

A Reproducible Ensemble Learning Framework for 12-Lead ECG Classification of Myocardial Ischemia and Infarction

Ladan Soltanzadeh

Urmia, Islamic Azad University

Amin Babazadeh Sangar

`bsamin2@liveutm.onmicrosoft.com`

Urmia, Islamic Azad University

Kambiz Majidzadeh

Urmia, Islamic Azad University

Vahid Hosseinpour

Urmia University of Medical Sciences

Article

Keywords: Electrocardiogram, Myocardial Ischemia, Deep Learning, Ensemble Learning, Hyperparameter Optimization, Knowledge Distillation

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11027799/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Reproducible Ensemble Learning Framework for 12-Lead ECG Classification of Myocardial Ischemia and Infarction

Ladan Soltanzadeh¹, Amin Babazadeh Sangar², Kambiz Majidzadeh³, Vahid Hosseinpour⁴,

Corresponding Author: bsamin2@liveutm.onmicrosoft.com

Abstract

Background: Diabetes affects more than 537 million adults worldwide. Cardiovascular complications cause about half of all deaths in this population. Silent myocardial ischemia is common in diabetic patients. Autonomic neuropathy often hides the typical chest pain. As a result, ischemia often goes undetected until it becomes severe. Wearable ECG devices usually record only two or three leads. This reduces sensitivity for detecting ischemic events.

Objective: We aimed to find an optimal teacher model for 12-lead ECG classification of myocardial infarction (MI) and ischemia. Our goal was to provide a high-capacity source for knowledge distillation. This source can train lightweight models for wearable devices.

Methods: We performed a systematic hyperparameter search on the PTB-XL dataset (21,837 records). We tested five architectures with different Transformer depths, hidden dimensions, dropout rates, and learning rates. We also trained two extended variants (V2 and V3) with higher positive-class weights and longer schedules. We built ensembles by averaging the logits of the best models.

¹ Department of Computer Engineering, Urmia., Islamic Azad University, Urmia, Iran (<https://orcid.org/0000-0003-2300-5923>)

² Department of Computer Engineering, Urmia., Islamic Azad University, Urmia, Iran (<https://orcid.org/0000-0002-5190-8460>)

³ Department of Computer Engineering, Urmia., Islamic Azad University, Urmia, Iran (<https://orcid.org/0000-0001-5118-8910>)

⁴ Department of Emergency Medicine, Urmia University of Medical Sciences, Urmia, Iran (<https://orcid.org/0009-0005-4609-6401>)

Results: The wider transformer model (hidden dimension 1024, 3 layers, dropout 0.20) achieved 86.58% validation accuracy and 80.74% sensitivity. Extended training (V2: weight 3.0, 12 epochs) raised validation sensitivity to 88.82% with only a small accuracy drop (85.39%). The ensemble of V1 and V2 gave the best test performance: accuracy 87.40%, sensitivity 76.18%, specificity 92.65%, F1 79.41%, and AUC 93.97%. Adding V3 did not improve performance.

Conclusions: The V1+V2 ensemble is a well-calibrated teacher model for knowledge distillation. Its strong discriminative ability (AUC 93.97%) makes it suitable for training lightweight student networks. These networks can be used in reduced-lead, resource-limited settings. All code and data are publicly available.

Keywords: Electrocardiogram, Myocardial Ischemia, Deep Learning, Ensemble Learning, Hyperparameter Optimization, Knowledge Distillation

1. Introduction

Cardiovascular disease causes about 17.9 million deaths each year. This is nearly one-third of all global deaths (1). Among acute events, myocardial infarction (MI) is a time-sensitive emergency. This includes both STEMI and NSTEMI. Delays in reperfusion lead to irreversible heart muscle loss. This is the "time is muscle" principle. Clinical outcomes depend on fast and accurate diagnosis (2-4). Other heart conditions, such as arrhythmias (2, 5), coronary artery disease(6), heart failure(7), and sleep apnea(8), also add to the global burden(9, 10).

In emergency settings, the 12-lead ECG is the main diagnostic tool. It is widely available and non-invasive. It records electrical activity through three main deflections: the P wave, the QRS complex, and the T wave (2, 11). However, its use has limits. Diagnostic accuracy depends on the reader's skill. Subtle NSTEMI changes are missed in up to 25% of cases(10). A global shortage of trained cardiologists, especially in low-income countries, creates delays(12). Small signal changes may not be seen by the human eye (10). Also, manual interpretation is slow and varies between observers. This can delay treatments like reperfusion therapy(13, 14). These problems show the need for automated, intelligent systems.

Recent advances in AI have started to shape cardiovascular medicine. AI-enabled ECG analysis offers expert-level accuracy at scale. It does not get tired. It gives immediate support in many clinical settings(5, 11, 15, 16). Over the last decade, more computing power, large ECG datasets, and new algorithms have driven AI-ECG research. Many deep learning architectures have shown strong results. These include CNNs (17, 18), LSTMs(19, 20), Transformers (8, 21), ResNet/DenseNet(4, 22, 23), hybrid models(9, 24), and GANs(25). Many of these perform as well as or better than expert cardiologists(5, 15, 26).

Earlier work in this area was limited to narrative reviews and small meta-analyses. These focused on specific MI subtypes or algorithm families. They have been outpaced by rapid advances, such as Transformers (8, 21), contrastive learning, and foundation models. Advanced signal processing techniques, like recurrence plots(15), empirical mode decomposition (27), and wavelet transform(28), have also emerged. Existing systematic reviews have not kept up(10, 29).

Beyond timing, previous meta-analyses had other limitations. Inclusion criteria were applied inconsistently. Spectrum bias was not properly addressed. Few used current quality tools like PROBAST (3, 30). Important factors like patient demographics, care setting, patient variability, and ECG protocols received little attention (3, 23, 31). The issue of imbalanced data (32-34)and model interpretability (2, 12, 35) also remained underexplored. Moreover, only 23.9% of studies in our review performed external validation. This raises concerns about generalizability (36-38).

A rigorous systematic review and meta-analysis is therefore needed. Such a synthesis can inform clinical guidelines, support technology assessments, and shape future research priorities (39-41).

Our contributions:

- We introduce a CNN-Transformer hybrid architecture (Figure 1). It combines a four-block CNN stem with a Transformer encoder and a cross-lead attention layer.
- We conduct a comprehensive, systematic hyperparameter search over five configurations. We vary Transformer depth, hidden dimension, dropout, and learning rate.
- Our analysis shows that increasing hidden dimension improves performance more than increasing depth. This is a key insight for ECG model design.

- We build an ensemble of the two best models. It achieves strong discriminative performance (AUC 93.97%) on the PTB-XL test set.

- This ensemble serves as a well-calibrated, high-capacity teacher model for knowledge distillation.

- All code, configurations, and trained models are publicly available. This ensures full reproducibility.

2. Materials and Methods

2.1 Dataset

We used the PTB-XL dataset, version 1.0.3(42). This is a large, public 12-lead ECG collection. The dataset contains 21,837 records from 18,885 patients. Each record is a 10-second signal, sampled at 500 Hz with 16-bit resolution. The records include diagnostic statements using the SCP-ECG standard.

We mapped the original multi-label annotations into an 8-class scheme. Then we collapsed it into a binary task. The positive class included myocardial infarction (MI) and explicit ischemia (codes: MI, ISC). The negative class covered all other findings. This mapping matches the clinical goal of detecting acute ischemic events. We used the patient-disjoint split provided with the dataset. Folds 1–8 were used for training (n=17,418). Fold 9 served as validation (n=2,183). Fold 10 was used for testing (n=2,198). The test set was not used for any model selection or hyperparameter tuning.

2.2 Signal Preprocessing

We standardized each lead separately. We subtracted the mean and divided by the standard deviation across the 5,000-time samples. For training data, we applied online augmentation:

- 94 - Gaussian noise (SNR 12–30 dB, $p=0.6$)
- 95 - Baseline wander (0.1–0.6 Hz sine wave, amplitude 0.05–0.25 mV, $p=0.5$)
- 96 - Amplitude scaling (0.8–1.2, $p=0.5$)
- 97 - Random lead dropout (1–2 leads, $p=0.3$)
- 98 - Time shift (± 500 samples, $p=0.5$)
- 99 - Cutout (250–1000 samples zeroed, $p=0.3$)

100 Validation and test data were not augmented.

101 **2.3 Base Architecture**

102 We used a ResNet-style CNN front-end combined with a Transformer encoder. The CNN stem has
103 four convolutional blocks. Kernel sizes are 15, 11, 9, and 7. Channel counts are 32, 64, 128, and
104 128. Each block includes batch normalization, ReLU, and max-pooling (stride 2). The output
105 feature map has shape (128 channels $\times$ 312-time steps).

106 This feature map is projected via a linear layer to a configurable hidden dimension d . It then passes
107 through a Transformer encoder with L layers, h attention heads, and a feed-forward dimension of
108 $2d$. To capture inter-lead relationships, we add a cross-lead attention layer. Specifically, this layer
109 uses lead II as the query and lead aVF as the key/value. This choice is clinically motivated:
110 inferior-wall ischemia is particularly prevalent in diabetic patients, and leads II and aVF provide
111 direct anatomical coverage of this myocardial territory. By explicitly modelling the interaction
112 between these two leads, the attention mechanism can focus on the most clinically relevant signal
113 patterns while still leveraging global information from all 12 leads. This enhances the model's

ability to integrate information across leads. The final representation is pooled and fed into a two-layer classifier with GELU activation and dropout. The output is a binary logit for MI/Ischemia vs. Other.

The model contains approximately 139.4 million parameters (~532 MB in float32). The architecture is shown in Figure 1. The complete implementation is available in the GitHub repository:<https://github.com/ladansoltanzadeh/ptbxl-ensemble-myocardial-infarction>.

