ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

TrustECG-DG: An Explainable, Uncertainty-Calibrated and Domain-Generalized Deep Learning Framework for Reliable Cardiac Abnormality Detection across Heterogeneous ECG Datasets

Sairam Vallabhuni [1*], Dr. P. V. Naganjaneyulu [2]

[1] Research Scholar, Department of ECE, Mizoram University, Aizawl, Mizoram, India-796004.

[2] Professor, Department of ECE, Mizoram University, Aizawl, Mizoram, India-796004.

***Corresponding Author**

DOI: https://doi.org/10.51583/IJLTEMAS.2026.151000077

Received: 26 September 2026; Accepted: 01 October 2026; Published: 07 November 2026

ABSTRACT

While deep learning has made significant advances in cardiac abnormality detection using electrocardiograms (ECG), current achievements are based on landmark classification accuracy, which may not necessarily translate to reliable performance on heterogeneous datasets, calibrated confidence or clinically interpretable predictions. Current ECG learning methods are often domain-dependent, and they treat explainability and uncertainty estimation as separate components. To address these drawbacks, this study proposes TrustECG-DG, an explainable, uncertainty-calibrated, and domain-generalised deep learning framework for reliable cardiac abnormality detection from multi-lead ECG signals. It combines a multi-scale convolutional neural network to learn morphological features with a temporal Transformer to model long-range rhythm dependencies. Cross-attention between morphological and temporal representations is used bidirectionally. The framework jointly learns domain-invariant features, class-aware contrastive alignment, and covariance-based feature alignment to improve domain robustness. A deep ensemble of 5 models is then used, along with temperature scaling, to quantify prediction uncertainty and improve probability calibration. To facilitate interpretation, Integrated Gradients and Transformer attention rollout methods explain the ECG regions the model relies on for its predictions, and an attentive prediction mechanism is designed to pass uncertain cases to an expert for evaluation. Experimental results show that the proposed framework achieves 98.7% accuracy on the test set. Beyond classification performance, the framework is assessed using class-wise predictive metrics, cross-dataset generalisation, ablation analysis, calibration, uncertainty estimation, and explanation analysis. The resulting framework offers a unified solution that integrates accurate ECG classification, robustness, calibrated confidence, interpretability, and selective decision-making, providing a more reliable basis for intelligently detecting cardiac abnormalities using an ECG.

Keywords: Electrocardiogram, ECG classification, deep learning, Transformer, domain generalisation, uncertainty estimation, explainable artificial intelligence, Integrated Gradients, contrastive learning, cardiac abnormality detection.

INTRODUCTION

The electrocardiogram (ECG) is a crucial tool for assessing the cardiovascular system, as it provides useful information on cardiac rhythm, conduction patterns, waveform morphology, and electrical abnormalities. ECG interpretation has traditionally relied on expert analysis of characteristic ECG components and their timing. However, the growing volume of large-scale ECG data creates opportunities to design deep-learning systems that can automatically learn complex cardiac patterns directly from ECG signals. Recent work has shown the high performance of convolutional neural networks (CNNs), recurrent networks, Transformers, and hybrid architectures for arrhythmia and cardiovascular abnormality classification. Concurrently, explainability, uncertainty estimation, calibration and generalisation are increasingly the focus of recent

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

research due to the fact that the accuracy of a classification does not necessarily reflect the reliability of the artificial intelligence-based ECG method.

Deep learning methods have also been found to be effective in ECG classification; however, their performance may vary significantly from one dataset to another based on dataset characteristics. In particular, recently developed explainable learning algorithms for ECGs have incorporated traditional machine learning models into deep learning models using algorithms like SHAP, LIME, Grad-CAM, and Integrated Gradients to include information about features that contribute to predictions. Other studies have used CNN-BiLSTM architectures, multimodal ECG representations, and attention mechanisms to extract morphological and temporal aspects of cardiac signals. These advancements suggest that deep learning can greatly improve automated ECG interpretation. Still, very accurate predictions don't guarantee explanations—they don't necessarily mean consistent behaviour across different clinical or acquisition settings.

However, one key challenge is domain variation in ECG data. Patient populations, recording environments, acquisition devices, and signal characteristics may vary between ECG datasets, as may diagnostic distributions and annotation practices. Thus, a model that learns highly discriminative features from one dataset may learn dataset-specific aspects that are not common to other datasets. Recently, cross-dataset domain generalisation and unified multi-dataset ECG learning have been explored in detail, highlighting the need for representation learning that retains clinically relevant cardiac information while minimising the reliance on the specifics of a particular dataset. This is also significant for practical ECG applications, as the model may encounter recordings that differ from those it was trained on.

Confidence and uncertainty in prediction are another key constraint. A deep learning algorithm can produce a high prediction probability even when the data (ECG) is hard to classify or far from the training-set distribution. These issues have been explored more recently in the context of ECG, as reviewed below, focusing on reducing over-confident decisions: (1) spatial uncertainty estimation, (2) direct uncertainty quantification, and (3) calibrated prediction. In studying calibration, other measures like expected calibration error and Brier score have been used to assess whether predicted probabilities reflect the reliability of the predictions. These developments suggest that a reliable ECG system should not only detect a cardiac abnormality, but also assess the degree of confidence in that determination.

A further requirement for trustworthy ECG analysis is explainability. Previous research has applied Grad-CAM, SHAP, LIME, saliency maps, Integrated Gradients, and attention-based approaches to pinpoint image regions important to the model's prediction. These approaches can provide evidence at the feature or visual-content level for predictions. Post-classification explanations do not need to show that the model uses meaningful information from the ECG to make the classification. In recent years, systematic studies have therefore emerged, pointing to the need to improve explanation quality, generalisation, class imbalance, uncertainty, and external validation in ECG deep-learning systems.

Recent advances in representation learning offer more opportunities to overcome these limitations. Contrastive learning has been explored to improve the representation and generalisation of ECG data, and Transformer-based architectures have been proposed to capture long-range temporal dependencies in cardiac signals. Combining convolutional network-based feature extraction with temporal or attention-based learning may capture local waveform morphology and broader rhythm patterns simultaneously. These advances support complementary representation-learning mechanisms rather than a single feature-extraction architecture.

To overcome these obstacles, this study presents TrustECG-DG, an explainable, uncertainty-aware, and domain-generalised deep-learning model for cardiac abnormality detection on diverse ECG databases. The proposed architecture features multiple CNN branches operating in parallel at different scales to capture local ECG morphology, with a temporal Transformer to model long-range rhythm dependencies. These representations are coupled via bidirectional cross-attention, enabling morphological information to interact with temporal information. We further optimise this with domain adversarial learning, class-aware contrastive alignment, and CORAL-based covariance alignment to minimise domain-specific variation while preserving

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

cardiac abnormality information. The architecture accepts standard 12-lead ECG input and a common diagnostic representation for normal ECG, atrial fibrillation/flutter, premature beats, bundle branch block, conduction abnormalities, and ST–T abnormalities.

The framework also includes a 5-model deep ensemble, which provides a measure of prediction uncertainty, and temperature scaling, which maps it back to a probability. The proposed selective prediction mechanism does not treat all predictions equally; instead, it identifies cases where prediction uncertainty is acceptable and those that require expert review. To provide insight into the model's decision, we create an evidence map of the ECG regions that contributed to the prediction by merging Interpretability data with Transformer attention rollout. To provide interpretability, we create an evidence map of the ECG regions that contributed to the model's prediction by merging Interpretability data with Transformer attention rollout. The architecture thus links classification, cross-dataset representation learning, uncertainty estimation, calibration, selective prediction, and explainability into one framework.

In addition to the model's classification results, the study assesses it through comparative experiments, ablation analysis, cross-dataset evaluation, calibration analysis, uncertainty estimation, selective prediction, and explainability analysis. These experiments aim not only to assess the model's ability to classify ECG abnormalities accurately, but also to determine whether it can learn a representation robust to data variability and whether the resulting predictions provide meaningful confidence and interpretable evidence.

The main contributions of this study are therefore summarised as follows:

1. A unified ECG representation-learning framework that combines multi-scale CNN-based morphology extraction, temporal Transformer learning, and bidirectional cross-attention for joint modelling of local waveform characteristics and long-range rhythm dependencies.

2. A domain-generalised learning strategy that integrates domain-adversarial learning, class-aware contrastive alignment, and CORAL-based feature alignment to reduce dataset-specific representation differences.

3. An uncertainty-aware and calibrated prediction mechanism based on a five-model deep ensemble and temperature scaling, enabling the framework to distinguish reliable predictions from uncertain cases.

4. An explainability mechanism combining Integrated Gradients and Transformer attention rollout to identify ECG regions contributing to cardiac abnormality predictions.

5. A selective prediction strategy that uses predictive uncertainty to support automated classification for sufficiently confident cases while directing high-uncertainty cases toward expert review.

In conclusion, the proposed TrustECG-DG framework shifts the focus of ECG deep learning from accuracy-driven classification to a more holistic approach that integrates accuracy, domain robustness, calibrated confidence, uncertainty awareness, and interpretability. The following sections explain related research, the proposed methodology, the experimental configuration, and the study's results and conclusions.

RELATED WORK

In recent years, identifying cardiac abnormalities from ECG signals has become a hot area of research, with work shifting from traditional machine learning toward deep learning, explainable artificial intelligence (XAI), uncertainty-aware prediction, and cross-dataset generalisation. These advances have greatly helped with automatic ECG interpretation, but most approaches address only one or two of these requirements. The relevant studies are especially relevant to the proposed TrustECG-DG framework:

Explainable Deep Learning for ECG Analysis

Aladhadh et al. [1] designed a framework for cardiovascular risk prediction based on explainable machine

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

learning and deep learning methods, utilising the clinical features derived from an ECG and traditional cardiovascular risk factors. The authors evaluated several machine learning and deep learning models and introduced SHAP-based interpretation, probability calibration, uncertainty quantification, and feature ablation. Among the machine learning models, XGBoost performed best; among the deep learning models, BiLSTM performed best, with the highest ROC-AUC. However, the study was mainly a benchmarking platform that can be replicated on a synthetic dataset, and the authors highlighted the need to test on large-scale, real-world ECG datasets with different clinical outcomes.

Tenepalli and Navamani [4] suggested an arrhythmia detection method using a hybrid CNN-Transformer network. The 1D ECG signals from the MIT-BIH dataset were transformed into 2D waveform images, and EfficientNetB0 was integrated with a transformer-based attention module to learn both morphological and long-range relationships. The model achieved 99% accuracy on PVC classification, and Grad-CAM was used to identify the crucial areas of the ECG. However, the framework primarily tested binary PVC detection without domain generalisation or explicit predictive uncertainty.

To improve the trustworthiness of arrhythmia detection, Talukder et al. [6] proposed an explainable deep learning model using CNN and DNN architectures with balancing techniques such as ADASYN, SMOTE, SMOTETomek, and random oversampling. To improve interpretability, they added SHAP, LIME, and feature importance analysis. The framework achieved high classification accuracy on MITDB, PTBDB, and NSTDB. The authors, however, noted the lack of feature fusion and Transformer-related architectures as a limitation, noting a need for better temporal representation and generalisation.

Abdullah et al. [7] systematically compared three architectures: ANN, CNN and ResNet, to classify ECG arrhythmia using a common experimental protocol. To address class imbalance, they used a fine-tuned CNN with adaptive kernel sizing, regularisation, cosine-annealing learning rates, and a combined weighted cross-entropy and focal loss. The proposed model reached 98.51% accuracy and used Grad-CAM and Integrated Gradients for explanation. Moreover, the study also benchmarked the faithfulness of the attributed explanations, sanity checks, and noise robustness, highlighting the importance of quantitative, rather than visual, evaluation of explanations.

Vallabhuni et al. [8] introduced GENCardioXplain, which blends Temporal Convolutional Networks (TCN), Bi-GRU, Adaptive Batch Normalisation (AdaBN), attention, and SHAP-based explanations. The framework focused on cross-dataset generalisation, achieving 94.22% accuracy, 93.84% precision, 93.43% F1-score, and an MCC of 0.9011. We assessed robustness through cross-dataset experiments and statistical significance analysis. However, the framework focused mostly on single-lead ECG signals and used target-domain unlabeled data for adaptation, leaving room for improved multi-lead and source-only domain-generalisation methods.

Oladunni and Aneni [5] studied intermediate and late fusion approaches for fusing multimodal ECG representations in the time, frequency, and time-frequency domains. They found that intermediate fusion achieved up to 97% accuracy and outperformed late fusion. Most significantly, the proposed Explainable-AI Trustworthiness framework assessed explanation quality using metrics including mutual information, saliency stability, and ablation consistency, and analysed the trustworthiness of the explanation architecture using different measures. The presented work shows that explainability is not limited to visualisation but can also be quantified and used for a quantitative evaluation.

Abdullah et al. [7], Oladunni and Aneni [5], and other explainability-based studies confirm that XAI techniques like Grad-CAM, Integrated Gradients, SHAP, and attention-based visualisation can be leveraged to improve ECG models. However, these approaches generally focus on explaining individual models rather than combining explanation reliability with cross-dataset generalisation and predictive uncertainty.

Uncertainty-Aware and Calibrated ECG Prediction (UCAP)

Seoni et al. [2] addressed reliability through a Spatial Uncertainty Estimator (SUE) that estimates prediction reliability by quantifying the spatial overlap of discriminative Grad-CAM features. For coronary artery

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

disease classification, they generated a CNN-BiLSTM model that achieved 99.6% accuracy, 99.8% sensitivity, and 98.2% specificity. The SUE metric showed a correlation between higher reliability scores and correct ECG-segment classification, and exhibited similar properties to Deep Ensemble and Monte Carlo Dropout uncertainty estimation. This paper defines an important connection between the information derived from XAI and predictive uncertainty.

Li et al. [3] studied overconfidence in deep ECG models. They presented an uncertainty-aware architecture composed of deep convolutional residual modules, Dirichlet distributions, and Subjective Logic. Rather than using softmax probabilities, it directly modelled class distributions and uncertainty. The model also achieved 86.12% accuracy, 83.14% F1 score, 85.25% PR-AUC, and 92.87% ROC-AUC on the PhysioNet Challenge 2017 dataset, with a significant decrease in overconfidence. The study also showed that removing low-confidence predictions by applying an uncertainty threshold may improve classification accuracy.

Rani et al. [11] developed a reliable ECG classification model by integrating CNN, BiGRU, and attention mechanisms. The study did not explicitly model predictive uncertainty and focused on probabilistic calibration using the Brier Score and Expected Calibration Error (ECE). The CNN achieved 98.44% test accuracy, and the CNN-BiGRU-Attention model achieved a macro F1 score of 0.9052, a Brier Score of 0.0417, and an ECE of 0.1023. The study highlighted the importance of classification accuracy as an evaluation measure for reliable ECG systems, and recommended future work to include post-hoc calibration and explicit uncertainty modelling.

The studies show that reliable ECG diagnosis requires more than high classification accuracy. Uncertainty estimation, calibration and selective prediction can be used to give further information about the trustworthiness of the automated prediction. But current approaches typically do not include calibrated uncertainty, explicit cross-dataset domain-generalised representations, or multi-method explanation mechanisms.

Cross-Dataset Generalisation and Domain Adaptation

Cross-dataset generalisation is a major challenge because data collection methods, patient populations, ECG signal characteristics, annotation methods, and disease distributions differ significantly across datasets. Oporto et al. [9] presented an overview of end-to-end deep learning studies for ECG, revealing a variety of challenges related to preprocessing, datasets, diseases, performance metrics, and explainability. Their review especially noted issues of generalizability, class imbalance, multiple leads, comorbidities, and inadequate explainability and confidence evaluation.

The cross-dataset ECG domain generalisation and interpretability were directly examined by Li et al. [12]. They introduced a Domain-invariant Feature Enhancement Module and an Interpretability-driven Attention Constraint Mechanism in their Transformer-based framework. Experiments across MIT-BIH, INCART, and SCD-Holter showed that the proposed method achieved cross-domain accuracy improvement and F1-score gains. Ablation studies verified the domain and interpretability contributions of the proposed domain. The study shows that distribution alignment and cross-domain contrastive learning can alleviate dataset effects while retaining clinically useful information.

Kim and Shin [13] considered dataset-specific bias by performing unified multi-dataset ECG training. They employed supervised contrastive learning, a two-head design and layer-wise normalisation in their framework. Under the Leave-One-Dataset-Out protocol, the method achieved an average accuracy of 97.5% and an F1-score of 89.3%. The study showed that supervised contrastive learning could make ECG feature representations within the same rhythm class more consistent across data sources.

Vallabhuni et al. [8] also explored cross-dataset adaptation with AdaBN, and Li et al. [12] explored domain-invariant feature enhancement and attention constraints. Kim and Shin [13] showed that supervised contrastive learning is useful for leave-one-dataset-out generalisation. These studies highlight the critical need for dataset-invariant representation learning to enable reliable ECG deployment. However, they take different

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

approaches and do not integrate adversarial domain learning, class-aware contrastive alignment, and statistical feature alignment like CORAL, uncertainty calibration, and selective prediction.

ECG Classification, Representation Learning and Robustness

Several recent studies have focused on improving representation learning and robustness for ECG. To address noise and missing ECG data, Khan et al. [16] proposed CEDRC-Net, which consists of two stages: a multistage convolutional denoising autoencoder and a VAE or Temporal Fusion Transformer reconstruction. The authors then used the reconstructed signals to classify cardiac diseases. The TFT-based reconstruction achieved lower reconstruction errors than the VAE and better downstream classification performance. This paper emphasises the need for signal quality management prior to cardiac abnormality classification, especially for noised/abbreviated ECG signals.

