VRIJE UNIVERSITEIT AMSTERDAM

Bachelor's Thesis
(Course Code: XB_40001)

Impact of Label Resolution on Deep Neural Networks for ECG Classification

by

Lara Ichli
(Student Number: 2730901)

Submitted in partial fulfillment of the requirements
for the degree of
Bachelor of Science
in
Computer Science
at the
Vrije Universiteit Amsterdam

July 18, 2025

Certified by .....
Dr. Kousar Aslam
Assistant Professor
First Supervisor

Certified by .....
Auke de Leeuw
Managing Director Annovaing B.V.
Daily Supervisor

Certified by .....
Zubaria Inayat
Lecturer
Second Reader

Impact of Label Resolution on Deep Neural Networks for ECG Classification

Lara Ichli
Vrije Universiteit Amsterdam
Amsterdam, NL
l.ichli@student.vu.nl

ABSTRACT

Deep learning has shown significant promise in automating the interpretation of electrocardiograms (ECGs). However, the impact of label granularity—ranging from detailed arrhythmia-specific codes to broader severity-based categories—has not been thoroughly explored. In certain triage settings, such as primary care, coarser labels could be advantageous, as models need to balance diagnostic precision with robustness against noise, artifacts, and variability among patients.

This study investigates how label granularity influences ECG classification performance, generalizability, and clinical applicability. A 1D ResNet model was trained on the SPH dataset (25,770 ECGs) using three labeling strategies: (1) 44-class American Heart Association (AHA) diagnostic codes, (2) a 3-tier severity-based system, and (3) a 4-tier hierarchical variant. The models were evaluated on internal (SPH) and external (PTB-XL) test sets, with additional analysis of rare and previously unseen pathologies.

The 3-tier model demonstrated the highest robustness, achieving 87% accuracy (F1 = 0.87) internally and 60.5% accuracy on external data. In contrast, the 44-class model failed to generalize to low-frequency conditions, yielding an external F1 score of 0. The 4-tier model maintained diagnostic nuance (e.g., differentiating between "minor" and "moderate" abnormalities) with only a 2% accuracy reduction. Attention-based interpretability further revealed that severity-tiered models consistently focused on diagnostically relevant waveform regions, while fine-grained models tended to overfit noise.

These findings suggest that coarse, severity-based labeling improves both generalizability and interpretability, making it a practical strategy for real-world deployment in primary care triage. The study offers empirical support for hierarchical labeling, a reproducible evaluation framework, and design insights for clinically robust ECG classification systems.

1 INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death worldwide. In 2021 they accounted for approximately 19.91 million deaths globally [4], and this figure is projected to reach nearly 23 million by 2030 [21]. Early detection of heart conditions is crucial for preventing severe complications and improving patient outcomes through timely interventions. Advances in artificial intelligence (AI) and portable ECG monitoring are transforming cardiac care [3]. AI-enabled Holter monitors—wearable devices that continuously record the heart's electrical activity—now not only track ECG signals in real time but also analyze large volumes of data to detect arrhythmias and urgent conditions in real time e [22].

This study addresses key challenges in today's clinical referral pathways. In many European countries, including the Netherlands, patients must first consult a general practitioner (GP) to obtain a referral to a cardiologist. GPs typically rely on patient-reported symptoms to assess urgency. An AI Holter device can empower GPs by providing data-driven severity scores, supporting evidence-based triage, reducing diagnostic uncertainty, avoiding unnecessary specialist consultations, and ensuring swift referral of urgent cases. Once a patient reaches a specialist, the same AI system can offer a more detailed diagnostic layer tailored to specific conditions. If clinical value is demonstrated, early AI-based detection may reduce dependency on traditional ECG machines, which often suffer delays due to high patient volumes and operational costs. A two-stage, hierarchical approach—first broad severity tiers for GP triage, then detailed labels for specialist analysis—aligns with regulatory, ethical, and practical constraints, simplifies certification, and mitigates risks such as patient anxiety or liability.

Although severity-based classification has attracted interest in other domains, its application to ECG analysis remains underexplored. No systematic study has yet examined how label granularity influences ECG model performance. In computer vision, finer labels (e.g., "Persian cat") can improve accuracy even on coarser tasks ("cat") [12], but ECG data present unique challenges: they are sequential, non-stationary time series subject to heart-rate variability, electrode placement shifts, motion artifacts, and nonlinear noise (baseline wander, muscle interference) that is difficult to remove without losing diagnostic detail. Severe class imbalance—normal rhythms vastly outnumber dangerous arrhythmias—further biases models. By grouping diagnoses into broader severity tiers, one can reduce overfitting, share representations across classes, and enable models to flag rare or novel arrhythmias as "severe." In primary-care settings, this strategy promises improved generalizability, focus on critical signal features, and enhanced performance in data-scarce regimes.