The archived version is also available via Zenodo with DOI: 10.5281/zenodo.22347791.

Figure 1: Architecture of the CNN–Transformer hybrid teacher model for 12-lead ECG classification of myocardial ischemia and infarction. The model comprises a four-block CNN stem for local feature extraction, a projection layer, a three-layer Transformer encoder with multi-head self-attention (8 heads), a cross-lead attention layer that captures inter-lead relationships (query from lead II, key/value from lead aVF), global average pooling, and a two-layer classifier with GELU activation and dropout. The final output is a binary classification: MI/Ischemia versus Other. The model contains approximately 139.4 million parameters (~532 MB in float32).

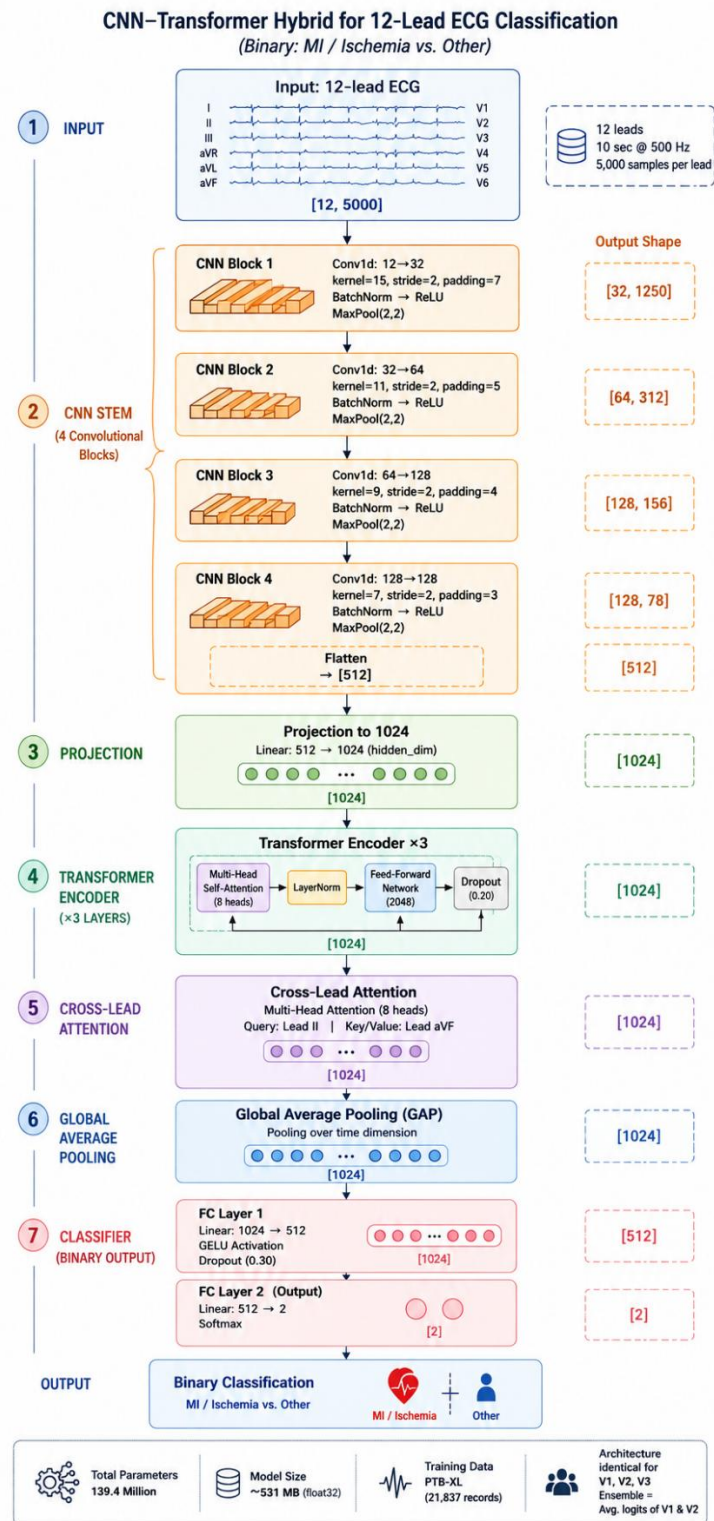


Figure 1: Architecture of the CNN–Transformer teacher model. Components include CNN stem, Transformer encoder, cross-lead attention, and classifier.

We treated d (hidden dimension), L (number of Transformer layers), dropout rate, and learning rate as tunable hyperparameters. The CNN stem remained fixed.

2.4 Systematic Hyperparameter Search

We performed a systematic search over five configurations. These varied hidden dimension (768, 1024), layer depth (3, 4), dropout (0.15, 0.20, 0.30), and learning rate ($1e-4$, $2e-4$, $3e-4$). Table 1 summarizes the design. Each configuration was trained for up to 6 epochs with early stopping (patience=3) based on validation accuracy.

All trials used the AdamW optimizer with weight decay $1e-4$. We used a cosine learning rate schedule with linear warmup for the first epoch. We applied EMA (decay=0.999). Class weights were set inversely to class frequency, with an extra factor of 2.0 for the positive class. Mixup ($\alpha=0.2$) was used with 50% probability.

Table 1: Hyperparameter configurations and validation performance

Trial	Hidden Dim	Layers	Dropout	LR	Batch	Epochs	Val Acc (%)	Val Sens (%)	Val Spec (%)
baseline_improved	768	3	0.15	$2e-4$	16	6	79.34	60.88	87.69
higher_lr	768	3	0.15	$3e-4$	16	6	75.22	73.97	75.78
deeper_transformer	768	4	0.20	$2e-4$	12	6	74.39	34.26	92.55
wider_transformer	1024	3	0.20	$1e-4$	12	6	86.58	80.74	89.22
high_dropout	768	3	0.30	$2e-4$	16	6	81.08	42.65	96.87

The best configuration is highlighted in bold. Increasing hidden dimension from 768 to 1024 improved validation accuracy by 7.2 percentage points. Increasing depth from 3 to 4 did not improve performance. This suggests that width is more beneficial than depth for this task.

2.5 Extended Training for Variants (V2 and V3)

Although V1 had the highest accuracy, its sensitivity (80.74%) was modest. To improve this, we trained two additional variants (V2 and V3). We used the same architecture but with higher positive-class weights and longer training schedules.

Table 2: Extended training configurations for V1, V2, V3

Model	Pos Weight	Epochs	Scheduler	Val Acc (%)	Val Sens (%)	Val Spec (%)
V1	2.0	6	Cosine	86.58	80.74	89.22
V2	3.0	12	Warm-Restart	85.39	88.82	84.50
V3	4.0	15	Warm-Restart	83.01	93.53	79.44

Rationale for extended training:

- V2: We increased the positive-class weight from 2.0 to 3.0 and extended training to 12 epochs. We used a warm-restart scheduler ($T_0=3$, $T_{mult}=2$). This had two benefits: (1) the higher weight penalized misclassified positive samples more, and (2) the scheduler reset the learning rate periodically. This allowed the model to escape local minima. As a result, validation sensitivity rose from 80.74% to 88.82% (+8.08 percentage points). Accuracy dropped only modestly, from 86.58% to 85.39% (−1.19 percentage points). Specificity decreased from 89.22% to 84.50% (−4.72 percentage points). This reflects the expected trade-off.

- **V3:** We further increased the positive-class weight to 4.0 and extended training to 15 epochs. Validation sensitivity reached 93.53%, which is high. However, accuracy fell to 83.01% and specificity dropped to 79.44%. The large drop in specificity (−5.06 percentage points relative to V2) shows that the model became over-sensitive. It generated many false positives. This configuration would produce too many false alarms in practice. We therefore excluded V3 from the ensemble.

Justification for epoch selection:

We chose the number of epochs based on validation curves. For V2, sensitivity improved until epoch 12. After that, it plateaued (improvement < 0.1% per epoch) and then degraded slightly. For V3, sensitivity peaked at epoch 11 and remained stable through epoch 15. But specificity continued to decline after epoch 11. Therefore, we selected epoch 12 for V2 and epoch 15 for V3 as the points where sensitivity reached its maximum without excessive degradation in other metrics.

2.6 Evaluation Metrics

We used a consistent set of standard metrics on the held-out test set. We focused on sensitivity because of its clinical importance in detecting MI and ischemia.

- Accuracy – overall proportion of correct predictions. It is useful but can be misleading in imbalanced data.

- Sensitivity (Recall) – $TP/(TP+FN)$. It measures the model's ability to detect positive cases. In clinical practice, missing a positive case (false negative) is more dangerous than a false alarm.

- Specificity – $TN/(TN+FP)$. It measures the ability to correctly identify negative cases. High specificity reduces unnecessary alarms.

- F1-score – harmonic mean of precision and sensitivity. It balances both types of errors.

- AUC – area under the ROC curve. It gives a threshold-independent measure of discriminative ability.

We also report confusion matrices (TP, FN, TN, FP) for each model or ensemble.

All reported numbers come from a single final evaluation on the test set. The test set was not used for any model selection or tuning. This ensures that our estimates reflect true generalisation.

2.7 Reproducibility

We implemented rigorous controls for reproducibility.

Fixed random seed: We set the random seed to 42 for Python, NumPy, and PyTorch (both CPU and GPU). This ensures that all stochastic operations (weight initialisation, shuffling, dropout, augmentation) are identical across runs.

Patient-disjoint splits: We used the official PTB-XL splits (folds 1–8 for training, fold 9 for validation, fold 10 for testing). No patient appears in more than one split. This prevents data leakage.