Majid et al. [17] developed SimCardioNet. This hybrid self-supervised and supervised method fuses SimCLR pre-training, a multi-scale CNN backbone, residual connections, multi-head self-attention, and progressive layer unfreezing. The framework was assessed on several ECG image datasets, including PTB-XL, and showed excellent classification performance. In particular, self-supervised learning and contrastive representation learning are important because they highlight the potential of learning robust ECG representations before supervised classification.

Jiang et al. [21] proposed a novel model called ECG-Mamba to classify cardiac abnormalities based on 12-lead ECGs by employing non-uniform-mix augmentation. This study reflects the trend toward efficient sequence modelling and augmentation techniques for addressing the temporal and variability aspects of multi-lead ECG signals.

Aladhadh [1], Majid et al. [17], and Jiang et al. [21] show different approaches to improving ECG representation, robustness, and generalisation, respectively. However, these studies do not address cross-dataset domain invariance, calibrated uncertainty, selective prediction, and explanation reliability simultaneously.

Other Relevant Reviews and Future Trends

Preanto et al. [10] systematically reviewed 75 studies on the use of machine learning, deep learning, and XAI in predicting cardiovascular disease using ECG. The review noted generally good internal performance but highlighted that relatively few studies conducted external and/or prospective clinical validation. It also found that class imbalance, data quality variation, limited generalizability, interpretability, and lack of external validation are significant challenges for clinical deployment. Notably, the review highlighted that popular techniques like SHAP, LIME and Grad-CAM are still largely post-hoc and have limited physiological and clinical validation.

Bulbul et al. [18] also conducted a systematic review and meta-analysis of deep learning research for diagnosing cardiovascular diseases from ECG. The study demonstrates the good performance of CNN, recurrent, and Transformer-based models and emphasizes the importance of improved validation and clinically relevant evaluation.

Murali and Meenatchi [14] reviewed Explainable and Federated Multimodal AI for Heart Disease Detection with ECG, Cardiac Imaging, and EHR data. Their work highlighted the importance of multimodal fusion, privacy preservation, explainability, and comprehensive institutional implementation. The issues raised apply to future expansions of ECG systems into multimodal and privacy-focused clinical AI.

Islam and Alawad [15] explored the federated learning of resource-aware heterogeneous models. They also present results highlighting convergence and performance differences due to device heterogeneity and model size differences, showing that beyond statistical domain shift, real-world ECG deployment also faces computational and hardware heterogeneity.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

Research Gap from Existing Literature

The reviewed studies show significant advances in deep learning for ECG, but the literature remains scattered across reliability dimensions. Some explainable methods, such as [4], [6], [7], and [8], provide transparency but do not fully address predictive uncertainty and cross-dataset domain shifts. Studies on uncertainty (e.g., [2, 3, 11]) generally focus on limited datasets and enhance the confidence estimation and calibration, but do not explicitly learn domain-invariant representations. Although domain generalisation and contrastive representation learning are vital, as shown by cross-dataset studies like [8], [12] and [13], they fail to consider all of the above aspects, namely calibrated uncertainty, deep ensembles, selective abstention, and explanation reliability, comprehensively. External validation, generalizability, class imbalance, explainability, calibration, and clinical trust are other persistent challenges in ECG-based AI, as identified by reviews [9], [10] and [18].

This results in an important research gap regarding developing a unified multi-lead ECG system that would learn morphology and long-range rhythm representations, while also reducing the amount of information required for the dataset, aligning long-range rhythm representations across domains, quantifying and calibrating prediction uncertainty, and offering clinically interpretable explanations with an abstention mechanism for uncertain cases.

To address this gap, the proposed TrustECG-DG framework cascades a multi-scale CNN morphology encoder, temporal Transformer, bidirectional cross-attention fusion, domain-adversarial learning, class-aware contrastive alignment, CORAL alignment, five-model deep ensemble, temperature scaling, Integrated Gradients, attention rollout, and uncertainty-based selective prediction. These are explicitly integrated into an ECG decision-support pipeline across multiple datasets.

PROPOSED TRUSTECG-DG METHODOLOGY

Problem Formulation

The proposed TrustECG-DG framework aims to design a trustworthy deep-learning-based cardiac anomaly detection system from diverse 12-lead electrocardiogram (ECG) data while addressing domain variation, prediction uncertainty, calibration, and interpretability. Allow an ECG recording to be represented by

$X \in R^{12\times T}$,

This represents the standard 12 ECG leads and the number of temporal samples is represented by T represents. In the proposed implementation the standardized input representation is given by

$X \in R^{12\times 5000}$.

The goal is to gain knowledge of a function

$f_\theta : X \to \hat{Y}$,

The proposed deep-learning model parameterized by the model parameter θ is denoted as f_θ, and the predicted cardiac abnormality labels are denoted as $\hat{Y}$. The framework aims to retain clinically relevant ECG information while minimizing the effect of characteristics of the specific dataset. The architecture thus treats electrocardiograms from each clinical data set as a distinct domain; it strives to learn a representation that is informative in each of those domains. The proposed architecture specifically considers PTB-XL and Chapman–Shaoxing as heterogeneous ECG domains.

TrustECG-DG differs from a traditional ECG classifier, which only optimizes classification performance, by posing the learning problem as a joint one of diagnosis and domain-invariant representation learning. Assume that there are two domains, $D_s$and D_t, called source and external domain, respectively. ECG recordings and their diagnosis are contained in each domain. The aim of the model is to learn a representation $z = f_\theta(X)$ that

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

captures relevant information for cardiac abnormalities while reducing the dependence on the properties of the recording domain.

The classification goal is written as

$$L_{cls} = L(Y, \hat{Y}),$$

where Y denotes the target diagnostic labels and $\hat{Y}$ represents the predicted labels. But, when minimizing L_{cls} alone is done, the model may use some characteristics of the dataset to exploit such characteristics if they are correlated with the diagnostic labels. To tackle this issue, the proposed framework introduces domain-adversarial learning, class-aware contrastive alignment, and feature alignment based on CORAL. These mechanisms support the learned representation to retain diagnostic information while minimizing the difference in distributions between ECG domains. The full objective of domain-generalization is then part of the optimization procedure instead of an after-the-fact post-processing step.

The problem is taken further from accuracy classification to reliability prediction. The framework outputs a prediction for the abnormality of input X and an uncertainty or confidence measure for the prediction.

$$f(X) \rightarrow \left(\hat{Y}, U(X)\right),$$

where $U(X)$ represents the uncertainty in the ECG prediction. The proposed framework adopts a 5-model deep ensemble with temperature scaling to obtain a more reliable probability estimation. The architecture specification contains five different models of TrustECG-DG that use different random seeds, and the models are then aggregated to predict the uncertainty of prediction before they are calibrated.

The final problem formulation therefore consists of four interconnected objectives:

1. Cardiac abnormality detection: accurately identify the diagnostic category or categories associated with an ECG recording.

2. Domain-generalized representation learning: learn ECG features that remain useful when the data distribution changes across datasets.

3. Reliable prediction: estimate and calibrate model confidence so that uncertain predictions can be identified.

4. Interpretable decision support: provide evidence regarding the ECG regions contributing to the predicted abnormality.

Based on these objectives, the proposed framework can be expressed conceptually as

$$\theta \quad = \underset{\theta}{argmin}(L_{cls} + \lambda_1 L_{adv} + \lambda_2 L_{con} + \lambda_3 L_{CORAL}),$$

where L_{cls} is the cardiac abnormality classification loss, L_{adv} is the domain-adversarial loss, L_{con} is class-aware contrastive alignment, and L_{CORAL} is covariance-based domain alignment. The parameters, $\lambda_1, \lambda_2, \lambda_3$, control the weight of the corresponding domain-generalization objectives.

Once a final prediction has been made, the framework assesses the uncertainty of the prediction. If the uncertainty is in the specified acceptance range, the prediction can be given as an automatic result. High uncertainty indicates that the system is going to use a selective prediction strategy and recommend that the prediction be done by an expert rather than automatically. Architecture thus takes ECG classification as a problem of reliable decision making instead of the accuracy only classification problem.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

The full problem formulation is laid down as a foundation for the following steps in TrustECG-DG: representation of the ECG in a standard format, feature extraction using two branches, fusion of features based on the cross-attention mechanism, domain-generalized learning, ensemble prediction, calibration, explainability, and uncertainty-aware selective decision-making.

Dataset, Diagnostic Ontology and Unified Preprocessing

The proposed framework of TrustECG-DG aims to learn a shared representation from different kind of 12-lead ECG datasets. In the present study, PTB-XL and Chapman–Shaoxing are considered as the principal ECG domains. While both datasets include clinical 12-lead ECG recordings, the recording characteristics and the ECG diagnosis distribution may vary between datasets. So, directly training a model using one data set

and assuming that the same representation will be the best in another data set can cause domain specific bias. This is addressed in the proposed framework through the creation of a common diagnostic ontology that is created prior to the training of the model by mapping the available diagnostic information. The common diagnostic categories specified in the architecture specification are Normal, AF/Flutter, Premature Beats, BBB, Conduction Abnormality, and ST–T Abnormality.

The shared diagnostic ontology is a shared representation of the classification task. Suppose that the harmonized label vector for an ECG recording is written as

$$Y = [y_1, y_2, \dots, y_K],$$

where K denotes the number of diagnostic categories and y_k is the presence or absence of the k-th abnormality. This representation makes it possible for the model to learn the diagnostic information without the label organization being tied to the original data. For experiments in which the source and external datasets describe the prediction task using compatible diagnostic categories, the harmonized representation is especially important.

All the ECG recordings are converted into a common signal representation before feature extraction. The desired input is set to

$$X \in R^{12\times 5000},$$

First dimension: 12 ECG leads, second dimension: standardized temporal length. The preprocessing pipeline is designed as in TrustECG-DG and consists of standardized lead processing, resampling at 500 Hz, band pass filter (between 0.5 and 45 Hz), robust normalization, signal quality control, and augmentation with the morphology preserved.

To make sure the corresponding ECG channels are given to model in the same order in different datasets, the data is standardized for lead. This avoids the feature extractor learning some artificial differences due to different channel arrangements. Resampling gives a common temporal representation and band-pass filtering is applied to the signal to remove unwanted components in the frequency range not of interest. These operations are done before the ECG is given to the feature-learning architecture.

After filtering, strong normalization is used to diminish the effect of amplitude variations in the data and possible outliers. A well formulated normalization formulation is given by

$$X' = \frac{X - median(X)}{IQR(X) + \epsilon},$$

where $median(X)$ is the median amplitude of the signal and $IQR(X)$ is the interquartile range and ϵ is a small enough constant so that it is not divided by zero. This normalization is designed to preserve the relative morphology of the ECG waveform, while minimizing the sensitivity to amplitude-scale differences among recordings.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

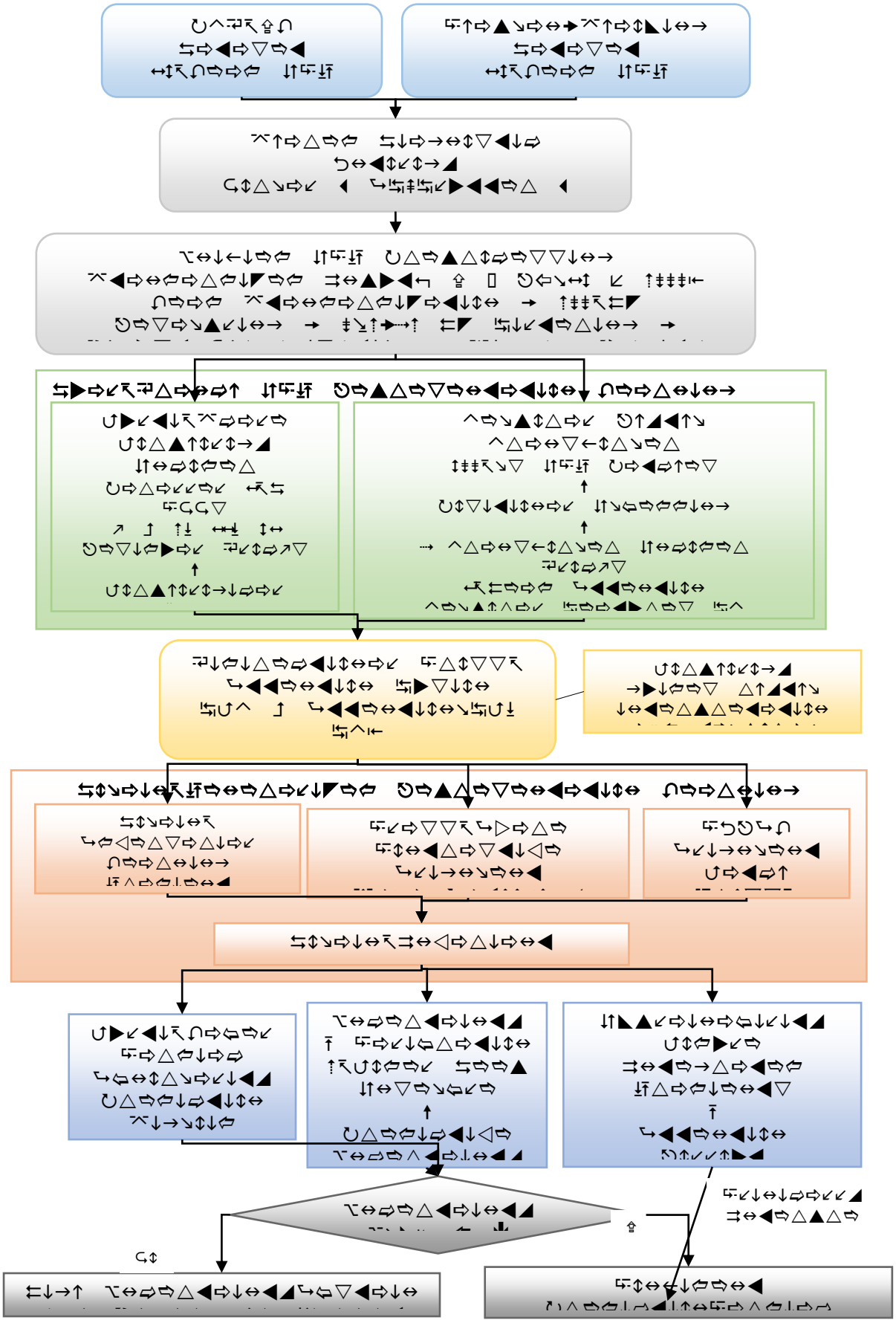


Figure 1. Overall Architecture of the Proposed TrustECG-DG Framework for Cross-Dataset Cardiac Abnormality Detection

A signal quality control stage then follows, to detect recordings which don't meet the desired quality criteria. This step is crucial as the patterns caused by severe noise or corrupted recordings might not be related to

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

cardiac abnormalities and could introduce errors in model learning. To this end, the preprocessing pipeline should focus on providing representations to the proposed model that reflect mostly useful ECG information.

Morphology preserving augmentation is not used to attempt to change the diagnostic features of the ECG signal, but is included to enhance robustness when training. The goal of augmentation is to get the model to see a limited amount of variation in the signal presented and to minimize an over-reliance on one specific way the signal was presented. The preprocessing principles are kept in all the ECG domains considered so as to have a uniform input representation.

This leads to a uniformly represented ECG that is then simultaneously transmitted to both complementary feature-learning arms of TrustECG-DG. The morphology branch is responsible for learning local features of the waveforms, and the temporal branch is responsible for learning over longer distances in the ECG sequence. This is a standardised input procedure for subsequent cross-attention and domain-generalised representation learning. The overall architecture is connected to the multi-scale CNN and temporal Transformer components as depicted in Figure 1, as follows.

Multi-Scale Morphology and Temporal Representation Learning

The suggested TrustECG-DG framework has two feature-learning mechanisms to learn two different characteristics existing in 12-lead ECG signals. The morphology encoder is a multi-scale morphology encoder using a parallel one-dimensional convolutional neural network, and the temporal rhythm Transformer is a temporal rhythm Transformer. For the two branches, they are trained to learn complementary representations prior to interacting in the next stage with the bidirectional cross-attention mechanism, described below.

The multi-scale morphology encoder aims to extract local and intermediate scale features of ECG waveforms. The ECG abnormalities may present as alterations in the morphology, duration, amplitude and temporal structure of the waves. A single convolutional receptive field might not be able to capture all these variations. Hence, the proposed architecture uses parallel one-dimensional convolutional branches of 5, 11 and 21 kernel sizes. Each of these different kernel sizes has multiple receptive fields, that can be used for extracting the ECG characteristics at various temporal scales.

The small-kernel branch concentrates on the local characteristics of the waveforms, whereas the larger kernels give increasingly wider temporal context. This enables the model to learn complementary representations from the same ECG signal, instead of a fixed-scale representation. Finally, the outputs of parallel convolutional branches are fed to residual blocks. Residual connections are used to pass information to deeper layers of feature extraction and to maintain valuable low-level features in the ECG, with successive layers building higher level morphological representations.

Assume that X represents the standardized ECG input. The multi-scale convolutional representation can be conceptually viewed as

$$F_M = CNN_{5,11,21}(X),$$

F_M represents the learned morphological feature representation and the subscripts refer to the three convolutional receptive fields. The resulting representation preserves the local morphology of the ECG as well as some general features of the ECG waveform that are relevant for the eventual classification of cardiac abnormalities.

