TABLE OF CONTENT

TrustECG: Explainable Multi-Label ECG Classification

SUMMARY

This report documents the development of TrustECG, a deep learning system that can automatically analyze 12-lead ECG recordings and detect five different cardiac conditions. What makes this project special is that it doesn't just give predictions, it also explains why it made those predictions. This is crucial in healthcare where doctors need to trust and understand AI recommendations.

Key Results:

- ***Achieved 92.1% validation AUROC and 91.2% test AUROC***
- Successfully classifies 5 cardiac conditions simultaneously
- Provides visual explanations through attention mechanisms

1. INTRODUCTION

1.1 The Problem

Electrocardiograms (ECGs) are everywhere in medicine. They're the first thing doctors look at when someone comes in with chest pain or heart problems. But here's the thing: reading ECGs properly takes years of training, and even experienced cardiologists can miss subtle patterns, especially when they're overworked or fatigued. Every year, millions of ECGs are recorded worldwide. Many of them are read by doctors who are juggling dozens of other patients. What if we could build an AI assistant that helps catch things that might be missed? Not to replace doctors, but to be a second pair of eyes.

1.2 Our Solution

We built TrustECG, a neural network that:

1. Reads 12-lead ECGs just like a cardiologist would
2. Detects multiple conditions at once (because patients often have more than one problem)
3. Shows its reasoning through attention visualization (this is the "Trust" part)

The "explainability" part is really important. We're not building a black box. Doctors can see exactly which parts of the ECG the model focused on and which leads were most important for each diagnosis.

1.3 The Dataset

We used the PTB-XL dataset, which is the largest publicly available ECG dataset. It contains:

- 21,801 clinical ECG recordings
- 12-lead format (standard clinical setup)
- 10 seconds each at 100 Hz sampling rate
- Verified by cardiologists with standardized diagnostic labels

This dataset comes from real hospital patients in Germany, so it represents actual clinical conditions, not synthetic or curated data.

2. UNDERSTANDING THE DATA

2.1 What Are We Predicting?

We focused on 5 main diagnostic categories (called "superclasses"):

Class	Full Name	What It Means
NORM	Normal	Healthy heart rhythm, nothing wrong
MI	Myocardial Infarction	Heart attack or signs of previous damage
STTC	ST/T Change	Abnormalities in specific waveform segments
CD	Conduction Disturbance	Problems with electrical signal transmission
HYP	Hypertrophy	Enlarged heart chambers

2.2 Class Distribution

Looking at our data, we found an interesting pattern:

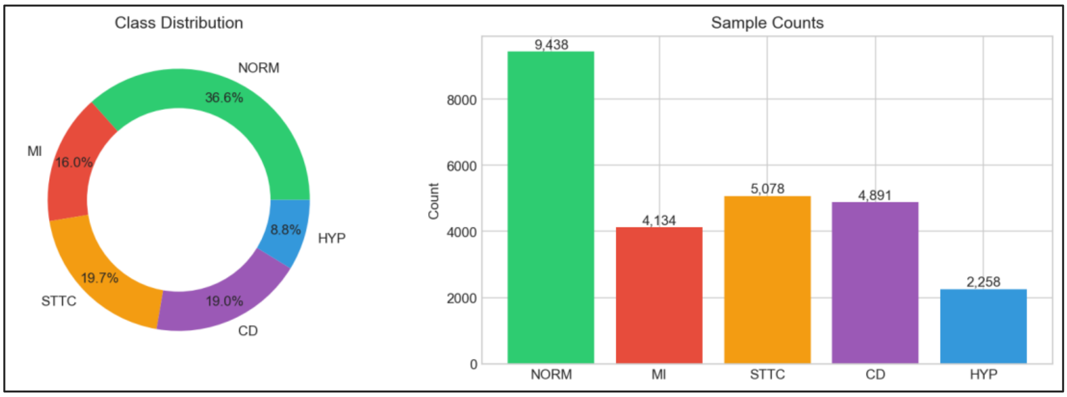


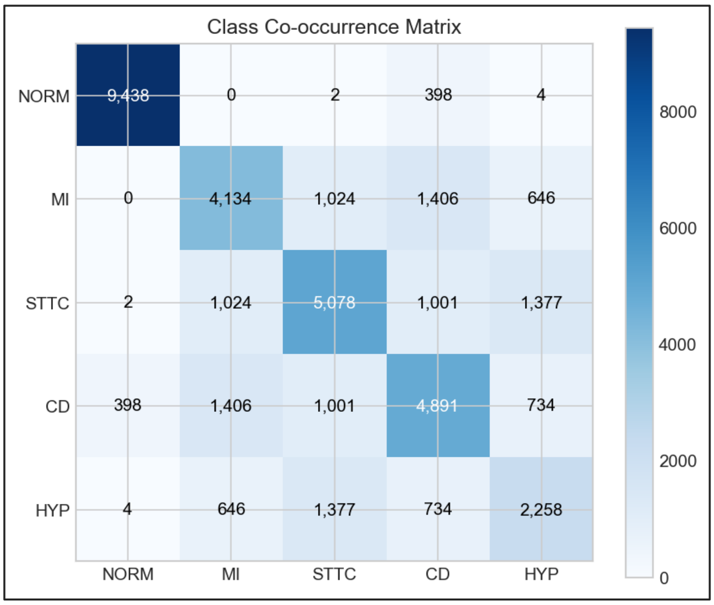
Figure 01: *Class Distribution*

This makes sense clinically. Normal ECGs are the most common (thankfully), while some conditions like hypertrophy are rarer. This imbalance is something we had to handle carefully during training.

2.3 Multi-Label Nature

Here's something important: patients can have multiple conditions at once. For example, someone might have both MI and STTC (heart attack often comes with ST changes). We found that:

- About 27% of recordings have multiple labels
- MI and STTC frequently occur together
- NORM rarely appears with other conditions (makes sense, right?)


Class Co-occurrence Matrix

	NORM	MI	STTC	CD	HYP
NORM	9,438	0	2	398	4
MI	0	4,134	1,024	1,406	646
STTC	2	1,024	5,078	1,001	1,377
CD	398	1,406	1,001	4,891	734
HYP	4	646	1,377	734	2,258

Figure 02: *Co-Occurrence*

This is why we use multi-label classification instead of picking just one class. We use sigmoid activation (not softmax) so each condition is predicted independently.

2.4 What Does an ECG Look Like?

A 12-lead ECG captures electrical activity from different angles around the heart:

- Limb leads (I, II, III, aVR, aVL, aVF): View from the frontal plane
- Chest leads (V1-V6): View from the horizontal plane