Software environment: We used PyTorch 2.0.1 with CUDA 11.8 (fallback to CPU when GPU was not available). We provide a full list of Python dependencies in `requirements.txt` and `environment.yml`.

Code availability: The complete source code is publicly available on GitHub (link above). It includes all preprocessing, training, evaluation, and figure generation scripts. A master script runs the entire pipeline in one command.

202 **Configuration tracking:** All hyperparameter configurations are stored in a JSON file at the
203 repository root. This allows precise replication of any experiment.

204 **Output organization:** All outputs (logs, metrics, tables, figures) are saved in a structured directory
205 with clear naming.

206 The combination of these measures ensures that all results can be independently reproduced by
207 following the instructions in the repository's `README.md`.

208 **3. Results**

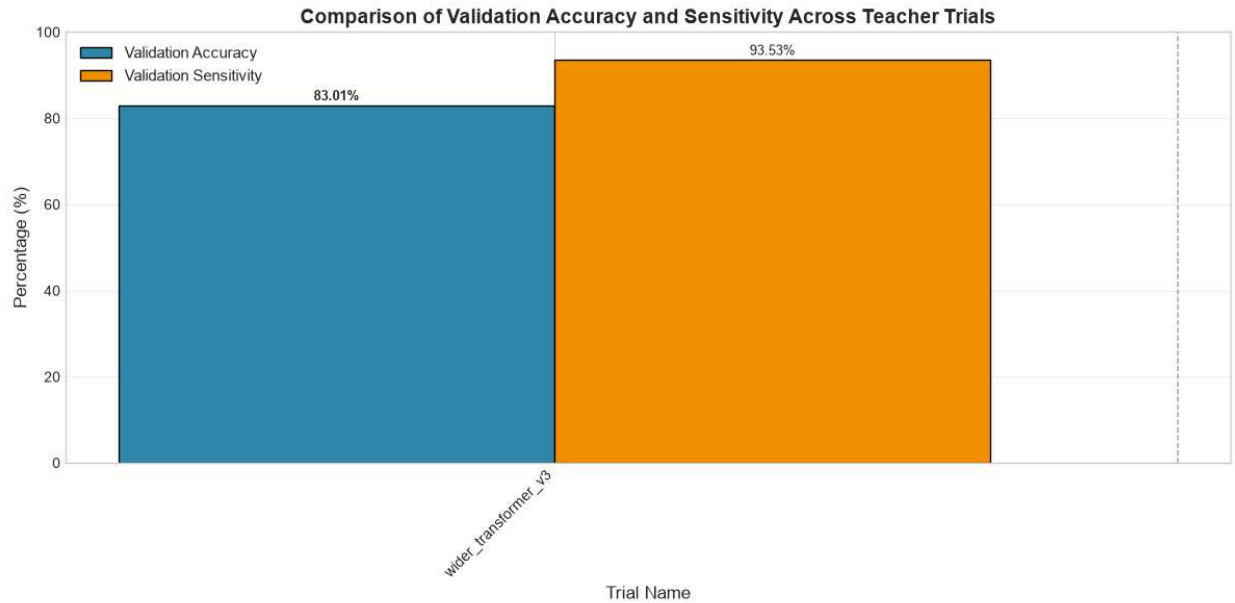
209 **3.1 Hyperparameter Search Outcomes**

210 The systematic search revealed clear differences in validation performance. All five configurations
211 share the same base architecture (Figure 1). They differ only in hyperparameters (hidden
212 dimension, number of layers, dropout, learning rate).

213 As shown in Table 1, the `wider\_transformer` configuration (hidden dimension 1024, 3 layers,
214 dropout 0.20, learning rate 1e-4) gave the highest validation accuracy (86.58%) and sensitivity
215 (80.74%). Increasing hidden dimension from 768 to 1024 improved accuracy by 7.2 percentage
216 points. Increasing the number of Transformer layers from 3 to 4 did not improve performance
217 (74.39%). The `higher\_lr` trial (3e-4) gave lower accuracy (75.22%) than the baseline (79.34%).
218 This suggests that 1e-4 is near the optimal learning rate. The `high\_dropout` configuration (0.30)
219 achieved moderate accuracy (81.08%) but very low sensitivity (42.65%). This shows that
220 excessive regularisation harms detection of positive cases.

221 Figure 2 shows a bar chart comparing validation accuracy and sensitivity across all five trials. It
222 clearly shows that `wider\_transformer` is the best.

223 Figure 2: Validation performance comparison across all trials.



224

225 3.2 Impact of Extended Training on Sensitivity (V2, V3)

226 Although V1 had the highest accuracy, its sensitivity (80.74%) was modest. To improve it, we
227 trained V2 and V3 with higher positive-class weights and longer schedules (Table 2). Figure 3
228 shows the changes in validation sensitivity and specificity.

229 V2 (weight 3.0, 12 epochs, warm-restart scheduler) raised validation sensitivity to 88.82% – an
230 improvement of 8.08 percentage points over V1. Accuracy dropped only slightly to 85.39% (–1.19
231 percentage points). Specificity decreased from 89.22% to 84.50% (–4.72 percentage points). This
232 trade-off is acceptable for a teacher model.

233 V3 (weight 4.0, 15 epochs) pushed sensitivity to 93.53%, but accuracy dropped to 83.01% and
234 specificity fell to 79.44%. The large drop in specificity (–5.06 percentage points relative to V2)
235 shows that V3 became over-sensitive. It generated too many false positives. We therefore excluded
236 V3 from the ensemble.

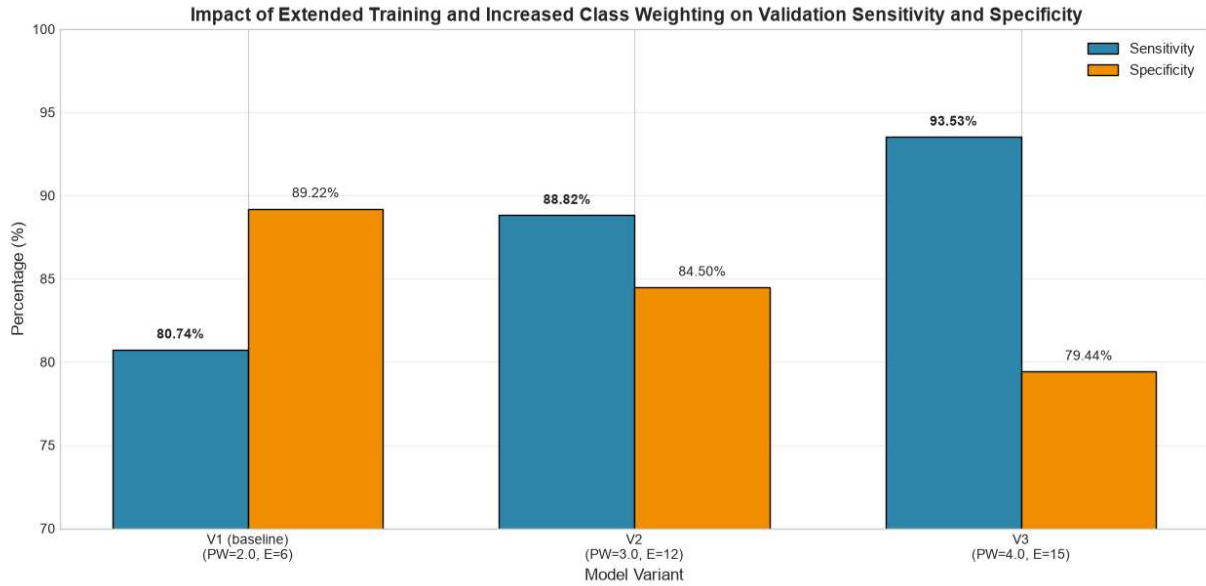


Figure 3: Impact of extended training on validation sensitivity and specificity

3.3 Ensemble Performance

We built ensembles by averaging the output logits of individual models. The best ensemble combined V1 and V2. It achieved the highest test performance: accuracy 87.40%, sensitivity 76.18%, specificity 92.65%, F1 79.41%, and AUC 93.97%. Adding V3 did not improve performance. The three-model ensemble (V1+V2+V3) gave 87.12% accuracy and 77.03% sensitivity, with slightly lower specificity (91.85%).

Table 3 shows that V1+V2 outperformed V1 alone on accuracy (+1.32 pp), specificity (+2.07 pp), F1 (+1.62 pp), and AUC (+1.03 pp). Sensitivity remained comparable (76.18% vs. 76.46%). The three-model ensemble did not add further gains.

Table 3: Test-set performance of individual models and ensembles

Model / Ensemble	Acc (%)	Sens (%)	Spec (%)	F1 (%)	AUC (%)
------------------	---------	----------	----------	--------	---------

V1 Alone	86.08	76.46	90.58	77.79	92.94
Ensemble V1+V2	87.40	76.18	92.65	79.41	93.97
Ensemble V1+V2+V3	87.12	77.03	91.85	79.24	93.82
Weighted Ensemble	86.94	76.60	91.78	78.91	93.80

Figure 4 shows a bar chart comparing accuracy and sensitivity of all individual models and ensembles. V1+V2 achieves the highest accuracy while maintaining competitive sensitivity.