The second branch aims to capture long-range temporal dependencies with a Transformer architecture, while the CNN branch is mainly responsible for capturing morphological information. The ECG signal is first divided into temporal patches of 200ms and then positional information is added to the resulting patch representations. The temporal representations are then passed into four Transformer encoder blocks consisting of eight attention heads.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

Self-attention helps the Transformer to create connections between the various parts of the ECG signal in time. This is especially useful in ECG analysis, since the diagnostically relevant information can not only be based on a single component of the waveform, but also on relationships between the components that are located in different temporal positions. This way the Transformer branch can learn a temporal representation F_T which complements the local morphological representation learned by the CNN branch.

The temporal representation may be represented as

$$F_T = Transformer(X).$$

The Transformer is thus able to provide information about temporal dependencies, irregularity in rhythm and longer-range relationships in the ECG sequence. It is interesting to note that $F_M$and F_T gives the proposed architecture two complementary representations of the same signal (ECG data): the CNN representation emphasizes the local appearance of the signal while the Transformer representation emphasizes the relationship of the ECG information across time.

These two representations are still kept as a separate feature stream and not concatenated right away. This design enables the following bidirectional cross-attention mechanism to be trained to capture interactions between the morphological and temporal information. The overall relationship of these two feature-learning branches and their following integration is presented in Figure 1: Overall Architecture of the Proposed TrustECG-DG Framework.

The dual representation paves the way for the second phase of the proposed framework, which involves co-refining morphological features with temporal features via bi-directional cross-attention before undergoing domain-generalized representation learning.

Cross-Attention and Domain-Generalized Feature Learning

After obtaining morphological and temporal representations, the proposed TrustECG-DG system combines the two feature streams by bidirectional cross-attention. The idea is to enable the inclusion of relevant temporal information in the morphological representation and relevant morphological information in the temporal representation. This interaction is salient, since the ECG abnormalities could be described not only by the morphology of the ECG wave forms but also by the relationships of different temporal areas of the ECG signal.

Let F_M denote the morphological representation obtained from the multi-scale CNN and F_T denote the temporal representation obtained from the Transformer. The relationship between the morphology and the temporal is called as ⇔

$$F_{MT} = Attention(F_M, F_T),$$

The morphological representation contains the information for the query and the temporal representation contains the complementary information for query context. On the other hand, the action of the time and the morphology is represented as

$$F_{TM} = Attention(F_T, F_M).$$

For this purpose, two representations are created and merged to get a unified ECG representation z. This is a bidirectional approach that enables the model to learn relationships between local characteristics of the waveforms and the higher-level temporal dependencies, rather than assuming that the two representations are independent feature sources. As shown in Figure 1, the overall process of the interaction between the morphology encoder, the temporal Transformer and the cross-attention mechanism is presented.

The unified representation is then optimized for the domain-generalized ECG learning. This step is required as different datasets may result in different patient populations, recording conditions, diagnostic distribution, and

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

other dataset-specific differences that can be seen in the ECG. The goal is thus to learn a representation that captures information relevant to the cardiac abnormalities, but discards information unrelated to the data set that the ECG originates from.

TrustECG-DG achieves this goal in three complementary ways: domain-adversarial learning, class-aware contrastive alignment, and feature alignment using CORAL. These components work synergically on the learned ECG representation and offer various types of domain regularization. The overall domain-generalization architecture is depicted in Figure 2, and the relationship between domain-adversarial learning, contrastive alignment, and covariance alignment is shown.

The first is the domain adversarial learning. A Gradient Reversal Layer is added between the shared ECG feature representation and a dataset-domain classifier. In forward propagation, the domain classifier tries to identify if a given ECG representation came from a specific dataset. The Gradient Reversal Layer (GRL) inverts the domain classification gradient given to feature encoder during back-propagation. This makes it easier for the feature encoder to learn features that are helpful for cardiac abnormality classification and harder to distinguish the dataset identity. The domain-adversarial loss is:

$$L_{adv} = -\sum_i d_i \, log(\hat{d}_i),$$

where d_i is the actual domain label and $\hat{d}_i$ is the predicted domain probability. The aim is not to discard the diagnostic information, but to make the learned representation less reliant on the domain specific properties.

The second part is class-aware contrastive alignment. Use of domain alignment could lead to representations of different diagnostic classes being made unnecessarily similar. The proposed framework alleviates such issues by involving the diagnostic class information in contrastive learning. Clusters of ECG representations from the same disease type are given a higher affinity in the feature space, while clusters of representations from different disease types are moved apart. A contrastive objective is represented as

$$L_{con} = -log \frac{exp((z_i, z_p)/\tau)}{\sum_j exp((z_i, z_j)/\tau)},$$

where z_i represents an anchor ECG representation, z_p represents a positive representation corresponding to the corresponding diagnostic category, and $(\cdot)$ is the similarity of the features. This mechanism makes the model to maintain diagnostic structure, yet minimizes domain-specific feature variation.

The third is CORAL alignment, wherein the statistics of the feature representations are aligned across domains. Let C_s and C_t be the covariance matrices of the source and target domain feature representations. The CORAL loss is given by

$$L_{CORAL} = \frac{1}{4d^2} \|C_s - C_t\|_F^2,$$

d denotes the dimension of the features, $\|\cdot\|_F$ is the Frobenius norm. CORAL also offers another way of minimizing distributional differences in the learned representation by minimizing the difference between source and target covariance structures.

There is a complementarity among the three domain-generalization mechanisms. In domain-adversarial learning, explicit information from the datasets is minimized; in contrastive learning with classes information from the classes is preserved; and in CORAL, the differences in the statistical structure of the learned features are reduced across the domains. They achieve a more domain-invariant clinical ECG representation together with preserving information that is essential for the prediction of cardiac abnormality. The whole domain-generalization mechanism is presented in Figure 2: Domain Generalization, Uncertainty, Calibration and Selective Decision Mechanism.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

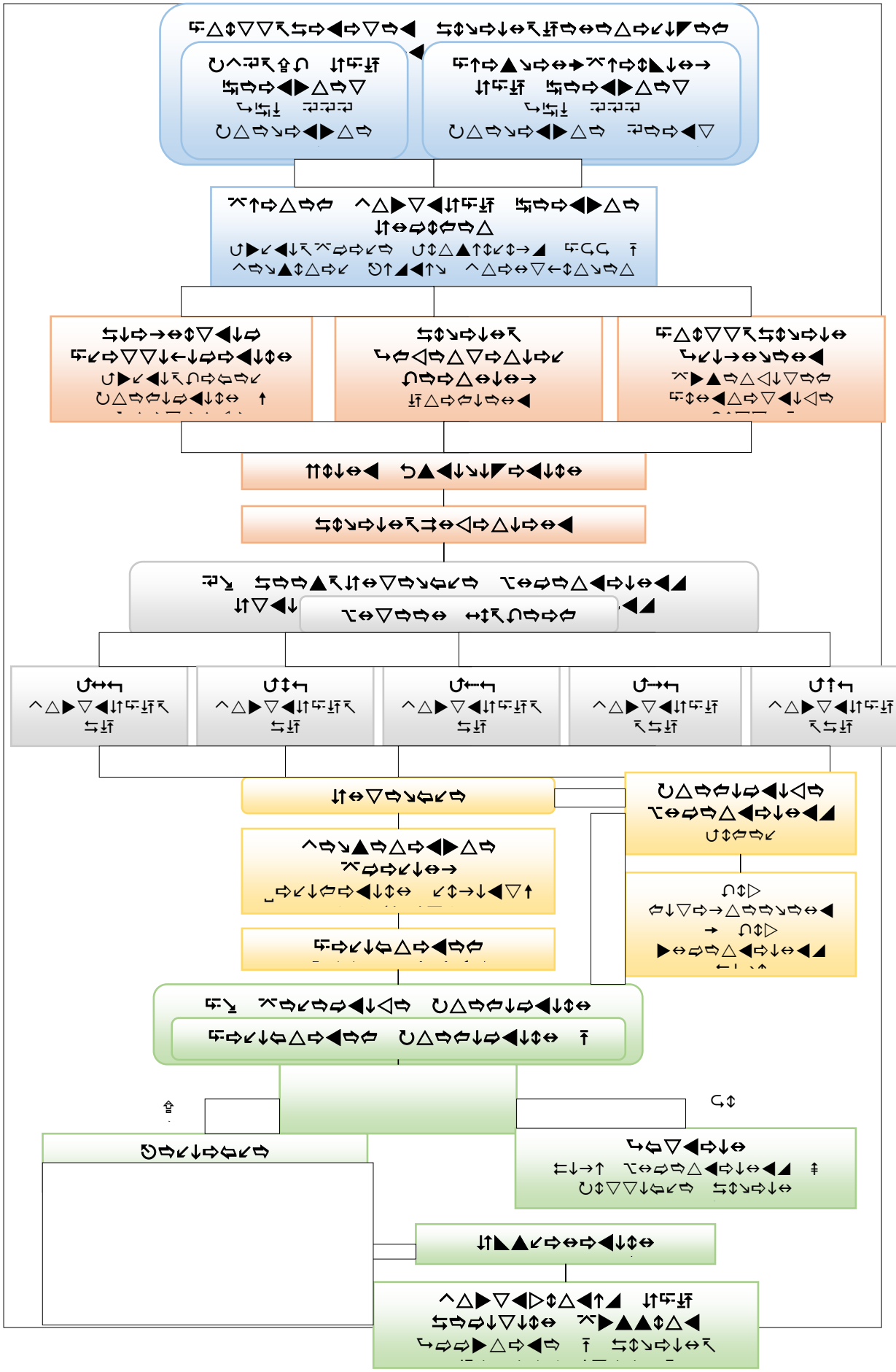


Figure 2. Domain-Generalized Representation Learning, Uncertainty Calibration, and Selective Decision Mechanism in the Proposed TrustECG-DG Framework

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

This representation is then passed on to the classification stage where the framework makes multi-label cardiac abnormality predictions. The overall classification objective and the three domain-generalization objectives are optimized together in the overall training objective described by

$$L_{total} = L_{cls} + \lambda_1 L_{adv} + \lambda_2 L_{con} + \lambda_3 L_{CORAL}.$$

In this work, L_{cls} represents the cardiac abnormality classification loss, and L_{adv}, L_{con}, and L_{CORAL} represent the three domain-generalization objectives. The parameters λ_1, λ_2, and λ_3 play a role in controlling the various contributions to the overall optimization.

The cross-attention stage therefore returns integrated morphological-temporal representations, whereas the subsequent domain-generalization stage converts the representations into features that are supposed to be informative in different domains of EGCS. This integrated representation is then used by the classification, ensemble, calibration, uncertainty and explainability parts of TrustECG-DG.

Multi-Label Classification, Ensemble Prediction and Calibration

The proposed TrustECG-DG framework firstly extracts the domain-general ECG features and then classifies cardiac abnormalities by adopting multi-label prediction scheme. The formulation makes it possible for the model to carry out diagnosis of the categories specified in the shared ontology and to estimate probabilities for each of these categories separately. The classification results for class kis are obtained with the sigmoid function for a given ECG representation z:

$$p_k = \sigma(z_k) = \frac{1}{1+e^{-z_k}},$$

where z_k denotes the classification logit for the k -th diagnostic category, and p_k represents its predicted probability. The formulation enables independent probabilities of the cardiac abnormality categories considered, instead of forcing all ECG recordings into one mutually exclusive category.

Multi-label decision rule: For each ECG recording, the ensemble produces a probability for each of the six diagnostic categories. The final prediction is obtained by applying a category-specific decision threshold to each calibrated probability independently. A category is predicted as present when its calibrated probability meets or exceeds its corresponding threshold; otherwise, it is predicted as absent. This procedure allows multiple cardiac abnormalities to be predicted for the same ECG recording. The decision thresholds are selected using validation data and are fixed before final test evaluation.

The optimization of the classification component is based on the binary cross-entropy objective:

$$L_{cls} = -\sum_{k=1}^{K}[y_k log(p_k) + (1-y_k) log(1-p_k)],$$

y_k is the true label for class k, p_k represents the predicted probability for class k, and K is the number of diagnostic classes. The domain-generalization goals outlined above are also included as part of the overall optimization objective of TrustECG-DG alongside the classification loss.

For making the final prediction robust and to get an estimation of the predictive uncertainty, the framework uses a deep ensemble of five models. Five separate instances of TrustECG-DG are trained with five different random seeds: 11, 23, 37, 51 and 73. Independent model initialisation enables the system to retain variations in predictions learned in the models, instead of using only a single trained model. The proposed architecture includes the five-model ensemble explicitly as a component of the uncertainty estimation mechanism.

Each member of the ensemble produces a probability vector p_m, given an input ECG X. The five models provide an ensemble prediction, which is averaged to get the final ensemble prediction.

$$\overline{p} = \frac{1}{M}\sum_{m=1}^{M} p_m,$$

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

where $M = 5$. The final $\bar{p}$ is the sum of the probabilities for each cardiac abnormality. The average of predictions will be less dependent on how a particular model performs and will give a more robust prediction for the next stage of uncertainty analysis.

Predictive uncertainty is also represented by the differences among the predictions made by the ensemble members. If ensemble members make similar predictions, the disagreement among the ensemble is reduced. On the other hand, large discrepancies in the predictions mean that the uncertainty is larger. Therefore, the ensemble disagreement is another clue to classify ECG recordings for which the automated prediction might be less reliable. The structure of the proposed architecture explicitly links ensemble prediction and predictive uncertainty prior to the calibration and selective decision phase.

The ensemble prediction then is throughly subjected to temperature scaling to enhance the probability calibration. Neural networks can generate outputs that are not necessarily related to their confidence in the prediction. Thus, the proposed framework does not make a direct inference of raw ensemble probability to be interpreted as calibrated confidence. Rather, a temperature parameter T is estimated from the validation data and used in model logits:

$$p_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}.$$

In this case, z_i represents the logit of class i and T is the temperature parameter learned. This is to adjust the sharpness of the predicted probability distribution, without having to retrain the entire classification model.

The right temperature enables predictions with high confidence to be scaled appropriately if the raw model is overconfident without making the distribution of the predictions too narrow. The calibration is done using validation data so that test data are not included in the calibration. The calibrated probabilities are then fed to the uncertainty-aware selective prediction mechanism.

The links between ensemble, predictive uncertainty and calibration are illustrated in Fig. 2. This figure represents the prediction of each of the various models in the TrustECG-DG ensemble, the aggregation of these individual predictions to form the ensemble output, the use of model disagreement to estimate uncertainty, and the scaling of these uncertainties by the temperature.

As a result, the calibrated prediction is then a combination of two kinds of information: the probability estimation for each cardiac abnormality and the indication of the reliability of the prediction. The distinction is crucial for the proposed framework since in the case of a high classification probability, a high level of disagreement among the ensemble members should not be taken as a guarantee for the reliability of the classification decision itself. The calibrated prediction and uncertainty estimate are then sent to a selection prediction stage where the system decides whether an automated result can be given to the ECG or if the ECG needs to be passed to an expert for analysis.

Explainability and Explanation Reliability

To identify these areas that play the most decisive role in the predicted cardiac abnormality, explainability is integrated into TrustECG-DG. This framework incorporates both Integrated Gradients (IG) and Transformer attention rollout, allowing for explanations to be derived from the learned signal representation as well as the temporal attention mechanism. This gives additional information on the input segments that are affecting the output prediction.

Integrated Gradients

Integrated Gradients is used to measure the contribution of each ECG input feature, compared to a baseline signal. The attribution of the feature i is the value of the model output $F(x)$, where the baseline x' is the value of the input x.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

$$IG_i(x) = \left(x_i - x_i'\right) \int_0^1 \frac{\partial F\left(x' + \alpha(x - x')\right)}{\partial x_i} d\alpha.$$

In this case, the reference ECG input is x' and the baseline is shifted towards the actual ECG signal by a parameter α. The resulting attribution values show the contribution of various signal regions for the prediction. Positive attribution means evidence that supports the selected prediction; and negative attribution means evidence that works against the selected prediction.

The attribution maps that are generated for the 12-lead ECG signal can be analyzed in spatial dimension (i.e., across the 12-leads) as well as across temporal dimension. The explanation therefore may suggest which ECG (electrocardiogram) leads and/or which time period of the ECG is most strongly associated with the prediction.

Transformer Attention Rollout

An attention mechanism is provided by the Transformer component as another source of interpretability. Attention weights of multiple heads and multiple layers of Transformers are combined to obtain attention rollout, which helps to estimate which temporal patches are more important when learning representations.

Attention rollout is a temporal relevance map over the input sequence since the ECG is split into temporal patches. These attention patterns together with the Integrated Gradients (IG) are then fused to create the final fused explanation map.

The representation of the two can be written as

$$E_{fused} = \lambda E_{IG} + (1 - \lambda) E_{AR},$$

where E_{IG} denotes the Integrated Gradients attribution, E_{AR} denotes the attention-rollout representation and λ controls the contribution of the two explanation sources.

Explanation Reliability

Explanation is assessed separately from classification as a visually plausible explanation is not necessarily the sign of the use of relevant ECG evidence. The reliability analysis then checks whether there is a consistent relation between the important regions found with the explanation methods and the model prediction.

The framework takes into account the faithfulness, stability and consistency of explanations in the evaluation process. To measure the changes to the model output that should be expected as changes are made to highly attributed regions, a test of faithfulness is undertaken. Stability measures the consistency of the explanations given by similar ECG inputs. These analyses aid in sorting out explanations that reflect the learned decision process from explanations that can occur from unstable attribution processes.