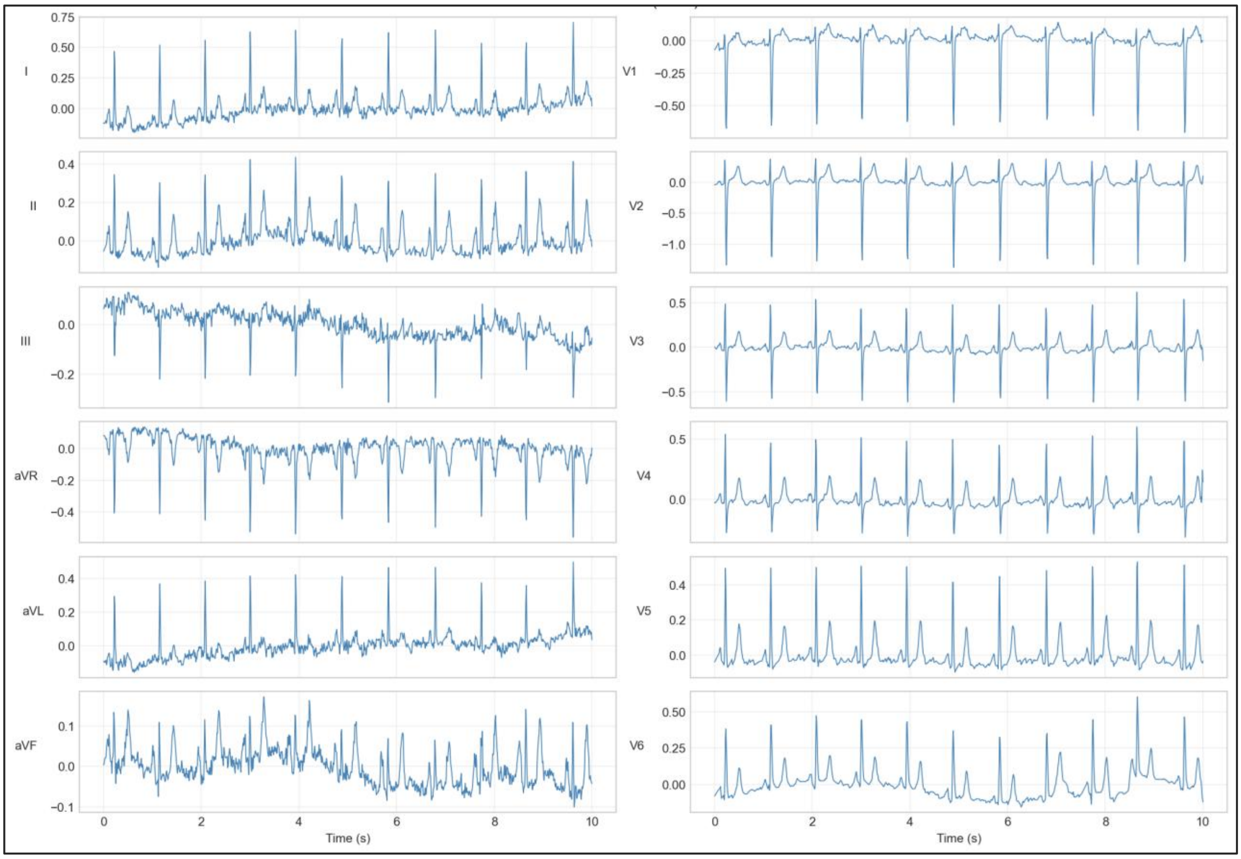


Figure 03: *Sample 12-Lead ECG*

Each lead tells a different story. For example:

- Lead II is great for rhythm analysis
- V1-V2 show the septum (middle wall)
- V5-V6 show the lateral wall

Different conditions affect different leads, which is why we built attention mechanisms to learn which leads matter for which diagnosis.

3. OUR APPROACH

3.1 Data Preprocessing

Before feeding ECGs to our model, we clean them up:

Step 1: Bandpass Filter (0.5-40 Hz)

- Removes baseline wander from breathing and movement (below 0.5 Hz)
- Removes high-frequency noise and muscle artifacts (above 40 Hz)
- Most clinically relevant ECG information is in this range

Step 2: Z-Score Normalization

- Subtracts the mean and divides by standard deviation
- Makes all ECGs comparable regardless of recording amplitude
- Formula: x_normalized = (x - mean) / std

3.2 Train/Validation/Test Split

We used the pre-defined stratification folds from PTB-XL:

Split	Folds	Count	Percentage
Training	1-8	17,441	80%
Validation	9	2,203	10%
Test	10	2,157	10%

This split ensures:

- No patient appears in multiple splits (prevents data leakage)
- Class distribution is similar across splits
- Results are reproducible

4. MODEL ARCHITECTURE

4.1 Overview

We designed "ExplainableECGNet", a neural network with 276,421 parameters. It has four main components:

1. Lead-wise Encoder: Processes each of the 12 leads separately
2. Temporal Attention: Learns which time points matter
3. Lead Attention: Learns which leads matter
4. Classification Head: Makes the final predictions