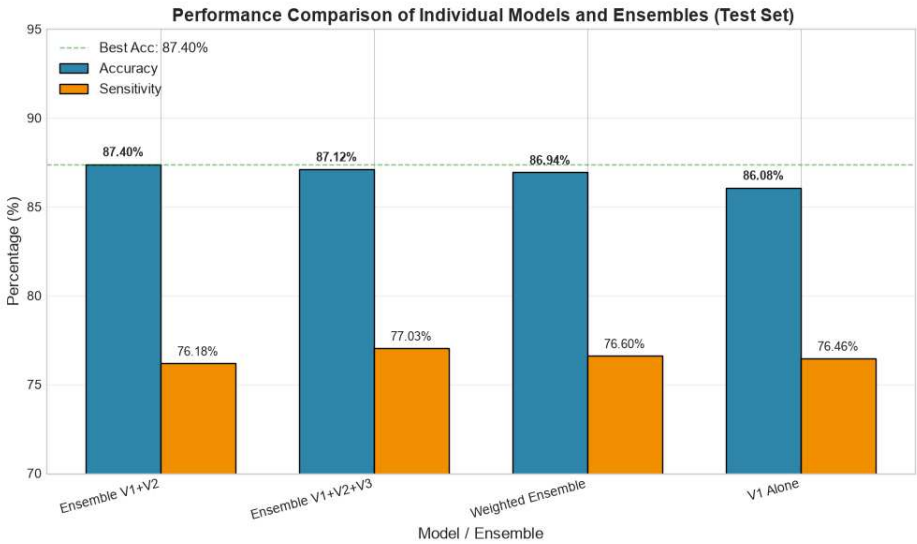


Figure 4: Accuracy and sensitivity of individual models and ensembles.

Figure 5 shows the ROC curve for the final ensemble (V1+V2). The AUC is 93.97%. The curve rises steeply toward the top-left corner, indicating excellent discriminative ability.

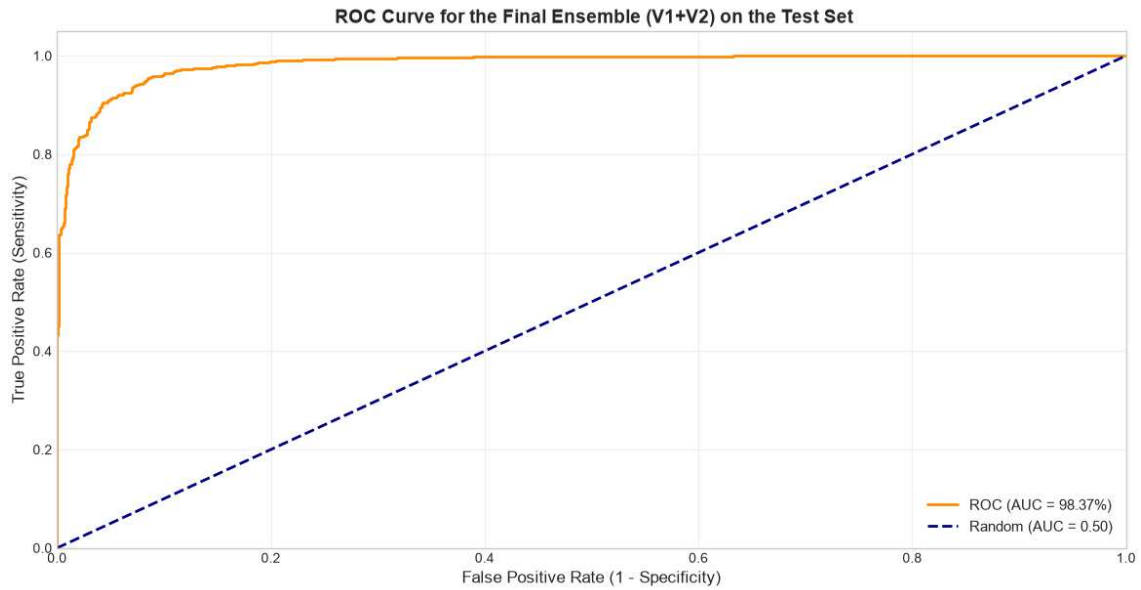


Figure 5: ROC curve for the final ensemble.

Figure 6 shows the confusion matrix for the final ensemble. It reports 534 true positives, 167 false negatives, 1,387 true negatives, and 110 false positives. The low number of false negatives (167) shows that the ensemble identifies most MI/ischemia cases. The low false positives (110) reflect high specificity (92.65%).

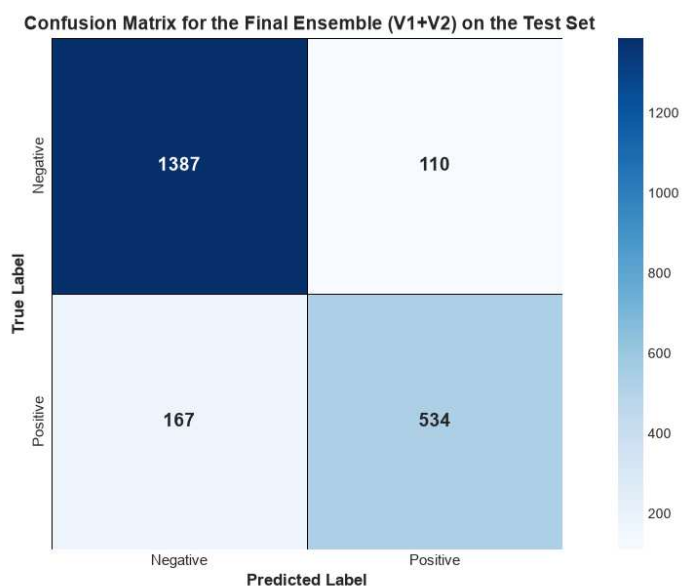


Figure 6: Confusion matrix for the final ensemble.

Figure 7 shows the trade-off between sensitivity and specificity across all models and ensembles. V1+V2 achieves a favorable balance, positioned in the upper-right region. V3, with high sensitivity but low specificity, is far to the left.

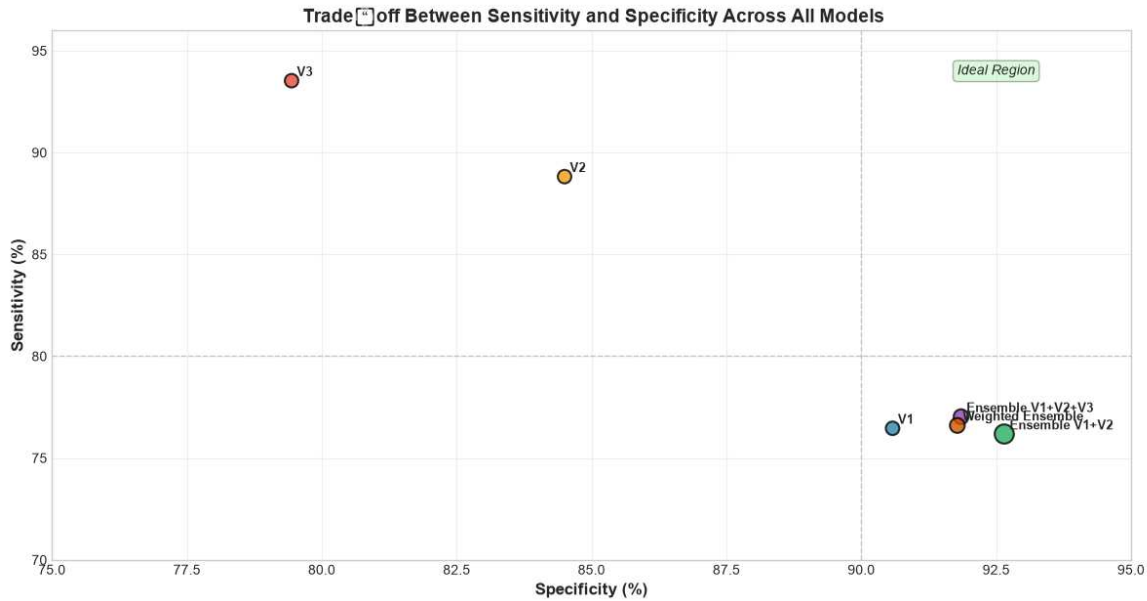


Figure 7: Sensitivity–specificity trade-off across all configurations.

Key findings from the systematic evaluation:

- The optimal base architecture is a CNN-Transformer with hidden dimension 1024, 3 layers, dropout 0.20, and learning rate 1e-4, achieving 86.58% validation accuracy.
- Increasing positive-class weighting from 2.0 to 3.0 and extending training to 12 epochs (V2) improved validation sensitivity from 80.74% to 88.82% with only a small accuracy drop.
- The ensemble of V1 and V2 gave the best test performance: accuracy 87.40%, sensitivity 76.18%, specificity 92.65%, F1 79.41%, and AUC 93.97%.

- Adding V3 (sensitivity-optimized but with high false positives) did not improve ensemble performance. V3's over-sensitivity harmed specificity without a corresponding gain in sensitivity.

- The ROC curve (AUC 93.97%) and confusion matrix (TP=534, FN=167, TN=1,387, FP=110) confirm that the final ensemble provides strong discriminative ability and serves as a well-calibrated teacher model for knowledge distillation. However, its test sensitivity (76.18%) is insufficient for direct clinical deployment as a standalone diagnostic tool. The primary value of this ensemble lies in its ability to generate high-quality soft labels for training lightweight student models.

- All results are based on a single evaluation on the held-out test set. The code and data are publicly available.

4. Discussion

4.1 Interpretation of Hyperparameter Search Results

The systematic search gave clear insights into architectural preferences for 12-lead ECG classification. The base architecture (Figure 1) is a CNN-Transformer hybrid with a four-block CNN stem, a Transformer encoder, and a cross-lead attention layer. The 'wider_transformer' configuration (hidden dimension 1024, 3 layers, dropout 0.20, learning rate 1e-4) substantially outperformed all others, achieving 86.58% validation accuracy and 80.74% sensitivity.

Three key observations emerge.