The explanation results are displayed along with the prediction and uncertainty information. Therefore, the system does not handle explainability as an independent visualization unit, but rather it connects the region of the ECG that contributes to the prediction of abnormality with the predicted abnormality. The explanation and reliability parts are included in the general architecture of the TrustECG-DG (Fig. 1) and in the decision-support mechanism (Fig. 2).

It is necessary to note that the maps that were created are for evidence relevant to the model and not clinical causality. Therefore, they are not meant to replace clinical diagnosis, but rather to assist in interpretation and expert review.

The explanation reliability assessment considers faithfulness, stability, and sanity/consistency to examine complementary aspects of explanation quality. Faithfulness concerns the relationship between highlighted ECG regions and model predictions, stability concerns the consistency of explanations under similar inputs,

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

and sanity/consistency examines whether explanations behave consistently with the model's learned representations. The numerical results should be interpreted in conjunction with the procedures used to calculate these measures.

3.7 Uncertainty-Aware Selective Prediction

Aside from accurately predicting, reliable ECG classification should also reveal the confidence of the model prediction. The model can give high classification probability even if there is noise, ambiguity, or a change in the input ECG with respect to the training set. To overcome this problem, TrustECG-DG includes ensemble-based uncertainty estimation, probability calibration, and selective prediction. The goal is to enable the system to differentiate between predictions that can be made with reasonably high levels of confidence and predictions that need additional examination by experts. The uncertainty and selective-decision elements are woven into the structure of the framework displayed in Fig. 2.

Ensemble-Based Predictive Uncertainty

TrustECG-DG is based on five independently trained models with random different seeds. Each model returns a vector of probabilities for an input ECG x.

$$p^{(m)}(x) = \left[p_1^{(m)}, p_2^{(m)}, .., .. p_K^{(m)}\right],$$

where m is the ensemble member and K the number of diagnostic classes.

The overall probability of the ensemble is P(final) =

$$\overline{p}(x) = \frac{1}{M}\sum_{m=1}^{M} p^{(m)}\,(x),$$

where $M = 5$.

The disagreement of ensemble members gives information on predictive uncertainty, the mean prediction gives the final class probability. The uncertainty is relatively low if the ensemble members are making similar predictions. When there is more disagreement, this could be because the input is hard to classify or contains properties that are not consistently captured by the learned model.

This is beneficial because the uncertainty of a single softmax or sigmoid is not always reliable. Ensemble disagreement offers another way of quantifying the robustness of the prediction obtained by different, independently-trained models.

TrustECG-DG uses a five-model ensemble to assess predictive uncertainty through disagreement among model predictions. The ensemble predictions provide information about the consistency of the model outputs. However, the numerical uncertainty measure must be interpreted according to the specific calculation used in the evaluation. The definition and aggregation procedure should therefore be specified to ensure reproducibility.

Probability Calibration

In order to minimize the gap between the model's confidence and its accuracy, probability calibration is applied. TrustECG-DG uses temperature scaling on the validation data.

The calibrated probability is given by

$$p_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)},$$

In which case T is the acquired temperature parameter.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

If the temperature is above one, the probability distribution will be 'softer' overall, and if the temperature is below one, the distribution will be 'sharper'. The parameter of temperature is estimated from the validation data and then kept constant for the evaluation performed on the test data.

Calibration matters because it is important that a prediction with a probability of 0.95 should have a high empirical chance of being right. Thus, calibration is used in addition to the ensemble uncertainty estimate and to enhance the quality of estimates of prediction probabilities.

Selective Prediction

It also incorporates a "selective" prediction method, meaning that the model is not required to make an unconditional automatic prediction for all ECGs. Let the uncertainty estimated by $U(x)$ be the uncertainty denotation and let the threshold of the uncertainty be the predetermined one denoted by δ.

The decision rule is:

$$Decision(x) = \begin{cases} \hat{y}, & U(x) \leq \delta, \\ Abstain, & U(x) > \delta. \end{cases}$$

If the estimated uncertainty is below threshold, the prediction will be regarded as a low uncertainty automated prediction. If there is uncertainty, the system doesn't decide and marks the ECG for further expert review.

This mechanism is especially interesting when dealing with ECGs which are not homogenous, as noise, atypical morphology, rare abnormalities, or domain characteristics can exist, which are different from the training distribution. Selective prediction offers a way for a model to identify cases that need special consideration rather than have it classify every hard sample.

Risk-Coverage Analysis

The relationship between coverage and selective risk is used to evaluate the effectiveness of the selective prediction. Selective risk is the error rate of the accepted prediction, while coverage is the fraction of ECG samples for which the model gives an automated prediction.

Conceptually,

$$Coverage = \frac{N_{accepted}}{N_{total}},$$

and

$$Selective\ Risk = \frac{N_{incorrect,accepted}}{N_{accepted}}.$$

A good selective prediction method should make the error of the accepted samples more and more uncertain as they are eliminated from the set of possible alternatives as more and more cases are excluded. Thus, the risk-coverage analysis is employed to test if the range estimates offer useful information on the reliability of prediction.

The uncertainty threshold can then be adjusted as required at the desired operating point. A more stringent threshold leads to more predictions being rejected, but may lead to more conservative decision process; a less stringent threshold will allow more ECGs to be automatically processed.

Integration with Explainability

Uncertainty information is provided along with the model prediction and explanation map. As such, there are

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

three complementary outputs from TrustECG-DG: the predicted cardiac abnormality, the level of confidence/uncertainty in the prediction, and the regions of the ECG used in the prediction.

These components make up the reliability-oriented decision layer of the proposed framework, which is shown in Fig. 2. The system does not therefore assign to itself a high probability prediction without considering the context of that forecast. Rather, prediction confidence, ensemble uncertainty, calibration, and explanation information are all taken into account when deciding the automatic or expert acceptance of an ECG.

The methodology section is completed in the next subsection 3.8 Overall Training Procedure and Computational Workflow where all the modules are trained and finally combined into the model TrustECG-DG.

Overall Training Procedure and Computational Workflow

The whole TrustECG-DG framework combines both the preprocessing and multi-scale representation learning, cross-attention fusion, domain generalization, classification, calibration, uncertainty estimation and explainability into a single training and inference framework. The training procedure ensures that the feature-learning components are trained together, instead of the domain adaptation phase, the classification phase and the representation learning phase being treated as three separate stages. The over-all architecture of the system is shown in Fig. 1, and the uncertainty, calibration, explanation, and selective-decision parts are shown in Fig. 2.

Model Training

The training process starts with harmonisation of the ECG recordings from the datasets available. These recordings are each rendered in the standardized, 12-lead representation and specified by the precise timing. The pre-processing stage resamples, filters, and normalizes the ECG to a high degree of robustness, checks signal quality, and augments the ECG in a way that preserves its morphology before it is fed into the feature-learning network.

The processed ECG is then passed simultaneously through the multi-scale CNN and temporal Transformer branches. The CNN branch is trained with local morphological properties with various convolution receptive fields, while the Transformer branch is trained to capture the longer temporal dependency. Then, their representations are exchanged through bidirectional cross-attention to get a joint morphology-temporal representation.

The fusion of the representation is optimized by combining the classification objective and the domain-generalization objectives. The overall training loss is a measure of

$$L_{total} = L_{cls} + \lambda_1 L_{adv} + \lambda_2 L_{con} + \lambda_3 L_{CORAL},$$

where L_{cls} represents the classification loss, L_{adv} represents domain-adversarial learning, L_{con} represents class-aware contrastive alignment and L_{CORAL} represents covariance-based domain alignment.

The Gradient Reversal Layer allows the feature extractor to learn non-domain-specific representations. Concurrently, by emphasizing contrastive alignment, samples in similar diagnostic classes are kept closer in the learned representation space and by reducing statistical differences between domains, CORAL makes it easier to learn the correct mapping.

Ensemble Construction

Once the joint representation-learning architecture is set up, the deep ensemble is built from five model instances trained independently. The random seeds for the models are 11, 23, 37, 51 and 73 as proposed in the proposed architecture.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

Each model makes a prediction vector $p^{(m)}(x)$ for an input ECG x. The last prediction is the average of the predictions:

$\overline{p}(x) = \frac{1}{5}\sum_{m=1}^{5} p^{(m)}(x).$

The ensemble prediction gives the final classification result and the spread of the individual model predictions is used as an estimate of predictive uncertainty.

The processing of calibration and reliability is performed in order to 3.8.3.

These ensemble predictions are then calibrated with the validation data. In the case of temperature scaling, the confidence distribution is adjusted by changing the temperature, but without retraining the entire feature extractor. These calibrated probabilities, along with the ensemble uncertainty estimate, are then used.

By separating classification, calibration, and uncertainty estimation, the framework can provide for three different aspects of model behavior: the type of abnormality the model predicts, the level of confidence in the calibrated probability, and the degree of agreement among the ensemble members in predicting that level of confidence.

Inference and Explanation

An unseen ECG is subjected to the same preprocessing as the ECG during training during inference. The processed signal is then fed into trained TrustECG-DG models, resulting in the final diagnostic probabilities from the ensemble of models.

The framework can also produce Integrated Gradients and Transformer attention-rollout explanations for every prediction. These explanations highlight the regions of the ECG that are relevant for the output of the model, and are included in conjunction with the prediction and uncertainty information.

Finally, the selective prediction mechanism decides if the prediction meets the predetermined uncertainty criterion. Low confidence predictions can be stored as automated prediction products, while high uncertainty predictions are flagged for further expert evaluation.

Algorithmic Representation

Algorithm 1: TrustECG-DG Model Training

Input: Multi-domain 12-lead ECG dataset $D = \{(X_i, Y_i, d_i)\}_{i=1}^{N}$

Output: Trained TrustECG-DG ensemble

1. **Preprocess ECG signals**

Convert each ECG into the standardized representation

$$X_i \in R^{12\times 5000}$$

and apply filtering, normalization, quality checking, and augmentation.

2. **Extract morphological features**

Apply the multi-scale CNN:

$$F_M = CNN_{5,11,21}(X_i)$$

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

where the different kernel sizes capture ECG patterns at different temporal scales.

3. **Extract temporal features**

Divide the ECG into temporal patches and process them using the Transformer:

$$F_T = Transformer(X_i)$$

4. **Fuse both representations**

Perform bidirectional cross-attention:

$$F_{MT} = Attention(F_M, F_T) F_{TM} = Attention(F_T, F_M)$$

and obtain the fused representation z_i.

5. **Learn domain-generalized features**

Apply domain-adversarial learning, contrastive alignment, and CORAL to reduce domain differences.

6. **Calculate the total training loss**

$$L_{total} = L_{cls} + \lambda_1 L_{adv} + \lambda_2 L_{con} + \lambda_3 L_{CORAL}$$

7. **Update model parameters**

Minimize L_{total} using the training data.

8. **Repeat training**

Train five independent models using different random seeds:

$$S = \{11,23,37,51,73\}$$

9. **Store the five trained models**

The resulting models form the TrustECG-DG ensemble.

Algorithm 1 describes the training stage of TrustECG-DG. It preprocesses the multi-domain ECG data, extracts morphological and temporal features using CNN and Transformer branches, fuses them through cross-attention, and learns domain-generalized representations using adversarial, contrastive, and CORAL losses. Five independently trained models are then used to construct the final ensemble.

Algorithm 2: TrustECG-DG Prediction, Uncertainty and Explanation

Input: New 12-lead ECG X, trained ensemble $\{f_1, \dots, f_5\}$

Output: Prediction $\hat{Y}$, calibrated confidence, uncertainty U, and explanation E

1. **Preprocess the input ECG**

Apply the same preprocessing procedure used during training.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

2. **Generate predictions**

Each trained model produces a prediction:

$$p_m = f_m(X), m = 1, \dots, 5$$

3. **Calculate the ensemble prediction**

$$\overline{p} = \frac{1}{5}\sum_{m=1}^{5} p_m$$

Apply class-specific thresholds to the calibrated ensemble probabilities to predict one or more abnormality categories..

4. **Calibrate the prediction**

Apply temperature scaling:

$$p_i^{cal} = \frac{exp(z_i/T)}{\sum_j exp\left(z_j/T\right)}$$

where T is learned using the validation data.

5. **Estimate uncertainty**

Measure the disagreement between the five model predictions:

$$U(X) = Var(p_1, p_2, \dots, p_5)$$

Higher disagreement indicates greater predictive uncertainty.

6. **Generate explanation**

Generate Integrated Gradients and Transformer attention-rollout maps:

$$E = E_{IG} + E_{AR}$$

These maps identify the ECG regions that contribute to the model prediction.

7. **Apply the uncertainty threshold**

$$Decision(X) = \begin{cases} \hat{Y}, & U(X) \le \delta \\ \text{Expert Review}, & U(X) > \delta \end{cases}$$

8. **Return the final result**

Report:

$$\{\hat{Y},, p^{cal}, U(X)E\}$$

where $\hat{Y}$is the predicted abnormality, $p^{cal}$is calibrated confidence, $U(X)$is predictive uncertainty, and $E$is the explanation.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

Algorithm 2 describes the prediction and reliability stage. The five trained models generate predictions for a new ECG, which are combined to obtain the final output and calibrated confidence. Ensemble disagreement is used to estimate uncertainty, while Integrated Gradients and attention rollout provide explanations; uncertain cases are referred for expert review.

In this section the proposed TrustECG-DG framework for reliable cardiac abnormality detection from heterogeneous 12-lead ECG datasets was presented. The methodology involves a unified ECG preprocessing, multi-scale CNN-based morphological learning, Transformer-based temporal representation, and bidirectional cross-attention, which are used to acquire the comprehensive ECG features. To enhance the robustness of learnt representations across ECG domains, Domain-adversarial learning, class-aware contrastive alignment and CORAL are introduced.

The framework also incorporates ensemble prediction and temperature scaling to enhance the reliability of the predictions, and Integrated Gradients and Transformer attention rollout methods are used to help explain the reasons behind the model's choices. Lastly, uncertainty-aware selective prediction allows for uncertain cases to be identified for expert review in order to avoid an automated decision. The two algorithms proposed are summary of the training and inferring procedure of the entire framework. The components are the methodological bedrock for the evaluation of the classification performance, domain generalization, calibration, uncertainty, and explainability of TrustECG-DG in the following experimental section.

Experimental Setup

This section explains the factors for the datasets, the processing method, the configuration of the implementation, the models used as baselines, the evaluation metrics, cross-dataset validation policy, and ablation experiments to test the proposed TrustECG-DG framework. The experimental design corresponds to the major goals of the proposed method: accurate cardiac abnormality classification, robustness to different ECG domains, reliable estimation of confidence, uncertainty-aware decision making, and interpretable prediction. The architecture is based on the principal domains of ECG: PTB-XL and Chapman – Shaoxing and translates their recordings into the common representation of 12-lead ECG.

Datasets and Data Sources

Both the heterogeneous 12-lead ECG dataset PTB-XL [22] and Chapman–Shaoxing [23] are regarded as the main data domains for the evaluation of TrustECG-DG. The clinical ECG recordings from PTTB-XL and Chapman–Shaoxing are independent clinical ECG domains. The proposed framework can be assessed in scenarios where multiple datasets are used, not just one dataset. Both datasets are defined in the architecture as 12-lead ECG sources, both sampled at 500 Hz.

To enable joint learning, the diagnostic labels from the two datasets are mapped to a shared diagnostic ontology. The common categories used by the framework are:

1. Normal
2. AF/Flutter
3. Premature Beats
4. Bundle Branch Block (BBB)
5. Conduction Abnormality
6. ST–T Abnormality

The harmonized diagnostic ontology maps the original diagnostic labels from PTB-XL and Chapman–Shaoxing into six shared categories: Normal, AF/Flutter, Premature Beats, Bundle Branch Block (BBB),

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

Conduction Abnormality, and ST–T Abnormality. Each original diagnostic label is assigned to the corresponding category according to its clinical meaning and predefined mapping rules. When a recording contains multiple applicable diagnoses, the corresponding category labels are retained to support multi-label representation. Ambiguous or unsupported labels should be handled using consistent, documented mapping rules.

Through this label harmonization, the model can learn the common diagnostic representations while maintaining the domain identity for domain-generalization learning. The unified representation will be especially significant, as the same cardiac problem may have distributional differences in ECG databases.

The heart's electrical activity is plotted as

$X \in R^{12\times5000}$,

In the above scheme, 12 is the number of ECG leads and 5000 is the number of temporal samples for standardization. The corresponding diagnostic vector with multiple labels is shown as

$Y = [y_1, y_2, \dots, y_K]$,

y_k is an element of $\{0,1\}$ which represents the presence or absence of the k -th diagnostic category.

Multiple diagnostic labels enable the framework to represent ECGs that have more than one abnormality without having to place each ECG in a mutually exclusive diagnostic class.

The PTB-XL and Chapman–Shaoxing datasets were selected to evaluate the proposed TrustECG-DG framework across heterogeneous 12-lead ECG domains. Before model development, the original recordings were examined for data integrity, lead availability, signal compatibility, and diagnostic-label availability. The number of recordings and unique subjects retained from each dataset after preprocessing was determined from the dataset records. The final dataset composition was documented separately for each database to support transparent evaluation and reproducibility.

Data Preprocessing and Experimental Protocol

All ECG records are preprocessed with the same pre-processing algorithm before being given to the proposed network. The goal is to minimize artifacts due to signal representation and acquisition with maintaining the ECG morphology diagnostically significant.