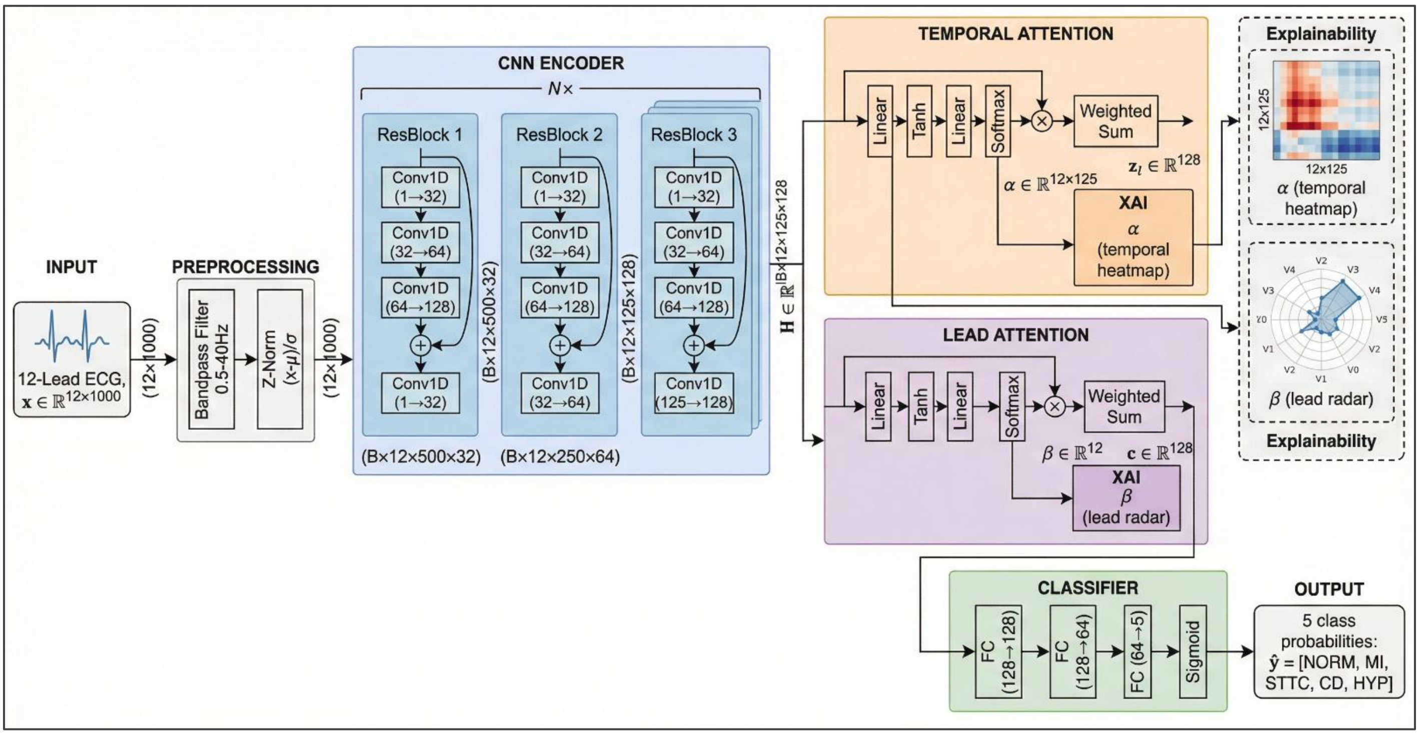


***Figure 04:** High-Level ExplainableECGNet Architecture*

4.2 Lead-wise CNN Encoder

Instead of treating the ECG as one big signal, we process each lead independently through the same encoder. This makes sense because:

- Each lead has the same structure (PQRST waves)
- Processing separately lets the model learn lead-specific features
- We can then compare across leads

The encoder uses Residual Blocks (skip connections):

- 3 blocks with channels: 32 → 64 → 128
- Kernel size: 7 (captures local patterns)
- Stride: 2 (downsamples by half)
- BatchNorm + ReLU + Dropout for regularization

After encoding, each lead goes from 1000 samples to 125-time steps with 128 features.

4.3 Temporal Attention

This is where it gets interesting. For each lead, we ask: "Which time points are most important?"

The attention mechanism learns to focus on:

- QRS complexes (the main spike)
- ST segments (elevation/depression indicates ischemia)
- T waves (inversions can indicate various conditions)

The attention weights (α) are exported for visualization. This lets doctors see exactly where the model looked.