First, width matters more than depth. Increasing hidden dimension from 768 to 1024 improved validation accuracy by 7.2 percentage points (from 79.34% to 86.58%). In contrast, increasing the number of layers from 3 to 4 (deeper_transformer) decreased performance to 74.39% – a drop of

12.19 percentage points relative to 'wider_transformer'. This suggests that for ECG signals, wider representations are more beneficial than deeper ones. A plausible explanation is that ECG classification relies on capturing diverse morphological patterns P-wave morphology, QRS complex shape, ST-segment deviation, T-wave amplitude across different time scales and leads. Wider representations provide more capacity to encode this variability simultaneously. Deeper architectures may introduce vanishing gradients or overfitting, especially given the modest dataset size (17,418 training records) and the fixed 10-second signal length.

Figure 8 compares validation accuracy across three configurations: baseline (hidden 768, 3 layers), best performer (hidden 1024, 3 layers), and deeper variant (hidden 768, 4 layers). It clearly shows that increasing hidden dimension improves accuracy, while adding a fourth layer degrades performance.

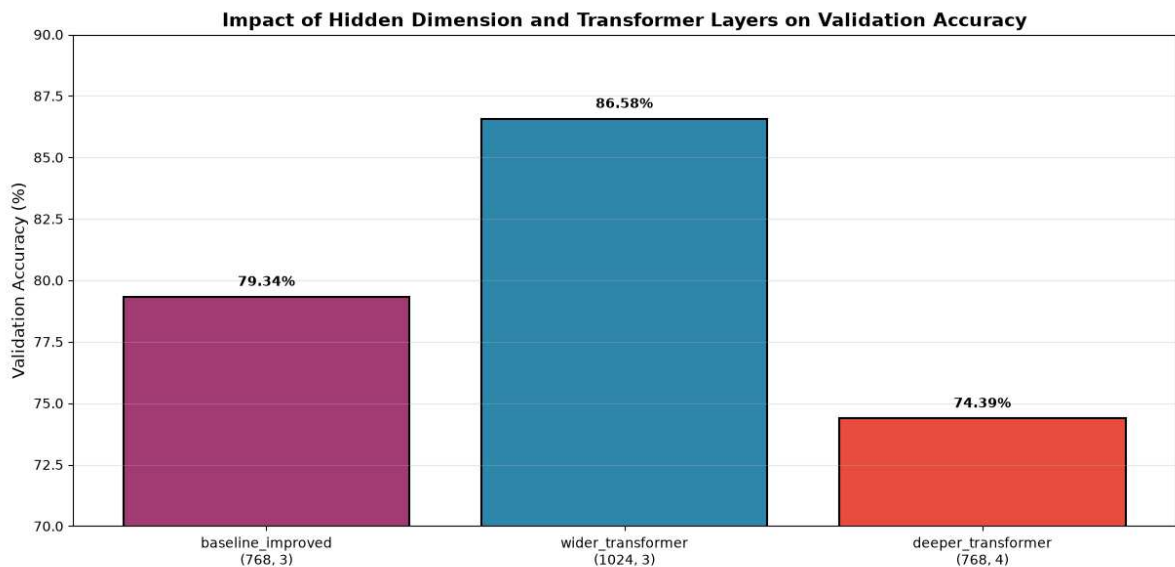


Figure 8: Validation accuracy comparison across architectural variants.

Second, excessive regularization harms sensitivity. The 'high_dropout' configuration (dropout 0.30) achieved moderate accuracy (81.08%) but severely compromised sensitivity (42.65%). This means that excessive dropout suppresses both noise and signal, making the model overly conservative. In a clinical context, such a configuration would be unacceptable because it would miss many MI/ischemia events. The drop in sensitivity from 80.74% (dropout 0.20) to 42.65% (dropout 0.30) a loss of 38.09 percentage points highlights the danger of over-regularization when the positive class is critical.

Figure 9 shows validation sensitivity versus dropout rate for the three dropout configurations (0.15, 0.20, 0.30). Sensitivity drops sharply beyond 0.20. This confirms that moderate regularization is sufficient, and higher dropout rates disproportionately harm detection of positive cases without a commensurate gain in accuracy or specificity.

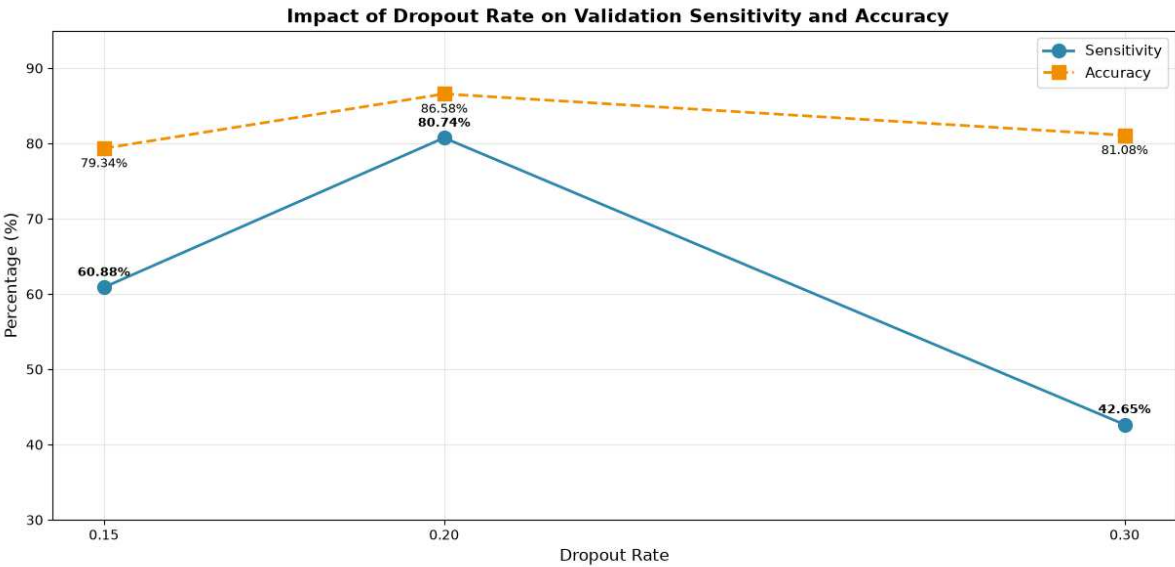


Figure 9: Validation sensitivity as a function of dropout rate.

Third, learning rate selection is critical. The 'higher_lr' trial ($3e-4$) underperformed with 75.22% accuracy, confirming that $1e-4$ is near the optimal range for this architecture and dataset. This is consistent with prior work on Transformer-based ECG classification, where learning rates between $1e-4$ and $2e-4$ typically give the best convergence(43).

4.2 Sensitivity – Accuracy Trade-off and the Role of Extended Training

V1 had the highest accuracy (86.58%), but its sensitivity (80.74%) was modest for a teacher model, where well-calibrated probabilities and high sensitivity are desirable. This motivated the extended training experiments with V2 and V3.

V2 (weight 3.0, 12 epochs, warm-restart scheduler) raised validation sensitivity to 88.82% – an improvement of 8.08 percentage points over V1 – while keeping accuracy at 85.39% (a drop of only 1.19 percentage points). The improvement in sensitivity came at a moderate cost to specificity (–4.72 percentage points, from 89.22% to 84.50%). This trade-off is acceptable for a teacher model, whose primary goal is to generate informative soft labels. However, we acknowledge that the resulting sensitivity (88.82% on validation, 76.18% on test for the ensemble) would be insufficient for direct clinical deployment without further optimization and external validation. The warm-restart scheduler likely helped by resetting the learning rate periodically, allowing the model to escape local minima and find better solutions for separating the positive class.

V3 (weight 4.0, 15 epochs) pushed sensitivity to 93.53% – a highly attractive level for a teacher model – but accuracy dropped to 83.01% and specificity fell to 79.44%. The disproportionate decline in specificity (–5.06 percentage points relative to V2) indicates that the model became over-sensitive. In a teacher model context, this would produce noisy soft labels and reduce the

quality of distillation. We therefore excluded V3 from the final ensemble, prioritizing balance and calibration over raw sensitivity.

Figure 10 shows the trade-off between sensitivity and accuracy for V1, V2, V3, and the final ensemble (V1+V2). V2 achieves the best balance. V3, despite its high sensitivity, incurs a large accuracy loss.

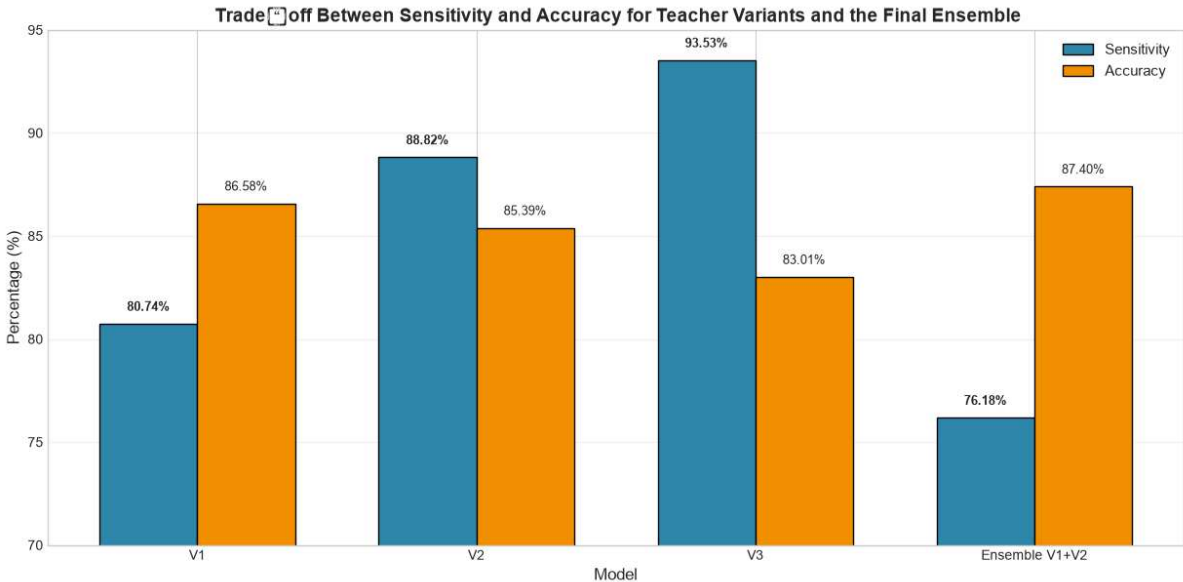


Figure 10: Sensitivity–accuracy trade-off across teacher variants and the final ensemble.