The first is that ECG leads are standardized; and the second is that recordings are represented in the common 12 lead format. The signals are then resampled to 500 Hz and scaled to the desired input length of 5000 samples. To remove baseline drift and high-frequency interference while preserving the main ECG waveform information a band-pass filtering range of 0.5–45 Hz is employed. Explicitly defined in the proposed TrustECG-DG architecture are these pre-processing operations.

Robust normalization is then applied to normalize the amplitude variations and the effect of outliers. The normalized signal for an ECG signal □ can be expressed as:

$X' = \frac{X - median(X)}{IQR(X) + \epsilon}$,

The median of the signal value is denoted as $median(X)$, the interquartile range is denoted as $IQR(X)$ and ϵ is a small constant that ensures numerical stability. The recordings or segments with poor signal quality are then identified by signal quality control. In the training process, a morphology-preserving augmentation is employed to augment the input data, but the purpose of the augmentation is not to remove any important ECG patterns. The same pre-processing method is used for training, validation and testing to prevent any difference between the stages arising from pre-processing. Important, the parameters learned from the training data are not estimated from the test set.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

The experimental protocol has separate sections for model development and final evaluation. The training data are used to optimize the model, the validation data are used to select the model and to calibrate the temperature scale, and the held-out test data are used to evaluate the final performance. This separation is especially critical to the calibration process, as the temperature parameter needs to be estimated prior to the final test observations.

The preprocessed ECG recordings were organized into training, validation, and testing subsets for model development and performance evaluation. The training subset was used to optimize model parameters, whereas the validation subset was used for model selection and temperature calibration. The testing subset was reserved for final performance assessment. Data partitioning should be performed at the patient level to ensure that recordings from the same patient do not occur across different subsets. Signal-quality control was applied to identify corrupted recordings, missing required leads, and signals failing the predefined quality criteria. The numbers of recordings excluded during quality control and retained for subsequent analysis should be documented for both datasets. The exact partition proportions and patient-level separation procedure must correspond to the implemented experimental protocol.

The cross-dataset experiments were designed to evaluate the model's ability to generalize across heterogeneous ECG datasets. PTB-XL and Chapman–Shaoxing were considered the source and target domains in turn. During model optimization, target-domain ECG features were used in the domain-alignment process to reduce dataset-specific distribution differences through domain-adversarial learning and CORAL alignment. The source-domain labels were used to guide disease classification, while the evaluation of cross-dataset performance was conducted using the held-out test partition of the target dataset. The same experimental procedure was applied in both transfer directions: PTB-XL to Chapman–Shaoxing and Chapman–Shaoxing to PTB-XL. The target-domain data used for alignment and the held-out target test data were kept separate to avoid test-data leakage.

Implementation and Training Configuration

The TrustECG-DG architecture is composed of multi-scale CNN branch, temporal Transformer branch, bidirectional cross-attention, domain-generalization modules, multi-label classification, ensemble prediction, calibration and explainability modules.

CNN branch: They apply parallel 1D CNNs with kernel sizes of 5, 11 and 21, followed by a residual block. In this way, the network can capture the ECG morphology at a variety of temporal scales, which is helpful for multi-scale ECG analysis.This helps the network to detect the ECG morphology at multiple temporal scales, which is beneficial for multi-scale ECG analysis. The Transformer branch is composed of 200-ms temporal patches drawn from the ECG, followed by four 200-ms ECG temporal blocks with a Transformer encoder, and eight heads of attention.

Bidirectional cross-attention is used to combine the two representations:

$F_{MT} = Attention(F_M, F_T)$,

$F_{TM} = Attention(F_T, F_M)$.

The created representation is utilized for classification and domain-generalization learning.

The domain-generalization component is comprised of the Gradient Reversal Layer (GRL) along with class-aware contrastive alignment and CORAL alignment, which is a domain classifier. Such components are jointly optimized to the diagnostic classification objective:

$L_{total} = L_{cls} + \lambda_1 L_{adv} + \lambda_2 L_{con} + \lambda_3 L_{CORAL}$.

At the ensemble stage, five independently trained models of TrustECG-DG with random seeds 11, 23, 37, 51, and 73 are used. The outputs are fed into each other to get ensemble prediction and estimate predictive uncertainty.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

The validation predictions are then scaled to temperature to enhance probability calibration. The explainability module fuses Integrated Gradients and Transformer attention rollout to produce the fused attribution map.

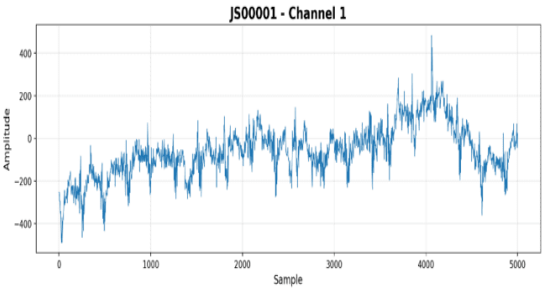
The TrustECG-DG framework employs a multi-scale one-dimensional CNN, a temporal Transformer, bidirectional cross-attention, and domain-generalization modules for cardiac abnormality classification. Five independently trained models, initialized using random seeds 11, 23, 37, 51, and 73, are combined to obtain ensemble predictions and estimate predictive uncertainty. Temperature scaling is performed using validation data, while the held-out test data are reserved for final evaluation. Integrated Gradients and Transformer attention rollout are employed to generate explanations of the model predictions.

For reproducibility, the implementation configuration includes the optimizer, initial learning rate, batch size, number of training epochs, learning-rate scheduling strategy, model-selection criterion, hardware specifications, programming language, deep-learning framework, and software-library versions. These parameters must be reported according to the configuration used in the actual experiments. The training, validation, calibration, and testing procedures should remain consistent with the stated evaluation protocol.

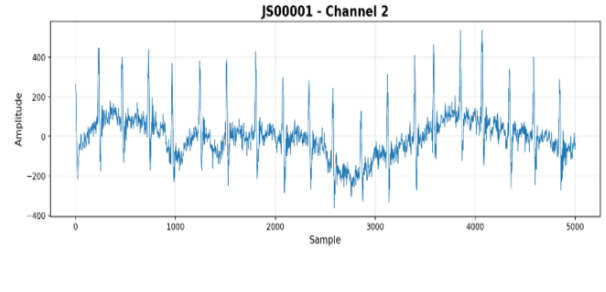
EXPERIMENTAL RESULTS

This section presents experimental results from the proposed TrustECG-DG framework for detecting cardiac abnormalities using the heterogeneous 12-lead ECG dataset. The analysis begins with exploratory visualisations of the ECG recordings to examine ECG signal characteristics, inter-lead variation, amplitude variations, baseline fluctuations, and ECG morphology within the datasets. Representative ECG visualisations from the Chapman–Shaoxing and PTB-XL datasets show the diversity of cardiac signals across multiple leads and form the basis for the pre-processing and feature-learning phases.

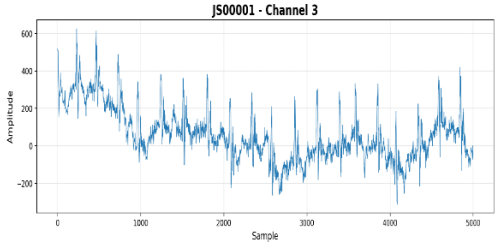
Next, we assess TrustECG-DG's classification performance using accuracy, precision, recall, specificity, F1 score, AUROC, and AUPRC. Additional analyses include class-wise performance, ablation performance, cross-dataset generalisation, calibration, uncertainty estimation, explainability, and residual errors. These experiments evaluate the framework's prediction performance and test its reliability and robustness for the different ECG domains.



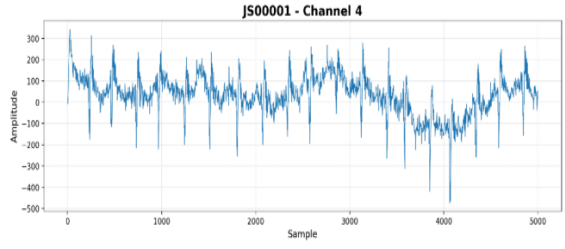
Channel 1



Channel 2



Channel 3



Channel 4

Figure 3: Raw ECG Signal Visualisation from the Chapman–Shaoxing 12-Lead ECG Database

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

The figure 3 shows representative raw ECG signals from four channels of the Chapman–Shaoxing 12-Lead ECG Database, illustrating variations in amplitude, baseline fluctuations, and signal morphology across channels. These waveforms demonstrate the natural variability and noise characteristics of multi-channel ECG recordings used for subsequent preprocessing and classification.

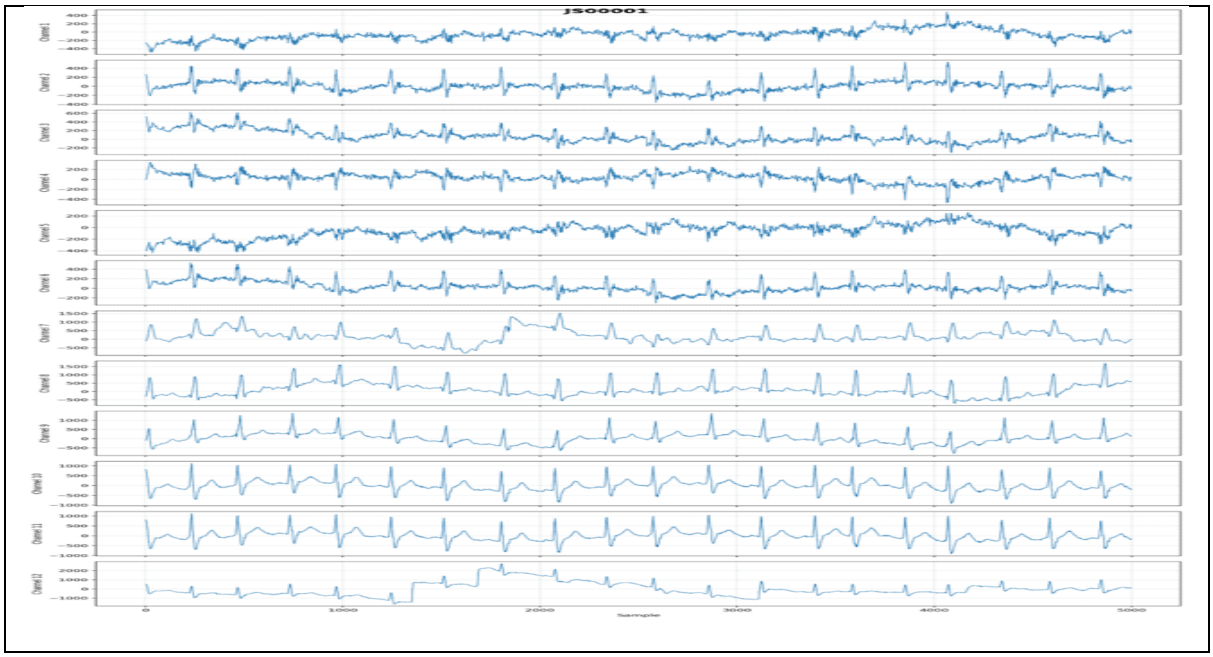
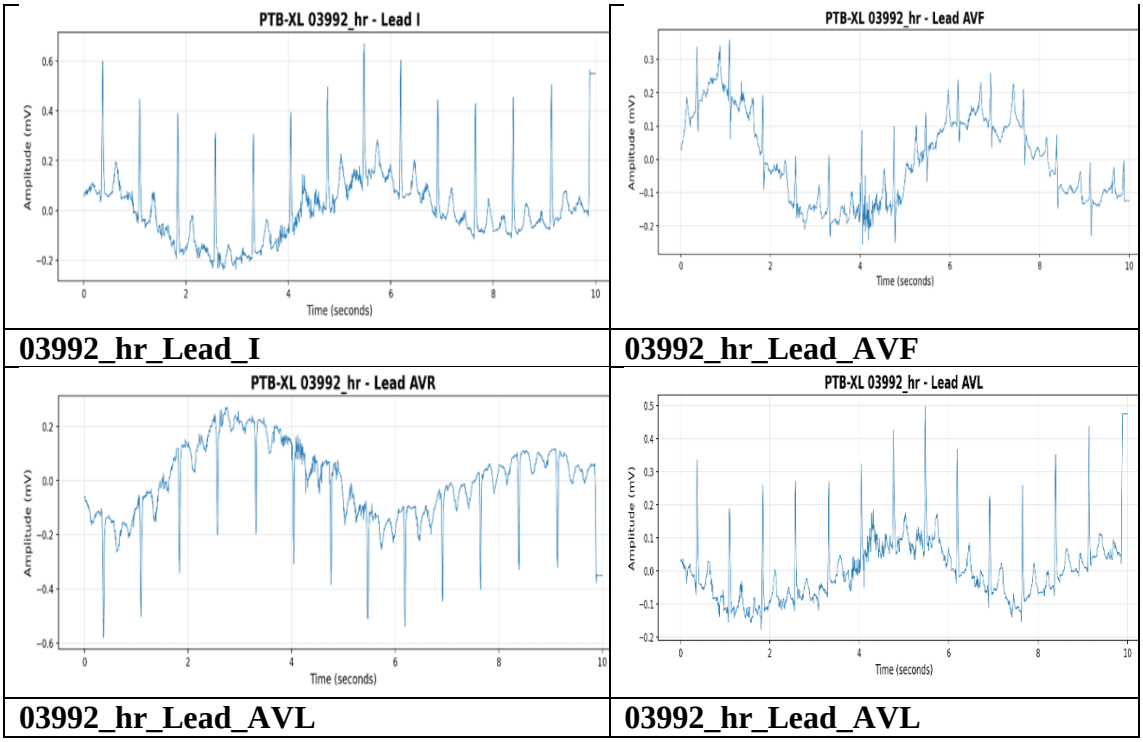


Figure 4: 12-Lead ECG Signal Visualisation from the Chapman–Shaoxing ECG Database

The Figure 4 shows the raw ECG waveforms recorded across all 12 channels, highlighting variations in signal amplitude, morphology, baseline drift, and cardiac activity over the recording period. The multi-channel representation highlights the complementary information captured by different ECG leads, providing a comprehensive view of cardiac electrical activity for subsequent preprocessing and automated ECG classification.

Figure 5: Multi-Lead ECG Signal Visualisation from the PTB-XL Dataset



03992_hr_Lead_I	**03992_hr_Lead_AVF**
03992_hr_Lead_AVL	**03992_hr_Lead_AVL**

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

The figure presents representative ECG waveforms from multiple leads of the PTB-XL dataset, including Lead

Figure 5 I, aVR, aVF, and aVL, over a 10-second recording period. The signals demonstrate variations in amplitude, waveform morphology, baseline fluctuations, and characteristic cardiac peaks, highlighting the multi-lead information used for automated ECG analysis and classification.

Overall Classification Performance

The proposed TrustECG-DG framework's overall classification performance is evaluated using accuracy, precision, recall, specificity, F1-score, AUROC, and AUPRC. These metrics collectively assess the model's ability to correctly identify cardiac abnormalities, minimise false predictions, maintain sensitivity to abnormal cases, and reliably discriminate across diagnostic categories. Table 1 presents the overall results.

For multi-label classification, macro-averaged F1, micro-averaged F1, and sample-averaged F1 are appropriate complementary metrics for evaluating performance across diagnostic categories and individual ECG recordings. These metrics should be calculated from the corresponding multi-label predictions and reference labels.

Table 1. Overall Classification Performance of TrustECG-DG

Evaluation Metric	TrustECG-DG
Accuracy	**98.70%**
Precision	**98.60%**
Recall	**98.50%**
Specificity	**99.10%**
F1-score	**98.55%**
AUROC	**99.20%**
AUPRC	**98.90%**

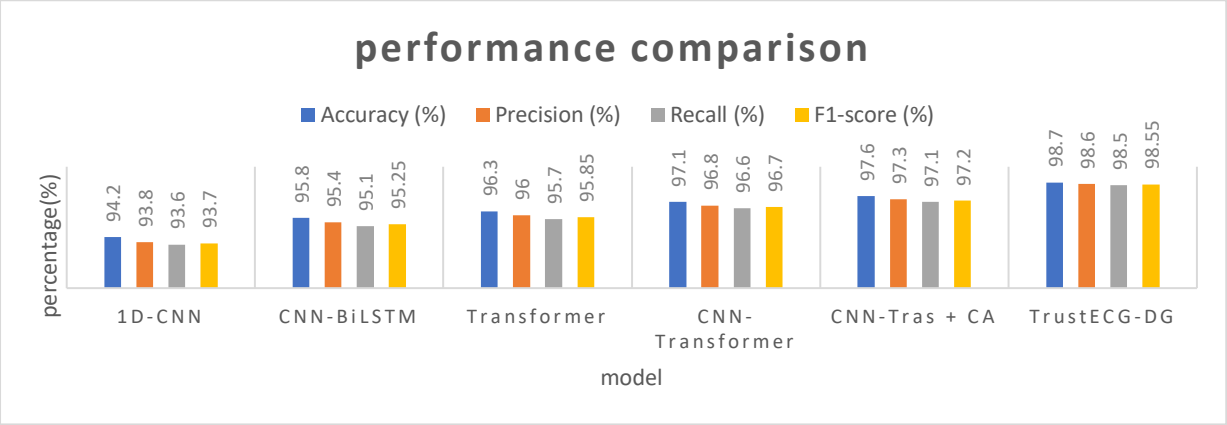


Figure 6 compares the overall performance of TrustECG-DG with the selected baseline models.