To address our research question—*How do fine-grained diagnostic labels versus coarser severity tiers affect deep-learning ECG classification in terms of F1 score, precision, recall, inference time, and resource efficiency?*—we conduct a comparative study using a 1D ResNet architecture trained under three labeling schemes: (i) fine-grained AHA diagnostic codes, (ii) three-tier severity labels, and (iii) four-tier severity labels, all derived from cardiologist-reviewed mappings. We evaluate each model on an external cohort and apply Grad-CAM to identify which ECG segments drive predictions under each labeling strategy.

We employ a deep-learning framework for its end-to-end capability and strong empirical performance in recent ECG studies [5]. By operating directly on raw signals, we avoid handcrafted features and

isolate the impact of label design on model behavior. Prior work has demonstrated that ECG deep models can balance computational efficiency and high accuracy for practical deployment [6, 8, 26], but clinical adoption also demands explainability. We integrate Grad-CAM—successfully applied in prior ECG analyses [15, 16]—to generate attention maps that highlight signal regions influencing decisions. This transparency aligns model outputs with clinical reasoning and may uncover novel diagnostic patterns.

Our systematic comparison reveals that coarser severity tiers yield superior robustness and interpretability for ECG triage. The three-tier ResNet1D model achieves 87% internal accuracy and maintains 60.5 % accuracy on an external cohort, whereas the four-tier variant provides finer stratification (mild vs. moderate) at only a modest 2 % performance decrease. Conversely, the 44-label Disease ResNet delivers high internal F1 scores (> 0.90) on common arrhythmias but fails to generalize to rare or unseen pathologies, underscoring the long-tail and distribution-shift limitations of fine-grained classification. Grad-CAM analyses further show that severity-tier models focus attention on clinically salient waveform features, while multi-label models risk over-attending to noise and non-diagnostic segments.

This work makes four principal contributions: **Empirical quantification** of the trade-off between label granularity and model generalizability across internal, external, and unseen-pathology evaluations. **A hierarchical, two-stage framework** aligned with European primary-care referral pathways—enabling GP-level severity triage followed by specialist-level diagnostic refinement—to facilitate regulatory and ethical deployment. **Open-source resources** including mappings from 44 AHA codes to three- and four-tier schemes, pretrained ResNet1D models, and Grad-CAM visualization tools to ensure reproducibility and support future research. **Practical Insights** for AI-enhanced Holter device models on when to aggregate rare diagnoses into broader tiers versus when to preserve label specificity given data and annotation constraints.

Collectively, these findings inform the design of enhanced Holter devices for both primary-care triage and specialist workflows and contribute to the growing body of evidence on how task framing and label structure influence machine-learning behavior.

2 RELATED WORK

2.1 Deep Learning Approaches for ECG-Based Disease Classification

Deep learning has significantly advanced the classification of diseases based on ECG (electrocardiogram) data, often outperforming traditional models by 10–15% in accuracy [17]. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) excel at capturing both spatial and temporal dependencies in ECG waveforms. In particular, ResNet architectures have gained widespread attention in ECG applications because their residual (skip) connections prevent vanishing gradients in very deep networks and allow for efficient fine-tuning on small or evolving datasets. Yang et al. [30] introduced 1D-ResNet-AdaSOM, which uses an adaptive learning rate technique, achieving an average F1 score of 0.862 on the CPSC2018 dataset. Li et al. [19] enhanced 1D ResNet-18 for ECG classification with a Convolutional Block Attention Module and auxiliary classifier, offering improved accuracy on the MIT-BIH Arrhythmia Database. Sakli et al. [27] created a ResNet-50 model that classifies 12-lead ECG signals into 27 categories, reaching 97.63% accuracy and 89.67 % precision. Boulkaboul et al. [9] developed D-Resnet, a deep network for six ECG rhythm types, demonstrating effective feature extraction for automated arrhythmia monitoring.