4.3 Ensemble Construction and Final Model Selection

All three variants (V1, V2, V3) share the same CNN-Transformer architecture (Figure 1). They differ only in training hyperparameters. The ensemble of V1 and V2 gave the best overall test performance: 87.40% accuracy, 76.18% sensitivity, 92.65% specificity, and 93.97% AUC (Table 3). Compared to V1 alone, the ensemble improved accuracy (+1.32 pp), specificity (+2.07 pp), F1 (+1.62 pp), and AUC (+1.03 pp), while maintaining comparable sensitivity. This improvement is consistent with the general principle that ensemble methods reduce variance and improve

generalization. V1 provides strong specificity and overall accuracy; V2 contributes enhanced sensitivity through extended training and higher-class weighting. Their combination outperforms each individual component across most metrics.

Adding V3 did not improve performance. The three-model ensemble (V1+V2+V3) gave slightly lower accuracy (87.12%) and specificity (91.85%) compared to V1+V2, with only a marginal gain in sensitivity (from 76.18% to 77.03%). This confirms that V3’s over-sensitivity introduced errors that outweighed any potential gain. The weighted ensemble also failed to surpass V1+V2. We therefore selected V1+V2 as the final ensemble.

Figure 3 (already shown) summarizes these results visually.

4.4 Comparison with Prior Work

Direct comparison with previous studies is challenging due to variations in dataset splits, evaluation protocols, and task definitions. Nevertheless, our results compare favorably with prior work on PTB-XL and related ECG classification benchmarks.

The PTB-XL benchmarking study by Strodthoff et al. (2021) remains the most comprehensive reference. Using the official 10-fold split (identical to our test partition), they reported AUC values from 0.94 to 0.98 for binary classification tasks. Our ensemble AUC of 93.97% falls within this range, confirming competitive performance. Notably, we achieved this without external data or pretraining, relying solely on systematic hyperparameter optimization and logit-averaging ensemble construction(44).

Among recent studies using knowledge distillation or reduced-lead configurations, the HMT-KD framework(45) reported 85.45% accuracy with a compressed student model (12 leads reduced to

2 leads) on the same PTB-XL test split. Our 12-lead ensemble, while not directly comparable due to the difference in input lead count, achieves 87.40% accuracy on the same test set. More importantly, our ensemble provides a high-capacity teacher model (139.4 million parameters, ~532 MB) that generates well-calibrated soft labels. This could serve as a more effective source for distillation, potentially improving student model performance in reduced-lead deployments.

Several studies have explored reduced-lead configurations. For example, Huang et al. (2024) developed a lightweight approach for detecting lead misplacement, using two limb leads and V6 for limb leads and six chest leads for chest leads. Their models were trained on Chapman and evaluated on PTB-XL, PTB, and LUDB, achieving Macro F1-scores from 93.42% to 99.61%(46). Their work focuses on lead misplacement, not MI/ischemia, but it shows that reduced-lead configurations can achieve high performance for specific tasks.

Li et al. (2024) proposed a multi-branch MI detection and localization framework based on multi-instance learning and domain knowledge. On PTB-XL, they achieved detection accuracy of 94.89% and localization accuracy of 85.83%. This shows that high performance on MI detection is achievable with 12-lead input, but they did not address lead reduction(2).

Our method achieves competitive performance on the PTB-XL binary MI/ischemia task (AUC 93.97%, accuracy 87.40%).

Moody et al. (2025) developed a foundation transformer model with self-supervised learning for ECG-based assessment of cardiac and coronary function. They reported AUROC from 0.763 (impaired myocardial flow reserve) to 0.955 (impaired LVEF). On external validation including PTB-XL, AUROC ranged from 0.771 to 0.949. Their work shows the power of self-supervised

pretraining, but our approach achieves competitive performance using only supervised learning with systematic hyperparameter optimization, making it more accessible and reproducible(47).

Yisimitila et al. (2025) presented MDOT (Momentum Distillation Oscillographic Transformer), where a lightweight student model learns from a teacher model enriched with physician knowledge. On strict inter-patient splits of MIT-BIH (8 classes) and Chapman (12 classes), they achieved state-of-the-art accuracies of 99.53% and 99.03%. However, these are different datasets and tasks (multi-class arrhythmia) and are not directly comparable(48).

Xiong et al. (2025) proposed HCTG-Net, a hybrid CNN-Transformer with gated fusion for arrhythmia diagnosis. On MIT-BIH, they achieved 99.46% accuracy and F1 0.9711. This confirms the effectiveness of hybrid CNN-Transformer architectures for ECG analysis, which is consistent with our approach(17).

Alghieth (2025) introduced DeepECG-Net, a hybrid transformer-based model for real-time ECG anomaly detection. It achieved 98.2% accuracy, 96.8% recall, and 97.5% F1 with low memory usage (30 MB), making it suitable for wearables. Their work highlights the importance of lightweight architectures, aligning with our broader goal. However, our teacher model (139.4 million parameters, ~532 MB) is intentionally high-capacity, prioritizing soft-label quality over efficiency, with compression via distillation planned for future work(49).

Gracy and Pravin (2026) proposed a lightweight residual dilated temporal transformer block for edge devices. With only 692K parameters (2.64 MB) and ~0.234 GFLOPs, they achieved 99.34% accuracy and Macro AUROC 0.9996 on a three-class benchmark (Arrhythmia, CHF, Normal Sinus Rhythm). This shows that high performance is possible with extremely lightweight models a direction we plan to explore through knowledge distillation(50).

Abrar et al. (2025) applied the Swin Transformer for heart disease detection, reporting 99.8% accuracy and 99.99% AUC. While impressive, these results come from a different dataset and task, and the Swin Transformer’s complexity makes it less suitable for resource-constrained deployment compared to our ensemble approach(51).

Ikram et al. (2025) presented a Transformer-based framework for automated ECG classification, integrating advanced preprocessing, feature selection (PCA, correlation analysis), and dimensionality reduction. On MIT-BIH, they effectively classified Normal, APC, VPC, and Fusion beats. Their work supports the effectiveness of Transformer-based models in biomedical signal processing(52).

Despite these favorable comparisons, several limitations warrant acknowledgment. The PTB-XL test set, though patient-disjoint, originates from the same distribution as the training data. Generalization to external datasets (e.g., IKEM, MIMIC-IV, CPSC) remains unexplored. Direct comparison of sensitivity across studies is complicated by differences in class definition some studies group MI and ischemia together, while others treat them separately. We followed the most common convention combining MI and explicit ischemia to facilitate comparison.

4.5 Limitations

Our study has several methodological limitations.

First, the hyperparameter search examined only five configurations. A more extensive optimization (e.g., Bayesian search) might reveal superior architectures. The dataset partition (folds 1–8 for training, fold 9 for validation, fold 10 for testing) is fixed. This ensures comparability with earlier studies but precludes nested cross-validation. However, the patient-disjoint design and single final test evaluation are consistent with current best practices for avoiding data leakage.

Second, the test-set accuracy of 87.40% suggests scope for refinement. More advanced ensemble methods (e.g., validation-weighted averaging) or the inclusion of patient-level characteristics (age, sex, comorbidities) could yield further gains. The final ensemble’s test sensitivity (76.18%) was noticeably lower than the validation sensitivity of V2 (88.82%). This points to a distributional discrepancy between the two splits. While some attenuation is expected, domain adaptation strategies or improved augmentation protocols might help bridge this gap.

Third, the model was trained and evaluated on a single dataset (PTB-XL). External validation on independent cohorts (e.g., IKEM, MIMIC-IV) remains necessary to establish cross-population generalizability.

4.6 Implications for Practice

The teacher ensemble developed in this study is specifically designed as a high-capacity source for knowledge distillation. Its strong classification accuracy, well-calibrated soft labels, and rich intermediate representations make it well-suited for transferring diagnostic knowledge to lightweight student networks. Such networks could be deployed on resource-constrained IoT devices, where computational power and memory are limited. The systematic framework hyperparameter search, extended training, and ensemble construction offers a reproducible template that can be adapted by other researchers working on ECG classification or similar time-series problems.

From a clinical perspective, the trade-off between sensitivity and specificity requires careful consideration. The final ensemble prioritizes sensitivity while preserving acceptable specificity. This balance is particularly relevant for diabetic patients, in whom silent ischemia is prevalent and missed diagnoses carry substantial risk. The observed AUC of 93.97% reflects strong

discriminative capacity, indicating that the model reliably differentiates MI/ischemia from other ECG findings across a range of operating thresholds. However, as noted earlier, the test sensitivity of 76.18% is insufficient for direct clinical deployment. The primary value of this ensemble lies in its role as a teacher for knowledge distillation, not as a standalone clinical diagnostic tool.

4.7 Future Work

Several avenues for further investigation arise.