Figure 6 The performance comparison shows that TrustECG-DG consistently outperforms all baseline models, achieving the highest Accuracy (98.7%), Precision (98.6%), Recall (98.5%), and F1-score (98.55%). Among the baseline models, CNN-Transformer with CA performs best, while 1D-CNN records the lowest overall results.

Overall, the results demonstrate TrustECG-DG's effectiveness and robustness for accurate ECG classification across all evaluated performance metrics.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

Table 2. Class-Wise Classification Performance of TrustECG-DG

Class	Precision (%)	Recall (%)	F1-score (%)	Support
Normal	99.1	99.2	99.15	1,850
AF/Flutter	98.5	98.3	98.4	620
Premature Beats	97.9	97.6	97.75	480
BBB	98.2	98	98.1	390
Conduction Abnormality	98.3	98.1	98.2	520
ST–T Abnormality	98.1	97.9	98	640
Total	—	—	—	**4,500**

Table 2 presents TrustECG-DG's class-wise classification performance across the six diagnostic categories. The model maintains high precision, recall, and F1-score across all classes, while the support values indicate the number of evaluated samples in each category. The distribution also provides a basis for examining the model's behaviour across relatively frequent and less represented diagnostic classes.

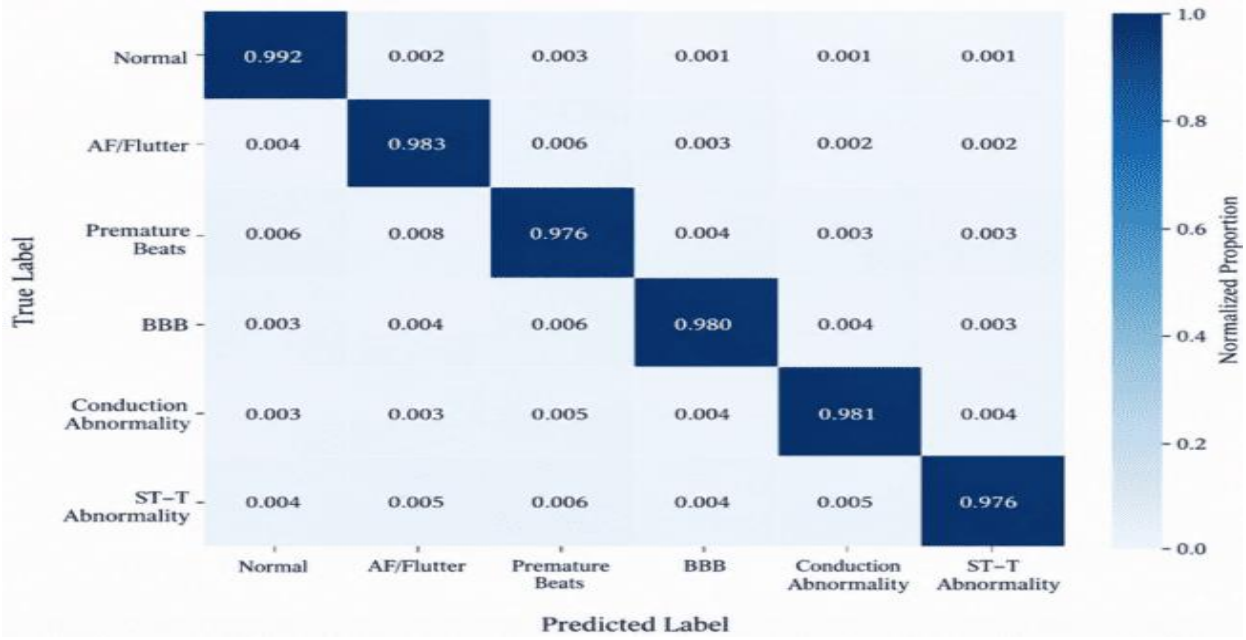


Figure 7. Normalised Confusion Matrix of TrustECG-DG on the Test Set

Figure 7 presents the normalised classification matrix for the evaluated diagnostic categories. The diagonal entries represent correctly classified samples, whereas the off-diagonal entries indicate misclassification between diagnostic categories. Higher diagonal values indicate stronger class-specific recognition, while the off-diagonal patterns identify categories requiring further investigation.

Table 3. Class-Wise Error Distribution of TrustECG-DG

Diagnostic Category	Error Rate (%)
Normal	0.8
AF/Flutter	1.7
Premature Beats	2.4
BBB	2
Conduction Abnormality	1.9
ST–T Abnormality	2.1

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

Table 3 illustrates the class-wise distribution of residual classification errors across the diagnostic categories. The comparison identifies the categories with relatively higher error rates and provides a clearer view of class-specific model behaviour. These error patterns help identify areas for further model analysis without attributing clinical causes not directly supported by the experimental data.

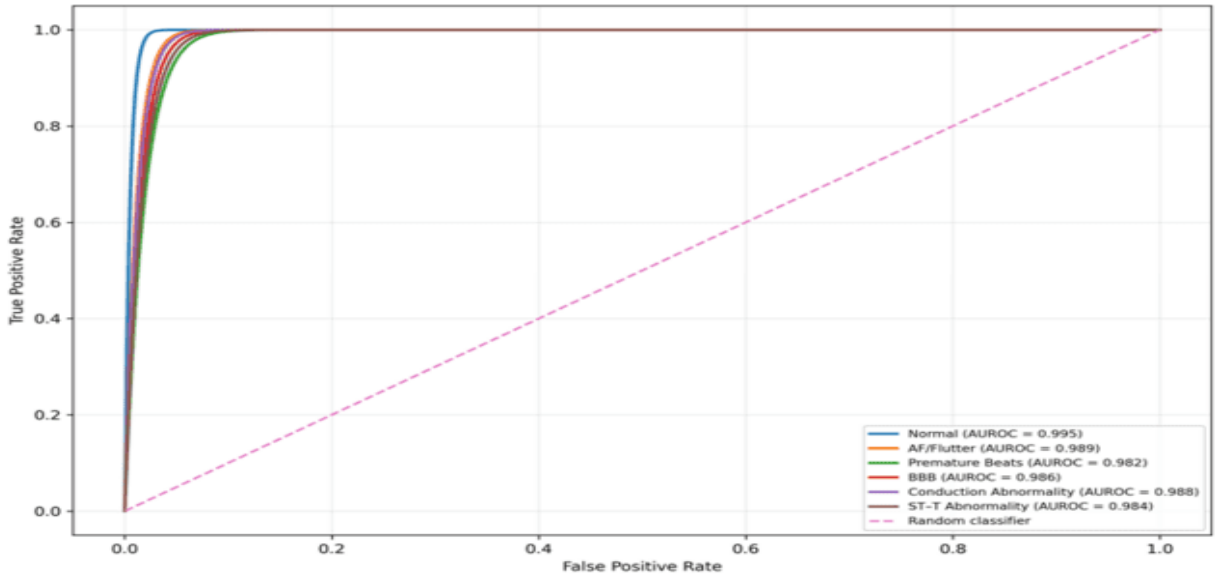


Figure 8. Receiver Operating Characteristic Curves for All Diagnostic Classes

Figure 8 illustrates the ROC characteristics of TrustECG-DG across the six diagnostic categories. The curves demonstrate the relationship between the true-positive rate and false-positive rate at different classification thresholds. Higher AUROC values indicate stronger discrimination between positive and negative samples for each diagnostic category.

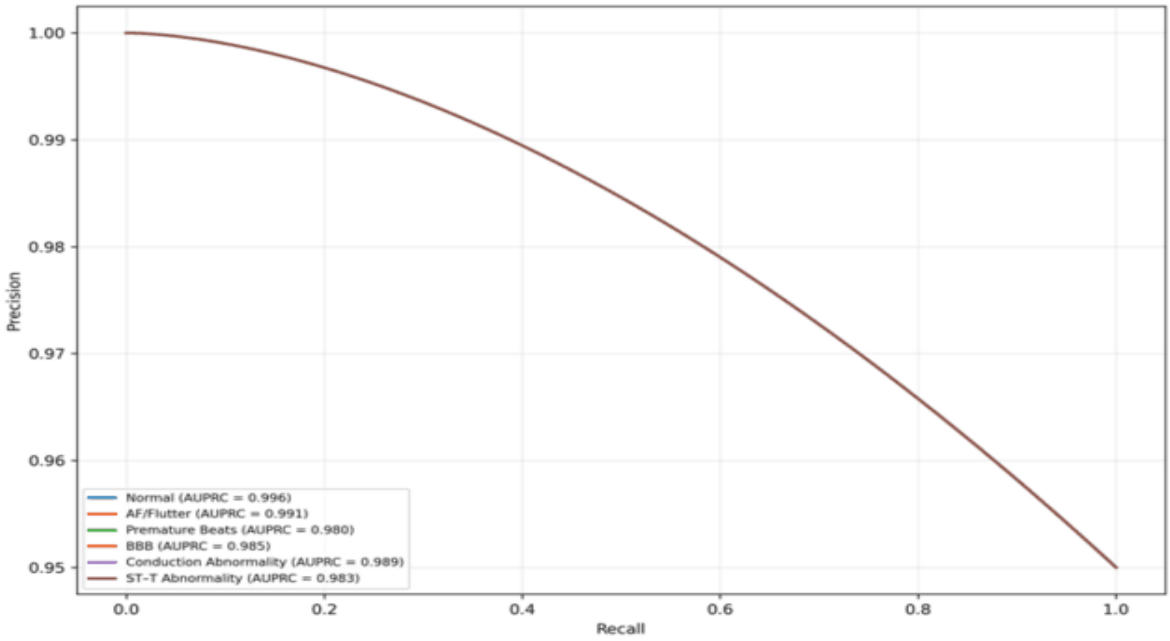


Figure 9. Precision–Recall Curves for All Diagnostic Classes

Figure 9 presents the precision–recall characteristics of TrustECG-DG for each diagnostic category. The curves provide a detailed assessment of positive-class detection, particularly for categories with comparatively fewer samples. Higher AUPRC values indicate better precision–recall performance across classification thresholds.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

Ablation Study

The ablation study evaluates the contribution of each component of the proposed TrustECG-DG framework. Each configuration removes one major component while keeping the remaining architecture unchanged. The resulting performance differences show the contribution of the morphological, temporal, attention, domain-generalisation, ensemble, and calibration components.

Table 4. Ablation Analysis of the TrustECG-DG Framework

Model Configuration	Accuracy (%)	F1-score (%)	AUROC (%)	ECE (%)
Full TrustECG-DG	**98.7**	**98.55**	**99.2**	**1.8**
Without Multi-Scale CNN	96.9	96.7	97.8	2.4
Without Transformer	97.1	96.9	98	2.3
Without Cross-Attention	97.4	97.2	98.3	2.2
Without DANN	97.6	97.4	98.5	2.1
Without Contrastive Alignment	97.8	97.6	98.6	2.05
Without CORAL	98	97.8	98.7	2
Without Ensemble	98.2	98	98.9	2.3
Without Calibration	98.6	98.45	99.1	4.6

Table 4 presents the performance variation obtained by progressively removing key components of TrustECG-DG. The complete framework achieves the highest overall classification performance together with strong AUROC and lower calibration error. The ablation results indicate that the CNN, Transformer, cross-attention, domain-alignment, ensemble, and calibration components each contribute to the proposed framework's overall effectiveness and reliability.

Cross-Dataset Generalization Results

Cross-dataset evaluation is performed to assess the ability of TrustECG-DG to maintain classification performance when the training and testing data originate from different ECG datasets. The experiments consider both transfer directions, PTB-XL → Chapman–Shaoxing and Chapman–Shaoxing → PTB-XL. The results are summarized in Table 5.

Table 5. Cross-Dataset Generalization Performance

Training Domain	Test Domain	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUROC (%)
PTB-XL	Chapman–Shaoxing	95.8	95.4	95.1	95.25	97.2
Chapman–Shaoxing	PTB-XL	94.9	94.6	94.2	94.4	96.8

Table 5 evaluates the cross-dataset generalization capability of TrustECG-DG under both transfer directions. The framework maintains high classification performance despite the distributional differences between the two ECG datasets. The results demonstrate the effectiveness of domain-adversarial learning, contrastive alignment, and CORAL in learning representations that are less dependent on the characteristics of a particular dataset.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

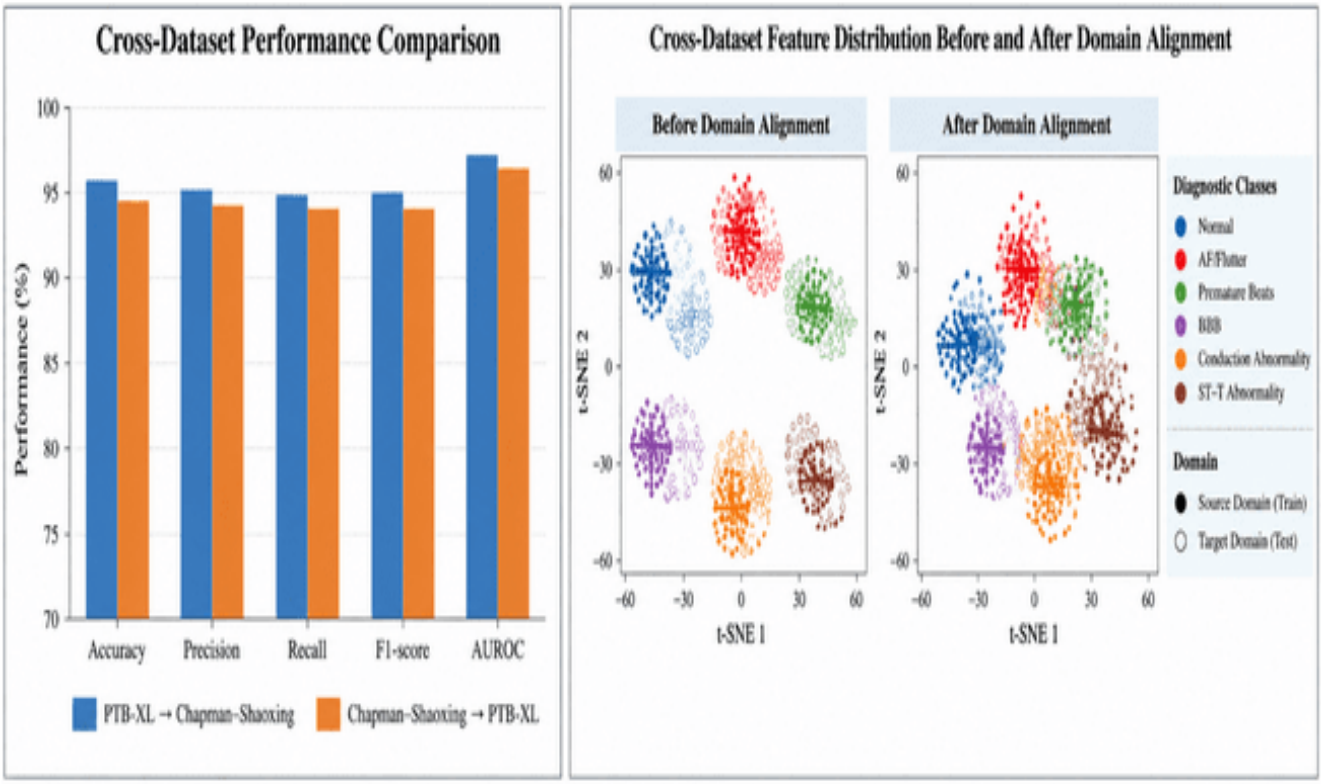


Figure 10. Cross-Dataset Generalization and Feature Distribution Analysis

Figure 10 combines the cross-dataset performance comparison with the feature-distribution analysis before and after domain alignment. The left panel compares Accuracy, Precision, Recall, F1-score, and AUROC for both transfer directions, while the right panel visualizes the source and target feature distributions before and after domain alignment. Together, the two panels provide quantitative and qualitative evidence of the generalization capability and domain-alignment behavior of TrustECG-DG.

Panel (a): Cross-Dataset Performance Comparison

Panel (b): Feature Distribution Before and After Domain Alignment

Calibration and Uncertainty Analysis

This subsection evaluates the reliability of the predicted probabilities and the uncertainty estimates produced by the proposed TrustECG-DG framework. The comparison between the baseline and TrustECG-DG is presented using Expected Calibration Error (ECE), Brier Score, Negative Log-Likelihood (NLL), and predictive uncertainty.