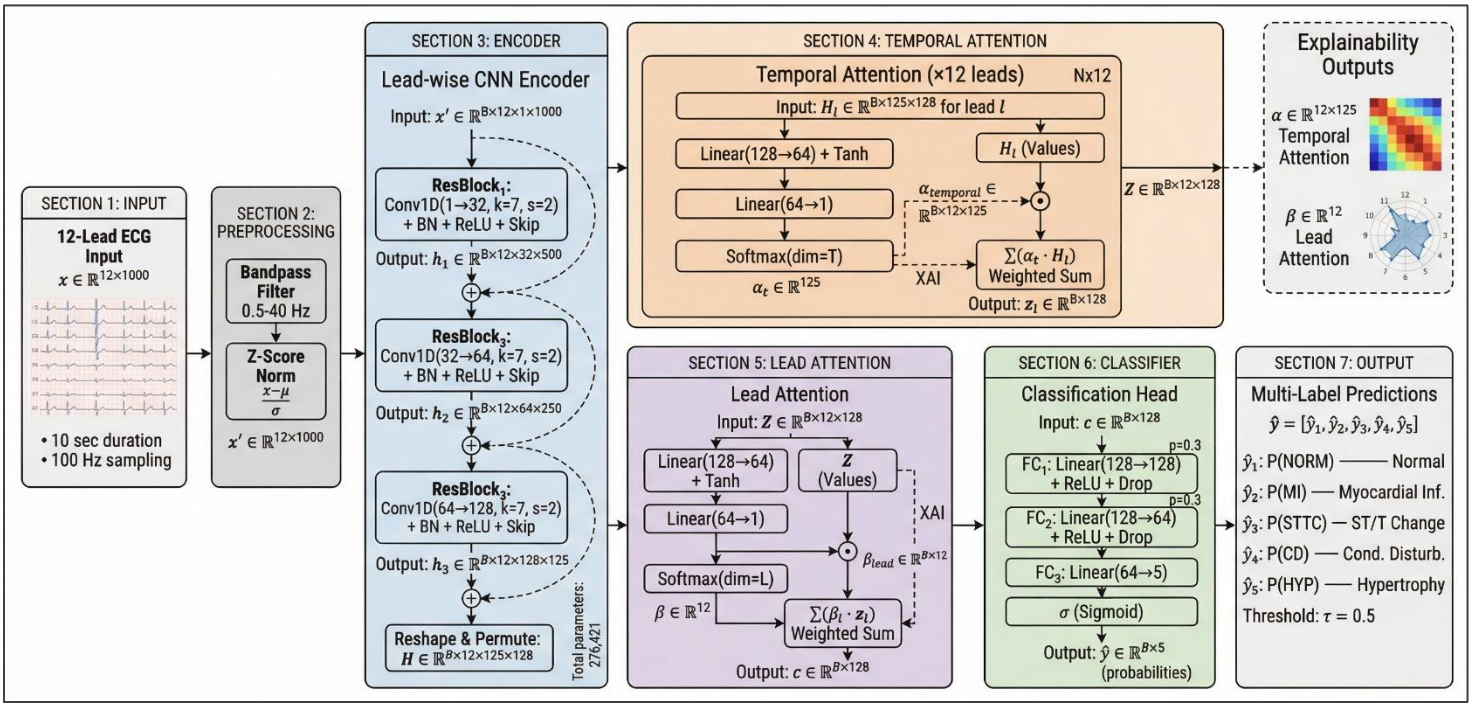

SECTION 1: INPUT
12-Lead ECG Input
$x \in \mathbb{R}^{12\times1000}$
• 10 sec duration
• 100 Hz sampling
SECTION 2: PREPROCESSING
Bandpass Filter 0.5-40 Hz
Z-Score Norm
$\frac{x-\mu}{\sigma}$
$x' \in \mathbb{R}^{12\times1000}$
SECTION 3: ENCODER
Lead-wise CNN Encoder
Input: $x' \in \mathbb{R}^{B\times12\times1\times1000}$
ResBlock$_1$: Conv1D(1→32, k=7, s=2) + BN + ReLU + Skip
Output: $h_1 \in \mathbb{R}^{B\times12\times32\times500}$
ResBlock$_2$: Conv1D(32→64, k=7, s=2) + BN + ReLU + Skip
Output: $h_2 \in \mathbb{R}^{B\times12\times64\times250}$
ResBlock$_3$: Conv1D(64→128, k=7, s=2) + BN + ReLU + Skip
Output: $h_3 \in \mathbb{R}^{B\times12\times128\times125}$
Reshape & Permute: $H \in \mathbb{R}^{B\times12\times125\times128}$
Total parameters: 276,421
SECTION 4: TEMPORAL ATTENTION
Temporal Attention (×12 leads)
Nx12
Input: $H_l \in \mathbb{R}^{B\times125\times128}$ for lead l
Linear(128→64) + Tanh
H_l (Values)
Linear(64→1)
$\alpha_{temporal} \in \mathbb{R}^{B\times12\times125}$
Softmax(dim=T)
XAI
$\Sigma(\alpha_t \cdot H_l)$ Weighted Sum
$\alpha_t \in \mathbb{R}^{125}$
Output: $z_l \in \mathbb{R}^{B\times128}$
$Z \in \mathbb{R}^{B\times12\times128}$
Explainability Outputs
$\alpha \in \mathbb{R}^{12\times125}$ Temporal Attention
$\beta \in \mathbb{R}^{12}$ Lead Attention
SECTION 5: LEAD ATTENTION
Lead Attention
Input: $Z \in \mathbb{R}^{B\times12\times128}$
Linear(128→64) + Tanh
Z (Values)
Linear(64→1)
XAI
$\beta_{lead} \in \mathbb{R}^{B\times12}$
Softmax(dim=L)
$\beta \in \mathbb{R}^{12}$
$\Sigma(\beta_l \cdot z_l)$ Weighted Sum
Output: $c \in \mathbb{R}^{B\times128}$
SECTION 6: CLASSIFIER
Classification Head
Input: $c \in \mathbb{R}^{B\times128}$
p=0.3
FC$_1$: Linear(128→128) + ReLU + Drop
p=0.3
FC$_2$: Linear(128→64) + ReLU + Drop
FC$_3$: Linear(64→5)
σ (Sigmoid)
Output: $\hat{y} \in \mathbb{R}^{B\times5}$ (probabilities)
SECTION 7: OUTPUT
Multi-Label Predictions
$\hat{y} = [\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4, \hat{y}_5]$
$\hat{y}_1$: P(NORM) — Normal
$\hat{y}_2$: P(MI) — Myocardial Inf.
$\hat{y}_3$: P(STTC) — ST/T Change
$\hat{y}_4$: P(CD) — Cond. Disturb.
$\hat{y}_5$: P(HYP) — Hypertrophy
Threshold: $\tau = 0.5$

Figure 05: *Detail ExplainableECGNet Architecture for Multi-label ECG Classification*

The Model processes 12-Lead ECG signals through a lead-wise CNN encoder with residual connections, followed by temporal attention to identify important time segments and lead attention to weight the contribution of each ECG lead. The attention weights (α, β) provide interpretability.