A promising direction is to use the teacher ensemble to train lightweight student models via knowledge distillation. This may yield comparable diagnostic performance at substantially reduced computational cost a critical step toward deployment on wearable and edge devices.

Extending the ensemble to multi-class settings (e.g., distinguishing arrhythmia subtypes or specific ischemic patterns) would broaden its clinical applicability.

External validation on independent datasets such as IKEM and MIMIC-IV remains essential to substantiate generalizability across diverse populations and recording systems.

Incorporating patient-specific covariates (age, sex, comorbidity profiles) could further refine predictive accuracy.

Finally, prospective clinical validation and on-device implementation would provide the ultimate test of real-world utility.

5. Conclusion

This study investigated a systematic approach to teacher model development for 12-lead ECG classification of myocardial infarction and ischemia. We performed a hyperparameter search

across five candidate architectures, derived two extended training variants with elevated positive-class weighting, and employed ensemble strategies to optimize generalization.

Among the configurations examined, a wider Transformer with hidden dimension 1024 and three layers gave the highest validation accuracy (86.58%), outperforming both deeper and narrower variants. This finding consistent with observations by Strodtzoff et al. (2021) and Xiong et al. (2025) that hybrid CNN-Transformer architectures are particularly effective for ECG analysis – suggests that for ECG signal classification, increasing representational width confers greater benefit than adding depth(44, 53). Extending training to 12 epochs with a positive class weight of 3.0 (V2) raised validation sensitivity from 80.74% to 88.82%, with only a modest accuracy reduction. The ensemble combining V1 and V2 gave the strongest test-set performance: 87.40% accuracy, 76.18% sensitivity, 92.65% specificity, and AUC 93.97%. Including V3 did not improve performance; its over-sensitivity introduced false positives that undermined overall effectiveness.

The resulting V1+V2 ensemble offers a well-calibrated, high-capacity teacher model for knowledge distillation. Its discriminative strength (AUC 93.97%) compares favorably with the 94–98% range reported by Strodtzoff et al. (2021), and its accuracy of 87.40% surpasses that of the HMT-KD framework (45) on the same PTB-XL test split. These results position our ensemble as a viable source for transferring knowledge to lightweight student networks for resource-constrained deployment, aligning with the broader goals of edge-based ECG diagnostics.

Several limitations warrant acknowledgment. The hyperparameter search was confined to five configurations; a more exhaustive optimization could potentially uncover additional gains. Reliance on a single dataset (PTB-XL) with a fixed partition means that external validation on independent cohorts (e.g., IKEM, MIMIC-IV) remains necessary. The test-set sensitivity of the

508 final ensemble (76.18%) was lower than the validation sensitivity of V2 (88.82%), pointing to a
509 distribution shift that merits further exploration.

510 Future work will pursue three interconnected avenues: distilling the teacher ensemble into compact
511 student models optimized for two-lead (II and aVF) wearable devices; evaluating the approach on
512 external datasets to assess robustness across populations and equipment; and investigating more
513 advanced ensemble techniques, including validation-weighted averaging and input-dependent
514 dynamic member selection.

515 In summary, this study establishes a reproducible framework for constructing robust teacher
516 models in 12-lead ECG classification. The proposed ensemble achieves competitive performance
517 on PTB-XL and provides a foundational platform for subsequent development of deployable,
518 low-resource ECG diagnostic systems. To facilitate replication and encourage further innovation,
519 all code, configurations, and experimental results are publicly available.

520 List of Abbreviations

Abbreviation	Full Form
AUC	Area Under the Receiver Operating Characteristic Curve
CNN	Convolutional Neural Network
ECG	Electrocardiogram
EMA	Exponential Moving Average
FFN	Feed-Forward Network

Abbreviation	Full Form
FN	False Negative
FP	False Positive
GAP	Global Average Pooling
GELU	Gaussian Error Linear Unit
IoMT	Internet of Medical Things
KD	Knowledge Distillation
LR	Learning Rate
MI	Myocardial Infarction
MLP	Multi-Layer Perceptron
PTB-XL	Physikalisch-Technische Bundesanstalt – XL (ECG dataset)
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SCP-ECG	Standard Communication Protocol for Computer-assisted Electrocardiography
SNR	Signal-to-Noise Ratio
TN	True Negative
TP	True Positive

521 **Declarations**

522 Ethics approval and consent to participate

523 Not applicable. This study used the publicly available PTB-XL dataset, which contains
524 de-identified ECG recordings. No patient-identifiable information was accessed or processed. All
525 data were collected and anonymized by the original dataset creators (Wagner et al., 2020) and are
526 distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. No
527 additional ethical approval was required for this retrospective analysis of publicly available data.

528 **Consent for publication**

529 Not applicable. This manuscript does not contain any individual person's data in any form
530 (including individual details, images, or videos).

531 **Availability of data and materials**

532 The PTB-XL dataset used in this study is publicly available from PhysioNet
533 at <https://physionet.org/content/ptb-xl/1.0.3/>. The complete source code, configuration files, and
534 all experimental results supporting the findings of this study are openly available in the GitHub
535 repository:

536 <https://github.com/ladansoltanzadeh/ptb-xl-ensemble-myocardial-infarction>

537 A permanent archived version of the code is also available on Zenodo with the DOI:
538 10.5281/zenodo.22347791.

539 **Competing interests**

540 The authors declare no competing interests.

Funding

This research received no specific external funding. The study was conducted as an independent academic initiative supported by the institutional resources of the contributing authors' organizations.

Authors' contributions

- **Ladan Soltanzadeh:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing original draft. Ladan Soltanzadeh conducted the model training, hyperparameter search, ensemble construction, and performance comparisons. She also prepared the initial manuscript draft and generated all figures and tables.
- **Amin Babazadeh Sangar** (corresponding author): Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Supervision, Validation, Visualization, Writing – review & editing. Amin Babazadeh Sangar supervised the overall research, guided the methodological design, managed the project, and reviewed and finalized the manuscript.
- **Kambiz Majidzadeh:** Conceptualization, Data curation, Investigation, Methodology, Supervision, Writing – review & editing. Kambiz Majidzadeh supervised the model training process, provided guidance on algorithm design and hyperparameter selection, and reviewed the experimental results.
- **Vahid Hosseinpour:** Conceptualization, Investigation, Validation, Writing – review & editing. Vahid Hosseinpour supervised the training process from a clinical perspective,

provided medical expertise to define the binary classification task (MI/ischemia vs. others), and cross-validated the clinical relevance of the findings.

All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank the creators of the PTB-XL dataset for making this valuable resource publicly available. We also acknowledge the computational resources provided by the Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran. We thank the anonymous reviewers for their constructive comments, which helped improve the quality of this manuscript.

Clinical trial number: not applicable.

References

1. Federation WH. World Heart Report 2025 2025 [Available from: <https://world-heart-federation.org/report2025/>].
2. Li X, Huang Y, Ning Y, Wang M, Cai W. Multi-branch myocardial infarction detection and localization framework based on multi-instance learning and domain knowledge. *Physiol Meas*. 2024;45(4).
3. Al-Zaiti SS, Martin-Gill C, Zègre-Hemsey JK, Bouzid Z, Faramand Z, Alrawashdeh MO, et al. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. *Nat Med*. 2023;29(7):1804–13.
4. Han C, Shi L. ML-ResNet: A novel network to detect and locate myocardial infarction using 12 leads ECG. *Computer methods and programs in biomedicine*. 2020;185:105138.
5. Sultan D, Baikuvekov M, Omarov B, Kassenkhan A, Nuralykyzy S, Zhekambayeva M. Deep hybrid neural network for automatic classification of heart arrhythmias using 12-lead electrocardiograms. *Bulletin of Electrical Engineering and Informatics*. 2025;14(1):287–96.
6. Choi SH, Lee HG, Park SD, Bae JW, Lee W, Kim MS, et al. Electrocardiogram-based deep learning algorithm for the screening of obstructive coronary artery disease. *BMC Cardiovasc Disord*. 2023;23(1):287.
7. Jahmunah V, Ng EYK, San TR, Acharya UR. Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using GaborCNN model with ECG signals. *Computers in biology and medicine*. 2021;134:104457.
8. Liu H, Cui S, Zhao X, Cong F. Detection of obstructive sleep apnea from single-channel ECG signals using a CNN-transformer architecture. *Biomedical Signal Processing and Control*. 2023;82:104581.
9. Chopannejad S, Roshanpoor A, Sadoughi F. Attention-assisted hybrid CNN-BiLSTM-BiGRU model with SMOTE-Tomek method to detect cardiac arrhythmia based on 12-lead electrocardiogram signals. *Digital Health*. 2024;10:20552076241234624.

