] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. 2019. p. 8024-8035.
[49] Thygesen K, Alpert JS, Jaffe AS, Chaitman BR, Bax JJ, Morrow DA, et al. Fourth universal definition of myocardial infarction (2018). Circulation. 2018;138(20):e618-e651. https://doi.org/10.1161/CIR.0000000000000617