4.4 Lead Attention

After temporal attention, we have one feature vector per lead. Now we ask: "Which leads contributed most to this diagnosis?"

For example:

- For inferior MI, leads II, III, and aVF should be important
- For lateral MI, leads I, aVL, V5, V6 should light up
- For hypertrophy, chest leads V1-V6 are typically relevant

The lead attention weights (β) create a radar chart showing lead importance.

4.5 Classification Head

Finally, a simple MLP (Multi-Layer Perceptron) takes the weighted combination and outputs 5 probabilities:

```
Linear(128 → 128) → ReLU → Dropout
Linear(128 → 64)  → ReLU → Dropout
Linear(64 → 5)    → Sigmoid
```

Sigmoid activation means each class is predicted independently (multi-label).

5. TRAINING

5.1 Handling Class Imbalance

Since some conditions are rarer than others, we used class weights in our loss function. We tried different approaches:

4. Full inverse frequency: Gave too much weight to rare classes
5. Square root scaling: Balanced approach that worked best

Final weights (sqrt-scaled):

- NORM: 1.14
- MI: 2.07
- STTC: 1.81
- CD: 1.86
- HYP: 2.93

5.2 Training Setup

- Optimizer: AdamW (lr=0.001, weight_decay=0.01)
- Scheduler: ReduceLROnPlateau (patience=3)
- Early Stopping: Patience=5 on validation AUROC
- Batch Size: 64
- Epochs: 50 (stopped early at best validation)

5.3 Training Progress

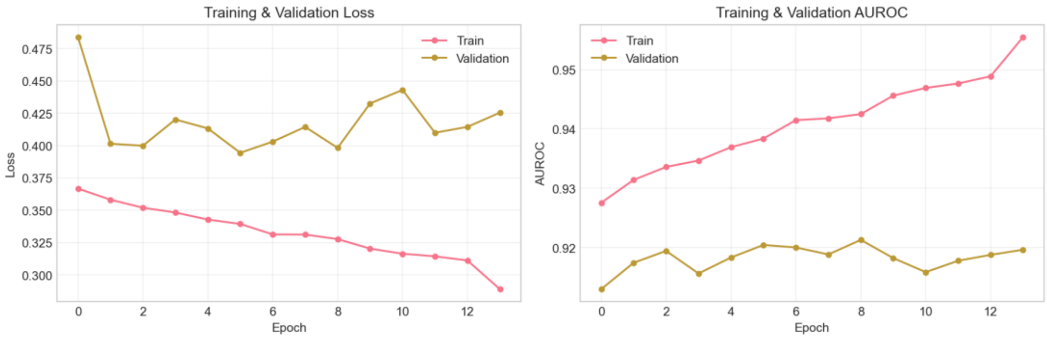


Figure 06: *Training Curves*

6. RESULTS

6.1 Overall Performance

Metric	Validation	Test
AUROC (Macro)	92.1%	91.2%
F1 Score (Macro)	69.4%	69.4%

6.2 Per-Class Performance

Class	Test AUROC	Test F1
NORM	93.5%	82.0%
MI	93.9%	70.6%
STTC	93.5%	66.6%
CD	91.0%	71.2%
HYP	84.9%	56.7%