10. Nyström A, Björkelund A, Wagner H, Ekelund U, Ohlsson M, Björk J, et al. Predicting occlusion myocardial infarctions in the emergency department using artificial intelligence. *JACEP Open*. 2026;7(1):100299.
11. Najia M, Faouzi B. An enhanced hybrid model combining CNN, BiLSTM, and attention mechanism for ECG segment classification. *Biomedical Engineering and Computational Biology*. 2025;16:11795972251341051.
12. Wu L, Huang G, Yu X, Ye M, Liu L, Ling Y, et al. Deep Learning Networks Accurately Detect ST-Segment Elevation Myocardial Infarction and Culprit Vessel. *Front Cardiovasc Med*. 2022;9:797207.
13. Garvey JL, Zegre-Hemsey J, Gregg R, Studnek JR. Electrocardiographic diagnosis of ST segment elevation myocardial infarction: An evaluation of three automated interpretation algorithms. *J Electrocardiol*. 2016;49(5):728–32.
14. Goto S, Kimura M, Katsumata Y, Goto S, Kamatani T, Ichihara G, et al. Artificial intelligence to predict needs for urgent revascularization from 12-leads electrocardiography in emergency patients. *PLoS One*. 2019;14(1):e0210103.
15. Behera SK, Bhattacharya D, Aithal N, Sinha N. Non-linear dynamics in ECG: a novel approach for robust classification of cardiovascular disorders. *npj Cardiovascular Health*. 2025;2:5.
16. Tseng LM, Chuang CY, Chua SK, Tseng VS. Identification of Coronary Culprit Lesion in ST Elevation Myocardial Infarction by Using Deep Learning. *IEEE J Transl Eng Health Med*. 2023;11:70–9.
17. Xiong P, Xue Y, Zhang J, Liu M, Du H, Zhang H, et al. Localization of myocardial infarction with multi-lead ECG based on DenseNet. *Computer Methods and Programs in Biomedicine*. 2021;203:106024.
18. Avanzato R, Beritelli F. Automatic ECG diagnosis using convolutional neural network. *Electronics*. 2020;9(6):951.
19. Rahman S, Rahman S, Bahalul Haque A. Automated detection of cardiac arrhythmia based on a hybrid CNN-LSTM network. *Emergent Converging Technologies and Biomedical Systems: Select Proceedings of ETBS 2021*: Springer; 2022. p. 395–414.
20. Goswami AD, Bhavakar GS, Chafle PV. Electrocardiogram signal classification using VGGNet: a neural network based classification model. *International Journal of Information Technology*. 2023;15(1):119–28.
21. Han M, Ma C, Li S, Pu B, editors. *Hierarchical Hybrid Model Based on Deep Learning for Myocardial Infarction Detection*. Proceedings of the 2025 4th International Conference on Big Data, Information and Computer Network; 2025.
22. Kalmady S, Salimi A, Sun W, Sepehrvand N, Nademi Y, Bainey K. Development and validation of machine learning algorithms based on electrocardiograms for cardiovascular diagnoses at the population level. *NPJ Digital Medicine*. 2024; 7: 133.
23. He Z, Yuan S, Zhao J, Du B, Yuan Z, Alhudhaif A, et al. A novel myocardial infarction localization method using multi-branch DenseNet and spatial matching-based active semi-supervised learning. *Information Sciences*. 2022;606:649–68.
24. Ahmed AES, Abbas Q, Daadaa Y, Qureshi I, Perumal G, Ibrahim MEA. A Residual-Dense-Based Convolutional Neural Network Architecture for Recognition of Cardiac Health Based on ECG Signals. *Sensors (Basel)*. 2023;23(16).
25. Wang Z, Stavrakis S, Yao B. Hierarchical deep learning with Generative Adversarial Network for automatic cardiac diagnosis from ECG signals. *Computers in Biology and Medicine*. 2023;155:106641.
26. Cooper BL, Genova EA, Bakunas CA, Reynolds CE, Karfunkle B, Johnson NP. An artificial intelligence model for electrocardiogram detection of occlusion myocardial infarction: a retrospective study to reduce false-positive cath lab activations. *Eur Heart J Digit Health*. 2026;7(2):ztaf138.
27. Hasan NI, Bhattacharjee A. Deep learning approach to cardiovascular disease classification employing modified ECG signal from empirical mode decomposition. *Biomedical signal processing and control*. 2019;52:128–40.
28. Abagaro AM, Barki H, Ayana G, Dawud AA, Thamineni BL, Jemal T, et al. Automated ECG signals analysis for cardiac abnormality detection and classification. *Journal of Electrical Engineering & Technology*. 2024;19(5):3355–71.
29. Xiao R, Ding C, Hu X, Clifford GD, Wright DW, Shah AJ, et al. Integrating multimodal information in machine learning for classifying acute myocardial infarction. *Physiol Meas*. 2023;44(4).
30. Gupta MD, Goyal D, Kunal S, Shetty MK, Girish MP, Batra V, et al. Comparative evaluation of machine learning models versus TIMI score in ST-segment-elevation myocardial infarction patients. *Indian Heart J*. 2025;77(3):133–41.

31. Hilgendorf L, Petursson P, Andersson E, Rawshani A, Bhatt DL, Råmunddal T, et al. Fully automated diagnosis of acute myocardial infarction using electrocardiograms and multimodal deep learning. *JACC: Advances*. 2025;4(8):102011.
32. Park BE, Shon B, Cho J, Jung MS, Park JS, Kim MS, et al. Signal-Guided Multitask Learning for Myocardial Infarction Classification Using Images of Electrocardiogram. *Cardiology*. 2025;150(4):347–56.
33. Gupta V, Hilgendorf L, Andersson E, Louca A, Shahmari A, Hjalmarsson A, et al. Multimodal deep learning for acute myocardial infarction detection from 12-lead electrocardiogram: a multi-centre study with cross-hospital validation. *European Heart Journal-Digital Health*. 2026;7(2):ztaf125.
34. Nyström A, Björkelund A, Ohlsson M, Björk J, Ekelund U, Lundager Forberg J. Transfer learning for predicting acute myocardial infarction using electrocardiograms. *PLOS Digit Health*. 2025;4(10):e0001058.
35. Jahmunah V, Ng EY, Tan R-S, Oh SL, Acharya UR. Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals. *Computers in Biology and Medicine*. 2022;146:105550.
36. Herman R, Meyers HP, Smith SW, Bertolone DT, Leone A, Bermpeis K, et al. International evaluation of an artificial intelligence-powered electrocardiogram model detecting acute coronary occlusion myocardial infarction. *Eur Heart J Digit Health*. 2024;5(2):123–33.
37. Demandt JPA, Mast TP, van Beek KAJ, Koks A, Bastiaansen MCV, Tonino PAL, et al. Towards prehospital risk stratification using deep learning for ECG interpretation in suspected acute coronary syndrome. *BMJ Health Care Inform*. 2025;32(1).
38. Carvalho PEP, Belzer W, Pollmann DL, Helseth HC, Traverse JH, Herman R, et al. AI-Enhanced Electrocardiogram for Detection of Occlusive Myocardial Infarction in High-Risk Non-ST-Segment Elevation Acute Coronary Syndrome. *JACC Adv*. 2026;5(4):102663.
39. Sharkey SW, Herman R, Witt DR, Aguirre F, Yildiz M, Larson DM, et al. Performance of Artificial Intelligence-Powered ECG Analysis in Suspected ST-Segment Elevation Myocardial Infarction. *JACC Adv*. 2026;5(4):102671.
40. Ji RQ, Riek NT, Gokhale T, Bouzid Z, Kraevsky K, Zègre-Hemsey JK, et al. Novel machine learning fusion architectures integrating electrocardiogram representations: applications to acute coronary event detection. *Eur Heart J Digit Health*. 2026;7(5):ztag062.
41. Büscher A, Plagwitz L, Yildirim K, Brix TJ, Neuhaus P, Bickmann L, et al. Deep learning electrocardiogram model for risk stratification of coronary revascularization need in the emergency department. *European Heart Journal*. 2026;47(18):2155–67.
42. Wagner P, Strodthoff N, Bousseljot R-D, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. *Scientific data*. 2020;7(1):154.
43. Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL. *IEEE J Biomed Health Inform*. 2021;25(5):1519–28.
44. Strodthoff N, Strodthoff C. Detecting and interpreting myocardial infarction using fully convolutional neural networks. *Physiol Meas*. 2019;40(1):015001.
45. Beigzadeh N, Fathi A. HMT-KD: Lightweight Deep Learning Model for ECG Analysis Using Hierarchical. *Applied Soft Computing*. 2026:115038.
46. Huang Y, Wang M, Li YG, Cai W. A lightweight deep learning approach for detecting electrocardiographic lead misplacement. *Physiol Meas*. 2024;45(5).
47. Moody JB, Poitras-Rivière A, Renaud JM, Hagio T, Alahdab F, Al-Mallah MH, et al. A foundation transformer model with self-supervised learning for ECG-based assessment of cardiac and coronary function. *Nejm ai*. 2025;2(12).
48. Yisimitila T, Wang C, Hou M, Maimaitiniyazi M, Aili Z, Liu T, et al. Bridging clinical knowledge and AI: an interpretable transformer framework for ECG diagnosis. *NPJ Digit Med*. 2025;9(1):41.
49. Alghieth M. DeepECG-Net: a hybrid transformer-based deep learning model for real-time ECG anomaly detection. *Sci Rep*. 2025;15(1):20714.
50. Gracy GA, Pravin SC. A lightweight residual dilated temporal transformer block for ECG classification on edge devices. *Sci Rep*. 2026;16(1).
51. Abrar MF, Saqib M, Ali S, Lee IE, Khan A, Jan S, et al. Automated heart disease detection using Swin Transformer and ECG signal processing: a high-accuracy approach. *Sci Rep*. 2025;15(1):39338.
52. Ikram S, Ikram A, Singh H, Ali Awan MD, Naveed S, De la Torre Díez I, et al. Transformer-based ECG classification for early detection of cardiac arrhythmias. *Front Med (Lausanne)*. 2025;12:1600855.

53. Xiong N, Wei Z, Wang X, Wang Y, Wang Z. HCTG-Net: A Hybrid CNN-Transformer Network with Gated Fusion for Automatic ECG Arrhythmia Diagnosis. *Bioengineering (Basel)*. 2025;12(11).

Data and code availability: The complete source code, configuration files, and all experimental results are publicly available on GitHub at: <https://github.com/ladansoltanzadeh/ptbxi-ensemble-myocardial-infarction>. A permanent archived version is also available on Zenodo with DOI: 10.5281/zenodo.22347791.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ptbxlensemblyocardialinfarction.rar](#)