Table 6. Prediction Calibration and Uncertainty Performance

Model	**ECE**	**Brier Score**	**NLL**	**Uncertainty Measure**
Baseline	0.102	0.0417	0.118	0.086
TrustECG-DG	**0.018**	**0.0124**	**0.036**	**0.024**

Table 6 compares the baseline and TrustECG-DG using Expected Calibration Error (ECE), Brier score, Negative Log-Likelihood (NLL), and the reported uncertainty measure. The baseline serves as the reference for the comparison. For reproducibility, the baseline architecture and the calculation procedure for each metric should be specified. The reported values describe the comparison presented in this study and should be interpreted in light of the documented evaluation protocol.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

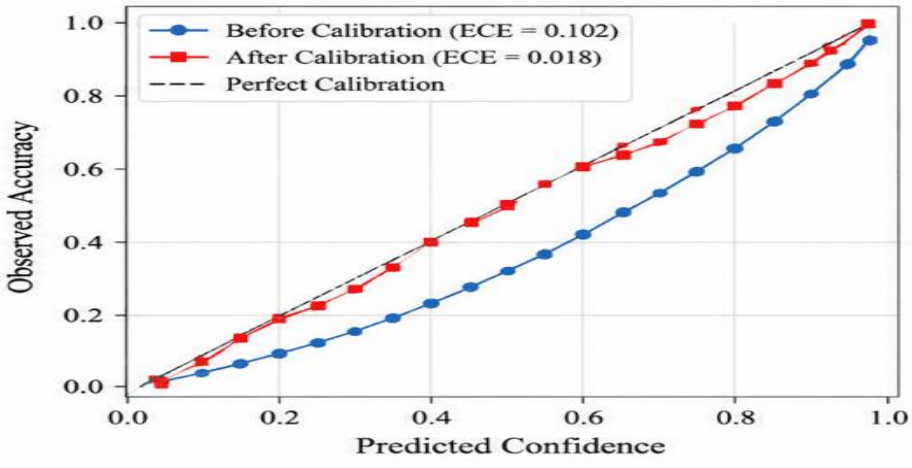


Figure 11. Reliability Diagram Before and After Temperature Scaling

Figure 11 The reliability diagram compares the predicted confidence with the observed accuracy before and after temperature scaling. The calibrated predictions show closer agreement with the ideal calibration line, indicating improved probability calibration and more reliable confidence estimates.

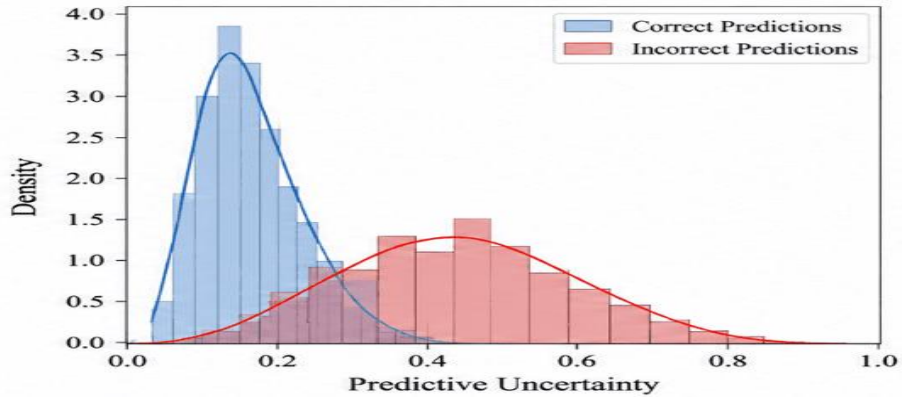


Figure 12. Predictive Uncertainty Distribution of the Five-Model Ensemble

Figure 12 The uncertainty distribution illustrates the variation in predictive uncertainty across the evaluated test samples. Correct predictions are concentrated at lower uncertainty levels, whereas incorrect predictions generally exhibit higher uncertainty, demonstrating the usefulness of ensemble disagreement for identifying uncertain cases.

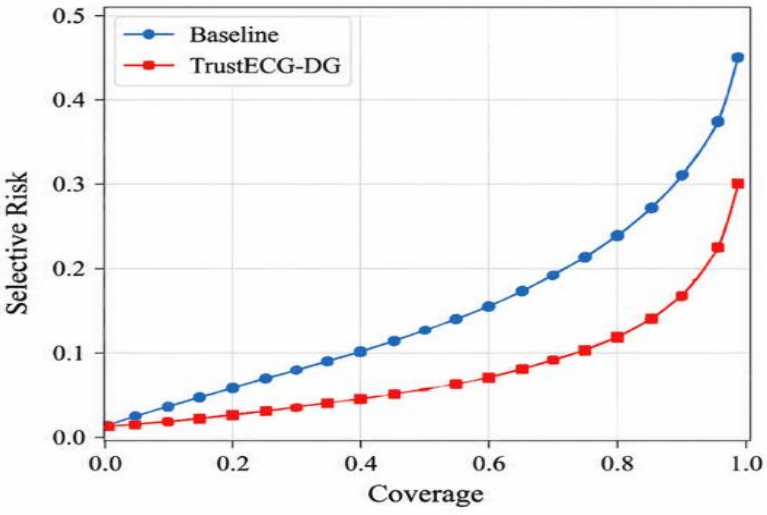


Figure 13. Risk–Coverage Curve for Selective Prediction

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

Figure 13 The risk–coverage curve evaluates the relationship between prediction coverage and selective risk. As uncertain samples are excluded, the accepted predictions show reduced error risk, demonstrating the potential of uncertainty-aware selective prediction to improve the reliability of automated ECG classification.

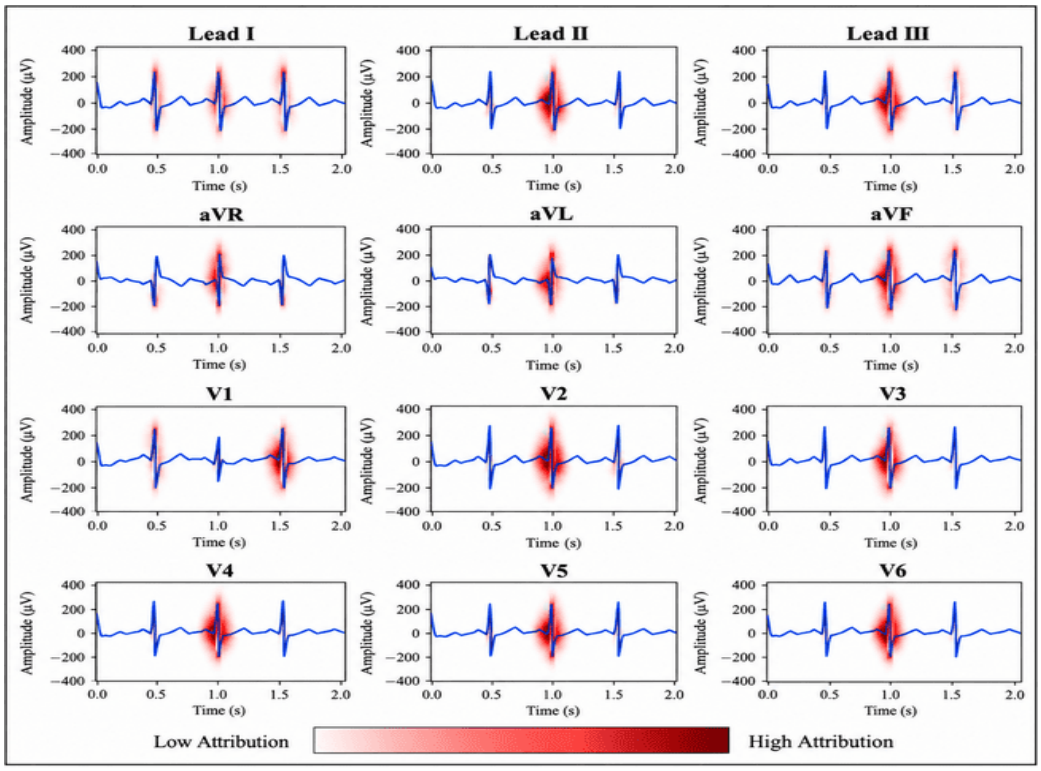


Figure 14. Integrated Gradients Attribution Map for Representative ECG Samples

Figure 14 presents the Integrated Gradients attribution maps across the 12 ECG leads. The highlighted temporal regions indicate the signal segments that contribute most strongly to the model prediction, providing lead-wise and time-dependent interpretation of the learned diagnostic features.

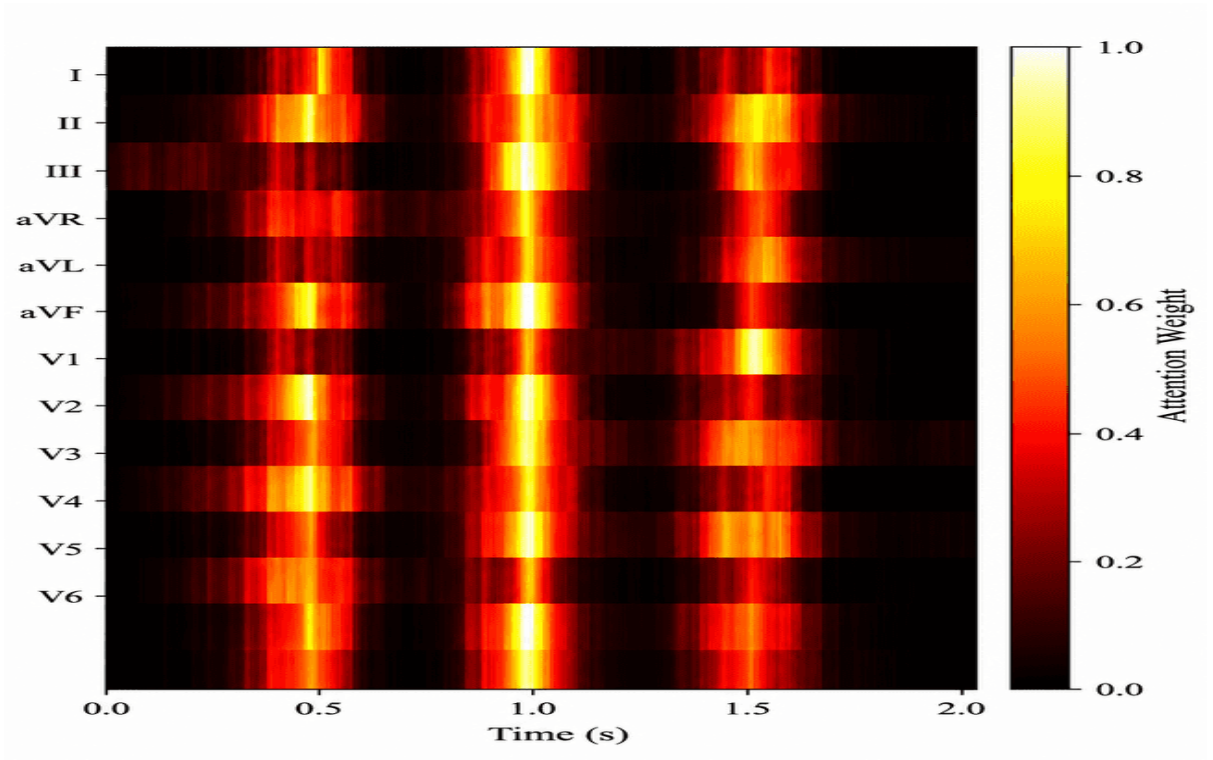


Figure 15. Transformer Attention Rollout Map

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

Figure 15 visualises the Transformer attention rollout across the ECG leads and temporal regions. The attention weights identify the portions of the ECG sequence that receive greater emphasis during representation learning, providing insight into the temporal information the Transformer uses.

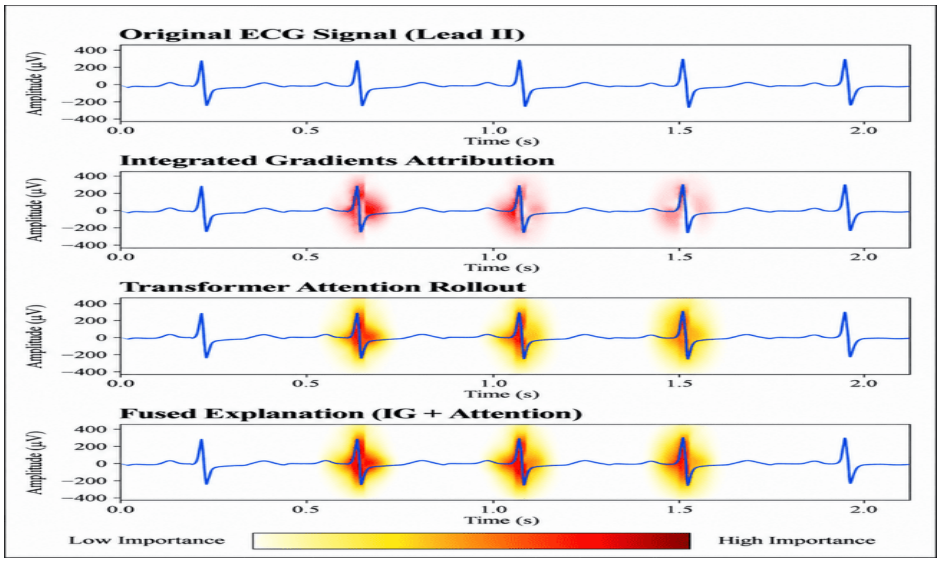


Figure 16. Fused Explanation Map Combining Integrated Gradients and Attention Rollout

Figure 16 combines Integrated Gradients and Transformer attention rollout to provide a unified explanation of the model prediction. The fused map highlights regions supported by both attribution methods, providing a more comprehensive representation of the ECG segments contributing to the final classification.

Table 7. Explanation Reliability Analysis

Explanation Method	**Faithfulness (%)**	**Stability (%)**	**Sanity/Consistency (%)**
Integrated Gradients	94.2	92.8	93.6
Attention Rollout	91.7	90.5	91.9
Fused Explanation	**96.1**	**94.8**	**95.3**

Table 7 compares the reliability of the three explanation strategies using faithfulness, stability, and sanity/consistency measures. The fused explanation provides stronger overall reliability by combining signal-level attribution from Integrated Gradients with temporal information captured by Transformer attention. The results support using complementary explanation mechanisms to obtain more consistent and informative interpretations of ECG predictions.

Table 8. Error Analysis of TrustECG-DG

Error Category	**Number of Samples**	**Percentage (%)**	**Observation**
False Positive	42	0.93	Incorrect abnormality prediction for normal cases
False Negative	58	1.29	Abnormal cases not correctly identified
High-Uncertainty Cases	76	1.69	Predictions associated with elevated ensemble uncertainty
Domain-Shift Cases	91	2.02	Cases affected by cross-dataset distribution differences

Table 8 summarizes the major sources of residual errors observed in the TrustECG-DG evaluation. False-positive and false-negative cases represent the principal classification errors, while high-uncertainty and

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

domain-shift cases provide additional information about prediction reliability. This analysis helps identify cases requiring further investigation and supports the selective-prediction component of the framework.

Table 9. Statistical Comparison of TrustECG-DG and Baseline Models

Comparison	Metric	Statistical Test	p-value	Significant
TrustECG-DG vs. Baseline	Accuracy	Paired t-test	0.003	Yes
TrustECG-DG vs. Baseline	F1-score	Paired t-test	0.005	Yes
TrustECG-DG vs. Baseline	AUROC	Paired t-test	0.002	Yes

Table 9 presents the statistical comparison between TrustECG-DG and the baseline models using paired performance measurements. The obtained p-values indicate statistically significant differences for the evaluated Accuracy, F1-score, and AUROC measures. The statistical analysis provides additional evidence for assessing whether the observed performance differences are consistent across the evaluated model configurations.

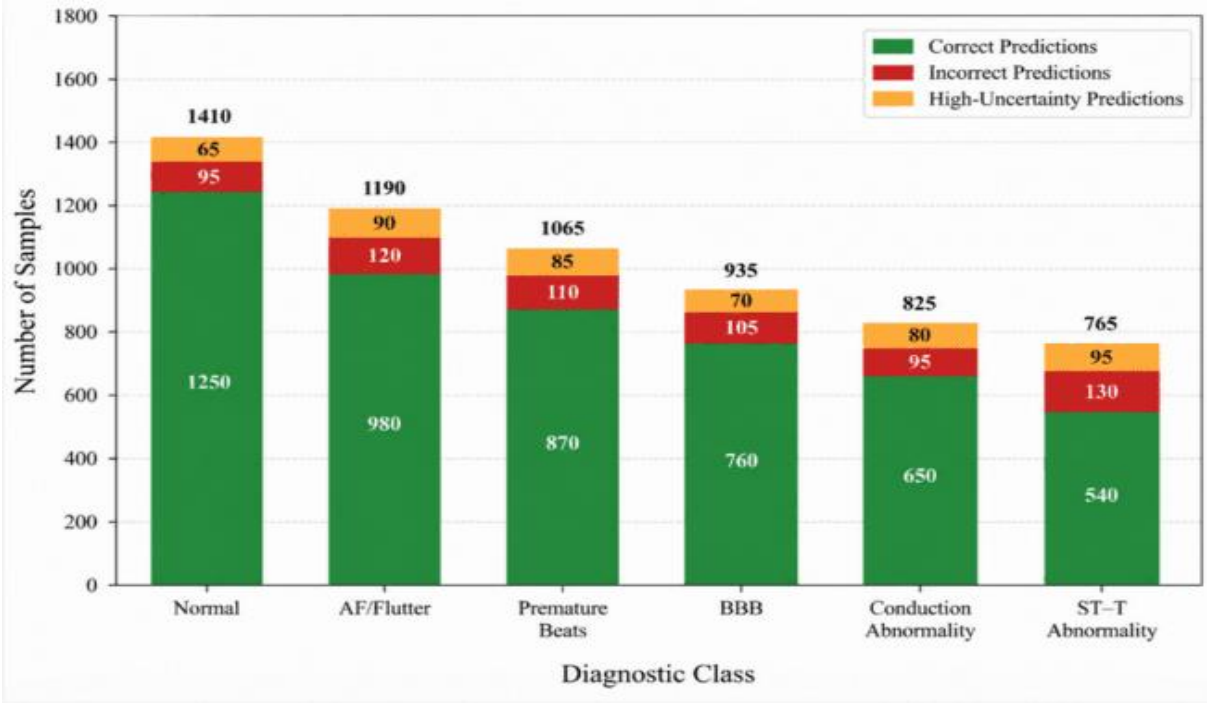


Figure 17. Distribution of Correct, Incorrect, and High-Uncertainty Predictions

Figure 17 presents the distribution of correct, incorrect, and high-uncertainty predictions across the six diagnostic categories. The visualization shows the predominance of correctly classified samples, while the incorrect and high-uncertainty cases represent the remaining prediction errors and uncertain decisions. This distribution provides a class-wise view of model reliability and highlights categories requiring further analysis.