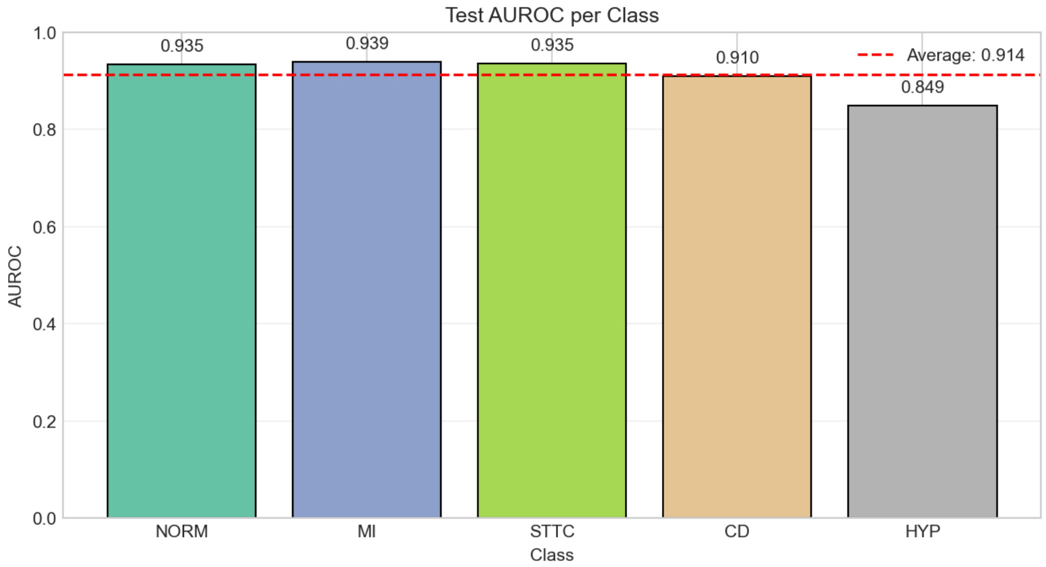


Figure 07: *AUROC Per Class*

6.3 ROC Curves

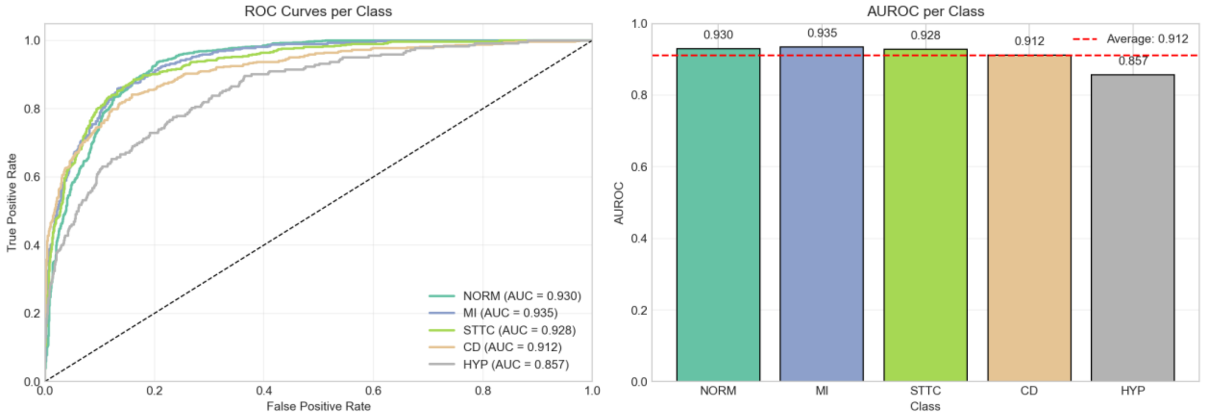


Figure 08*: Roc Curves*

6.4 Confusion Matrices

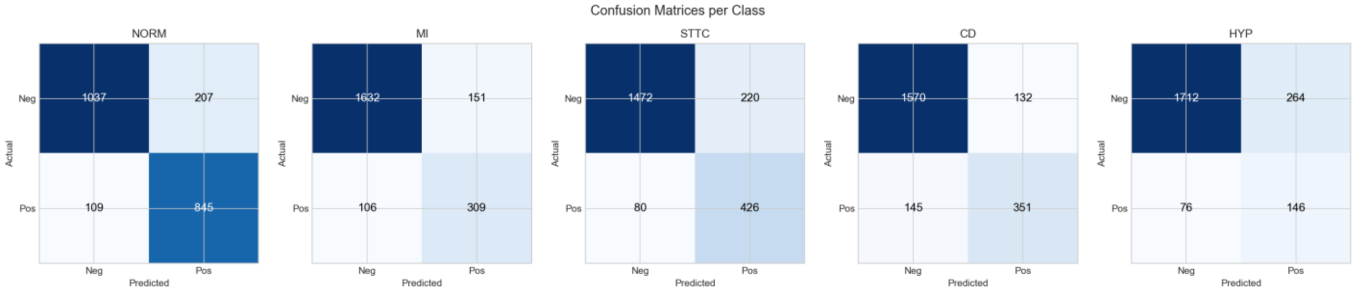


Figure 09: *Confusion Matrices*

The confusion matrices (one per class) show:

- True Negatives (top-left): Correctly identified negatives
- False Positives (top-right): Incorrectly flagged as positive
- False Negatives (bottom-left): Missed cases
- True Positives (bottom-right): Correctly identified positives

7. EXPLAINABILITY

This is what makes TrustECG different from a black box. We can visualize exactly what the model focuses on.

7.1 Lead Importance

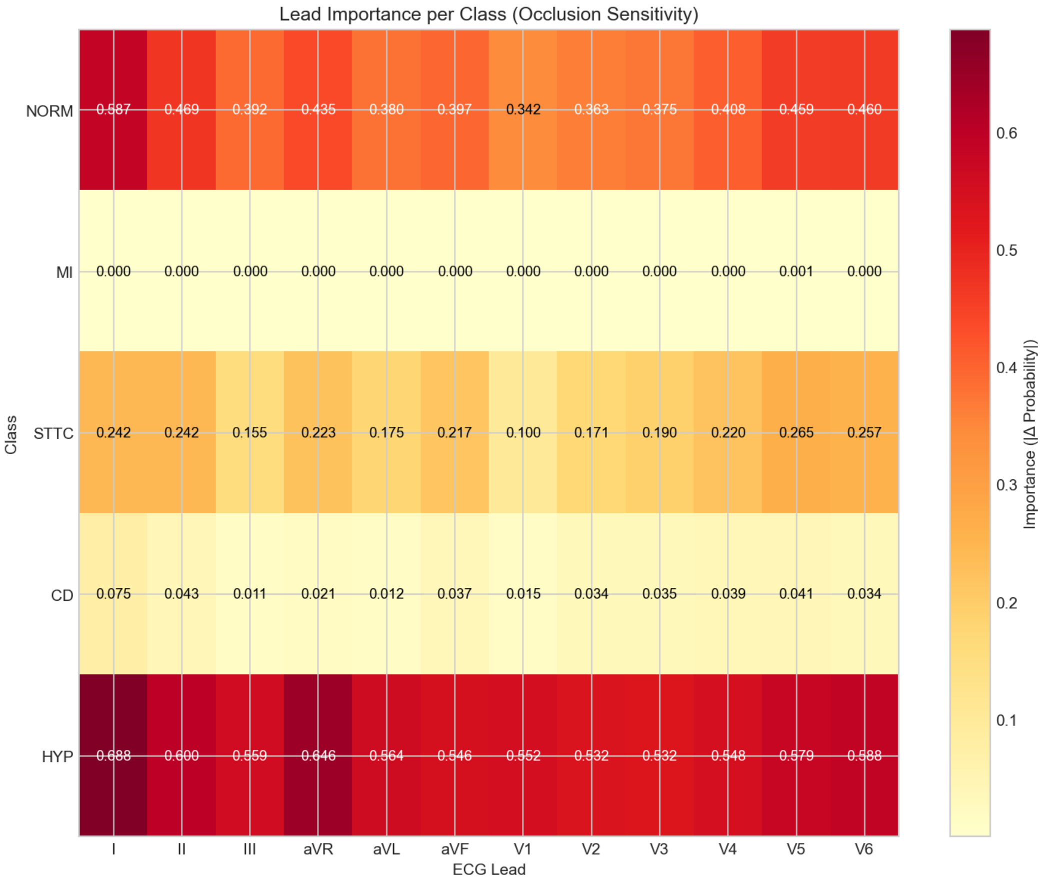


***Figure 10:** Lead Importance*

The radar chart shows which leads the model considers most important for a given prediction. This helps doctors:

- Verify the model is looking at clinically relevant leads
- Catch cases where the model might be using spurious correlations
- Build trust in the AI's reasoning

7.2 Temporal Attention Heatmap

The heatmap shows attention across time for each lead. Hot spots indicate:

- Where the model focused most attention
- Which ECG segments drove the prediction
- Potentially clinically relevant regions

7.3 Occlusion Sensitivity

We also implemented occlusion analysis: "What happens if we mask out each lead?"

This gives another view of lead importance by measuring how much the prediction changes when each lead is removed. Higher change = more important lead.

8. LESSONS LEARNED

What Worked Well

1. Attention mechanisms provide genuine explainability, not just post-hoc explanations
2. Lead-wise processing makes biological sense and the model learned relevant patterns
3. Square-root class weighting balanced rare class performance without hurting common classes
4. Preprocessing (bandpass filter + normalization) was crucial for stable training

Challenges We Faced

1. Class imbalance: Full inverse weighting hurt performance; needed careful tuning
2. HYP detection: Hardest class, likely needs more specific features
3. Preprocessing mismatch: Initially forgot to preprocess in the app, causing wrong predictions
4. Threshold optimization: 0.5 threshold isn't optimal for all classes

Future Improvements

1. Per-class threshold optimization using validation set
2. Larger model or ensemble for better HYP detection
3. Additional explainability methods (SHAP, Grad-CAM)
4. External validation on other ECG datasets
5. Real-time deployment on edge devices

9. CONCLUSION

We successfully built TrustECG, an explainable AI system for multi-label ECG classification. The model achieves strong performance (91.2% test AUROC) while providing interpretable predictions through attention mechanisms. This project demonstrates that AI in healthcare doesn't have to be a black box. By building explainability into the architecture itself, we create systems that doctors can actually trust and verify.

REFERENCES

1. *Wagner, Patrick, et al. "PTB-XL, a large publicly available electrocardiography dataset." Scientific data 7.1 (2020): 1-15.*
2. *Vaswani, Ashish, et al. "Attention is all you need [J]." Advances in neural information processing systems 30.1 (2017): 261-272.*
3. *He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.*