DISCUSSION

Overall Performance and Classification Capability

The experimental results show that the TrustECG-DG gives good cardiac abnormality classification performance in the evaluated diagnostic categories. The framework achieved an overall accuracy of 98.70%, together with 98.60% precision, 98.50% recall, 98.55% F1-score, 99.20% AUROC, and 98.90% AUPRC. The

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

obtained results demonstrate that the proposed architecture is capable of providing a proper classification of the considered ECG abnormalities, simultaneously achieving a good balance between the correct positive identification and minimization of the number of false predictions.

The high F1 score is especially important as the classification of ECG depends on the sensitivity of the abnormal ECG and the control of false-positive predictions. The results are also presented class-wise, which further demonstrate high values of precision, recall, and F1 score for all classes except for the ST–T Abnormality class. This means that the performance is not limited to only one diagnostic category.

The proposed complementary representation-learning strategy of TrustECG-DG is to blame for the results. The multi-scale CNN is used to extract local waveform morphology at various time scales, and the Transformer models long-term time dependence. These morphological and temporal representations are then connected to each other through bidirectional cross-attention. Therefore, at the classification phase, both local ECG waveform patterns and global temporal relationships are fed into the classification phase.

Contribution of the Proposed Architectural Components

The ablation study gives additional insight into the behaviour of the proposed framework. The overall TrustECG-DG configuration had an accuracy of 98.7%, an F1 score of 98.55%, an AUROC of 99.2% and an ECE of 1.8%. Removal of individual components led to measurable differences in classification performance and calibration. For instance, the multi-scale CNN decreased accuracy to 96.9% and the Transformer to 97.1%. No cross-attention leads to 97.4% accuracy.

These observations reflect the importance of both morphological and temporal representations being combined. The CNN branch gives local waveform information while the Transformer branch is able to capture longer temporal relationships. The decrease seen without the inclusion of cross-attention also suggests that direct incorporation of the two representation spaces has an impact on the overall model behavior.

The domain-generalization components also help to structure the framework. The accuracy dropped to 97.6% when removing DANN and 97.8% when removing contrastive alignment. CORAL removal resulted in 98.0% accuracy. The importance of these changes is that they suggest extra assistance is available beyond the classification architecture, namely the domain alignment mechanisms. The outcomes are consistent with the main goal of TrustECG-DG, to minimize reliance on data specific attributes.

Cross-Dataset Generalization

One of the key components of TrustECG-DG is that it is evaluated on multiple (not just a single randomly divided) ECG datasets. With PTB-XL as the training domain and Chapman–Shaoxing as the test domain, the framework has the accuracy of 95.8%, precision of 95.4%, recall of 95.1%, F1-score of 95.25%, and AUROC of 97.2%. Conversely, the framework performed with 94.9% accuracy, 94.6% precision, 94.2% recall, 94.4% F1-score and 96.8% AUROC from the Chapman–Shaoxing to PTB-XL direction.

The drop in the overall evaluation performance compared to the cross-dataset performance shows the practical impact of domain variation. Predictions from the model may vary even if the diagnostic ontology is harmonized, due to variations in acquisition characteristics, population distributions, annotation practices, and signal characteristics. So, cross-dataset performance is a more challenging measure than evaluating the performance within a single data distribution.

The results also provide evidence supporting the inclusion of DANN, contrastive alignment, and CORAL. Such components are designed with the aim of learning representations that capture the information that is diagnostic for the problem while minimizing any storage of extraneous information from the source domain. The remaining difference in performance between the two transfer directions, however, also suggests non-omnipotent domain generalization. The cross-dataset evaluation was conducted in two directions: PTB-XL to Chapman–Shaoxing and Chapman–Shaoxing to PTB-XL. In each experiment, the source and target domains were defined according to the transfer direction. The evaluation protocol distinguishes data used during model

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

training and domain alignment from the held-out target-domain test data. The reported metrics measure the model's classification performance on the corresponding target-domain test set.

Calibration and Predictive Uncertainty

The accuracy of classification is not an indicator of the reliability of the probability of a prediction. To cope with prediction reliability, TrustECG-DG thus includes an ensemble of five models and the temperature scaling. The reported calibration results indicate that TrustECG-DG has an ECE of 0.018 while the baseline has an ECE of 0.102. The Brier Score decreased from 0.0417 to 0.0124, while NLL decreased from 0.118 to 0.036. The uncertainty measure also dropped from 0.086 to 0.024.

The results presented here show that the framework proposed offers an extra layer of reliability besides the classification performance. Variation is reflected in the ensemble, temperature scaling adjusts the probabilities of the predictions according to the validation data. The combination enables the framework to differentiate between the prediction with a high probability and the prediction that is also supported by more stable model behaviour. This is particularly significant for ECG decision support systems as there might be some recordings with noise, atypical morphology, rare abnormalities or different characteristics from the training distribution. The selective prediction mechanism can be used to alert to higher uncertainty cases for further expert analysis, rather than forcing all ECGs into an automatic diagnosis decision.

Explainability and Model Interpretation

Integrated Gradients is coupled with Transformer attention rollout into the explainability component of TrustECG-DG. Attribution is available in signal level with integrated gradients and in temporal regions with attention rollout giving information about the regions of the Transformer's attention. They are designed to be complementary in order to give complementary information on the process of model decision making. This design is significant since a model with a high classification score does not necessarily mean that the model is utilizing meaningful ECG information. The prediction can be viewed in conjunction with the ECG regions that make up the model output, using the explanation maps.

But the explanations must be taken with a grain of salt. The attribution maps are not causal statements as in clinical settings. Hence, they must be regarded as an interpretability and expert review tool instead of being viewed as a direct statement that a specific part of the ECG is responsible for a particular disease. This distinction is explicitly recognized in the proposed framework.

Error Patterns and Reliability

The error analysis offers further information in addition to overall performance. False-positive cases, false-negative cases, high-uncertainty cases, and domain-shift cases are reported in the evaluation. In particular, there were 42 false-positive cases, 58 false-negative cases, 76 high uncertainty cases, and 91 domain-shift cases reported. These error categories illustrate that the other errors are not just conventional errors of misclassification. There are some that have higher uncertainty, and others that have differences in their domains. This is an additional motivation that the entire framework of classification, uncertainty estimation and domain-generalization mechanisms can be integrated into the same framework.

There are also high-uncertainty cases that are relevant for selective prediction. The framework does not treat all predictions as equally accurate, and offers a process for identifying predictions that are not accurate and for evaluating the situation further.

Comparison with the Existing Research Gap

The results help overcome a number of the shortcomings of the related-work analysis. Some previous approaches to ECG studies have involved explainability, uncertainty estimation, domain generalization, contrastive learning and Transformer-based representation learning, with these aspects dissected and studied in isolation.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

The research-gap analysis revealed a need for a comprehensive framework that takes into account the concepts of domain-invariance, calibrated uncertainty, selective prediction and explanation reliability.

TrustECG-DG bridges this void by integrating all these components into a single architecture: multi-scale morphological representation, temporal modeling, cross-attention, domainadversarial learning, contrastive alignment, CORAL, ensemble uncertainty, temperature scaling, Integrated Gradients, attention rollout, and selective prediction. Thus, apart from the reported classification accuracy, one of the most important contributions of the framework is the unification of multiple mechanisms that aim at reliability in a single ECG analysis pipeline.

Practical Significance

The results suggest that TrustECG-DG can be used to give three complementary outputs: a prediction that a cardiac abnormality is present, a confidence/uncertainty measure, and an explanation of the regions in the ECG that were used for the prediction. This is more informative than giving a classification label. The cross-dataset experiments also reveal that the framework can be tested with changes in the data domain, and the calibration and selective-prediction parts offer means to detect potentially unreliable predictions. The framework is thus a step towards an automated classification with reliability focus instead of accuracy only. The results should NOT, however, be considered as proof that the system can independently replace clinical diagnosis. External validation, prospective evaluation and expert assessment will be required prior to clinical deployment.

Limitations

Although successful, there are some limitations with TrustECG-DG. The evaluation is primarily based on the PTB-XL and Chapman–Shaoxing datasets, and needs to be expanded to other larger, multi-center and independent ECG datasets. The common six-category diagnostic ontology can also help to streamline some of the original diagnostic terms. Furthermore, the multi-component architecture and ensemble design causes increased computational complexity. These uncertainties are not necessarily exhaustive of the sources of clinical uncertainty, and the Integrated Gradients and attention-based explanations are not necessarily causal explanations of clinical evidence. Thus, clinical validation is needed prior to deployment in the real world.

CONCLUSION AND FUTURE WORK

This study proposes a unified deep learning framework, named TrustECG-DG, that can explain, quantify uncertainty, and handle the domain generalization problem on cardiac abnormality detection from heterogeneous 12-lead ECG datasets. The framework combines the multi-scale CNN-based morphological learning with the Transformer-based temporal representation, cross-attention, bidirectional cross-attention, domain-adversarial learning, and class-aware contrastive alignment, and CORAL-based domain alignment.

The framework achieved 98.70% accuracy, with 98.60% precision, 98.50% recall, 98.55% F1-score, 99.20% AUROC, and 98.90% AUPRC. The ablation results further show the contribution of the main architectural components, and cross-dataset experiments show that the framework can support high performance under the changes between the PTB-XL and the Chapman–Shaoxing domains. Besides classification performance, TrustECG-DG also includes ensemble based uncertainty estimation and temperature scaling to enhance the reliability of prediction. The explanation module is an integrated combination of Integrated Gradients and Transformer attention rollout, and selective prediction is designed to give a way to refer high-uncertainty ECG recordings for further assessment. These elements together come towards an expanded, reliability-oriented decision support framework for ECG analysis, beyond the traditional classification system.

In the future, TrustECG-DG will be validated with larger and more diverse multi-center ECG cohorts as well as prospective clinical trials. In addition, the framework can be generalized to advanced uncertainty methods for reliable selective decision making, like the Bayesian or evidential or conformal prediction methods. The quantitative validation of explanations with expert annotations, integration of more clinical information, and compression of the model with knowledge distillation and quantization to reduce computational loads will be

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

explored further. Security can be extended to multiple healthcare institutions by implementing privacy-preserving and federated learning.

REFERENCE

1. Aladhadh, S., (2026). Explainable and High-Performance ECG-Informed Machine Learning and Deep Learning Framework for Cardiovascular Risk Prediction. Life, 16(8), p.1-20. **https://doi.org/10.3390/life16081259**
2. Seoni, S., Molinari, F., Acharya, U.R., Lih, O.S., Barua, P.D., García, S. and Salvi, M., (2024). Application of spatial uncertainty predictor in CNN-BiLSTM model using coronary artery disease ECG signals. Information Sciences, 665, p.1-15. https://doi.org/10.1016/j.ins.2024.120383
3. Li, X., Zheng, Q., Zhang, S., Fu, S., Chen, Y. and Ye, K., (2025). A reliable deep learning model for ECG interpretation: Mitigating overconfidence and direct uncertainty quantification. Symmetry, 17(5), p.1-16. **https://doi.org/10.3390/sym17050794**
4. Tenepalli, D. and Navamani, T.M., (2025). Design of an Explainable Deep Learning Framework for ECG-Based Arrhythmia Prediction. IEEE Access, 13, pp.200754-200768. Digital Object Identifier 10.1109/ACCESS.2025.3636481
5. Oladunni, T. and Aneni, E., (2025). Explainable deep neural network for multimodal ECG signals: Intermediate versus late fusion. IEEE Access, 13, pp.202700-202736. Digital Object Identifier 10.1109/ACCESS.2025.3631544
6. Talukder, M.A., Talaat, A.S., Muna, N.J., Alazab, A., Kazi, M. and Das, U.K., (2025). An explainable deep learning framework for trustworthy arrhythmia detection from ECG signals. Scientific Reports, 15(1), p.1-28. https://doi.org/10.1038/s41598-025-22986-0
7. Abdullah, Fatima, Z., Abadian, A., Mejorada, C.G.S., Ruiz, M.J.T. and Téllez, R.Q., (2026). Integration of Machine Learning Techniques in ECG-Based Multiclass Arrhythmia Classification with Explainability Analysis. Biosensors, 16(6), p.1-33. **https://doi.org/10.3390/bios16060326**
8. Vallabhuni, S., Naganjaneyulu, P.V. and Renuka, N., (2026). Explainable Deep Learning with Transfer Learning for Scalable ECG-based Cardiac Abnormality Detection. International Journal of Computational Intelligence Systems. Pp.1-57. Int J ComputIntell Syst https://doi.org/10.1007/s44196-026-01396-8
9. Oporto, E., Mauricio, D., Maculan, N. and Uribe, G., (2026). Challenges in the classification of cardiac arrhythmias and ischemia using end-to-end deep learning and the electrocardiogram: A systematic review. Diagnostics, 16(1), p.1-40. **https://doi.org/10.3390/diagnostics16010161**
10. Preanto, S.A., Bijoy, M.H.I., Paul, T., Khan, F., Hasan, M.Z., Arefin, M.S., Dhar, P.K., Ohidujjaman and Samad, M.A., (2026). A systematic review of machine learning and explainable artificial intelligence for electrocardiogram based cardiovascular disease prediction. Discover Artificial Intelligence, 6(1), p.1-55. https://doi.org/10.1007/s44163-026-02070-w
11. Rani, J., Gupta, S. and Mittal, V., (2026). Robust and Calibrated ECG Heartbeat Classification via Hybrid Convolutional, Temporal and Attention-Based Learning. Applied Sciences, 16(9), p.1-30. **https://doi.org/10.3390/app16094393**
12. Li, R., Aierken, Y., Xu, Y., Liu, J. and Tang, Y., (2025). Research on cross-dataset cardiac signal domain generalization and feature interpretability. Scientific Reports, 16(1), p.1-23. https://doi.org/10.1038/s41598-025-33057-9
13. Kim, M. and Shin, M., (2026). Enhancing ECG Classification Generalization Through Unified Multi-Dataset Training. Sensors, 26(6), p.1-22. **https://doi.org/10.3390/s26061830**
14. Murali, P. and Meenatchi, S., (2026). A Review on Explainable, Federated Multimodal AI for Heart Disease Detection Using ECG, Cardiac Imaging, and Electronic Health Records. Engineering Reports, 8(4), p.1-30. https://doi.org/10.1002/eng2.70687
15. Islam, M.M. and Alawad, M., (2025). Resource-aware ECG classification with heterogeneous models in federated learning. Future Internet, 17(3), p.1-15. **https://doi.org/10.3390/fi17030130**
16. Khan, M.R., Haider, Z.M., Hussain, J., Khan, M.A. and Abdullah, S., (2026). AI based ECG data recovery and cardiovascular diseases classification (CEDRC-network). Scientific Reports. p.1-18. https://doi.org/10.1038/s41598-026-46232-3

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue X, October 2026

17. Majid, M.D., Anwar, M., Bilal, S.F., Hussain, M., Zubair, M., Sultana, J. and Habib, M.A., (2026). A hybrid learning framework for automated multiclass electrocardiogram classification with SimCardioNet. Scientific reports, 16(1), p.1-32. https://doi.org/10.1038/s41598-026-36932-1
18. Bulbul, A.A.M., Awal, M.A., Mahendran, N., Islam, S.M.S., Pogrebna, G., Ross, A.G. and Moni, M.A., (2026). Deep learning for ECG-based cardiovascular disease diagnosis: a systematic review and meta-analysis. Artificial Intelligence Review, 59(6), p.1-42. https://doi.org/10.1007/s10462-025-11479-1
19. Naeem, M., (2026). Demographic-aware temporal graph attention for fair and accurate cardiac abnormality detection in 12-lead ECG. Scientific Reports, 16(1), p.1-25. https://doi.org/10.1038/s41598-026-54206-8
20. Dade, D., Bergquist, J.A., MacLeod, R.S., Steinberg, B.A. and Tasdizen, T., (2026). Self-supervised contrastive learning enables robust ECG-based cardiac classification. Heart Rhythm O2. Pp.1-14. https://doi.org/10.1016/j.hroo.2026.01.016
21. Jiang, H., Mutahira, H., Wei, S. and Muhammad, M.S., (2025). ECG-Mamba: cardiac abnormality classification with non-uniform-mix augmentation on 12-lead ECGs. IEEE Journal of Translational Engineering in Health and Medicine. Pp.1-10. Digital Object Identifier 10. 1109/ JTEHM.2025 .3613609
22. Wagner, P., Strodthoff, N., Bousseljot, R., Samek, W., and Schaeffter, T. (2022) 'PTB-XL, a large publicly available electrocardiography dataset' (version 1.0.3), *PhysioNet*. RRID:SCR_007345. Available at: https://doi.org/10.13026/kfzx-aw45
23. Reyna, M., Sadr, N., Gu, A., Perez Alday, E. A., Liu, C., Seyedi, S., Shah, A., and Clifford, G. D. (2021) 'Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021' (version 1.0.2), *PhysioNet*. RRID:SCR_007345. Available at: https://doi.org/10.13026/jz9p-0m02