Beyond ResNets, a variety of other deep architectures and design choices have shown promise. Cai et al. [10] proposed a Deep Densely Connected Network (DDNN) for atrial fibrillation detection, achieving 99.35% accuracy and fast inference suitable for real-time deployment. Acharya et al.[2] used a nine-layer CNN trained on raw and denoised ECG signals for multi-class heartbeat classification, achieving accuracies of 94.03% and 93.47%, respectively. They addressed class imbalance by applying Z-score transformations to augment minority classes. Chen et al.[11] combined CNNs with LSTMs to classify six rhythm types using the MIT-BIH dataset. The hybrid architecture reached 99.32% accuracy, demonstrating how CNNs aid in feature extraction while LSTMs excel in modeling temporal sequences. Narotamo et al.[23] compared one-dimensional (1D) and two-dimensional (2D) ECG representations using GRU, LSTM, and CNN models, as well as multimodal fusion. They found that GRU-based 1D models performed best (79.67% sensitivity, 81.04% specificity), suggesting that retaining the native 1D structure preserves more diagnostic information than converting signals to images. Lai et al.[18] developed a self-supervised, multi-label learning framework using momentum contrastive learning and a Siamese CNN. Pretrained on over 639,000 ECGs, their model achieved an AUROC of 0.975 offline, and maintained 0.736 sensitivity and 0.954 specificity in a two-month real-world test on wearable devices.Peng et al. [24] introduced EGCNet, a hierarchical graph convolutional network that used 1D residual convolutions and a disease graph, achieved 0.753 precision on the PhysioNet dataset. These studies highlight the importance of architectural design in model performance, guiding our focus on essential components. However, most research concentrates on rhythm or disease detection, leaving a gap in severity-based stratification, which this study aims to address.

2.2 Strategies for Class Imbalance and Structured Label Learning

Several researchers have tackled class imbalance and label complexity through augmentation or structured learning. As noted earlier, Acharya et al.[2] used Z-score transformations to augment underrepresented classes. Fan et al.[14] addressed class imbalance through an active subset-selection strategy integrated into a modified Broad Learning System. Their iterative data refinement and majority-vote aggregation improved the detection of minority classes. Peng et al. [24] leveraged label relationships in a graph-based structure to capture contextual dependencies between disease classes, which helped improve generalization in multi-label settings. While these approaches offer meaningful solutions to class imbalance, they often treat disease classes as mutually exclusive labels. In contrast, our approach models hierarchical severity groupings alongside disease types, allowing for more nuanced triage-oriented classification and better alignment with clinical decision-making in primary care.

2.3 Progress and Gaps in Severity-Based ECG Classification

Recent work has begun to explore ECG classification by disease severity, though significant gaps remain. Abdellatif et al.[1] applied SMOTE to improve performance across both severity and disease classes using six traditional ML algorithms. However, they did not specify the criteria for defining severity. BhanuSri et al.[7] also incorporated severity but lacked transparency in stratification and performance reporting. Both used the Cleveland Heart Disease dataset, which relies on interpreted rather than raw ECG signals, limiting its utility in real-time clinical scenarios. Prabhakararao et al.[25] proposed a multi-lead attention-based RNN (MLDA-RNN) for staging myocardial infarction (early, acute, chronic). The model incorporated intra- and inter-lead attention and achieved 97.79% accuracy, with attention maps aligning well with expert annotations. Togo et al.[28] developed a DNN to assess heart failure (HF) severity in hypertrophic cardiomyopathy patients using raw ECGs. Using Grad-CAM and Integrated Gradients, they highlighted clinically relevant QRS regions, achieving a weighted F1 score of 0.745 and precision of 0.750. Diware et al. [13] designed a hierarchical arrhythmia classifier optimized for energy efficiency, but their method did not evaluate severity detection, and was restricted to rhythm abnormalities. Our study explicitly defines and validates severity tiers, addresses primary care constraints, and extends severity classification to include a broader set of cardiovascular conditions.

3 RESEARCH METHOD

3.1 Research Design

In this study, we take a comparative experimental approach by training three ResNet1D models—each using a different labeling scheme (full AHA codes, three-level severity, and four-level severity)—on the same SPH ECG dataset [20]. After optimizing each model under identical conditions, we evaluate their performance both on the held-out SPH test set and on the external PTB-XL [29] cohort to assess how label granularity influences accuracy and robustness to new patient populations.

3.2 Labeling Approach

We compared three labeling schemes:

(1) *High-Resolution Diagnostic*: All original AHA codes (44 diagnoses, 15 modifiers) are kept, preserving maximum pathophysiological detail.
(2) *Three-Level Severity*: We group codes into {normal, non-urgent abnormal, urgent}, matching common triage categories in primary care.
(3) *Four-Level Severity*: We split "non-urgent abnormal" into "minor" and "major" to allow finer risk assessment.

A board-certified cardiologist mapped each high-resolution code into the three- and four-level schemes based on shared ECG features, typical time-to-intervention, and standard referral pathways. The complete mapping is shown in Table 2, and the disease names corresponding to the AHA codes are listed in Table 1.

ECG recordings were obtained from two publicly available cohorts. For model development (training and internal testing), the SPH dataset [20] (China; August 2019–August 2020) was employed, comprising 25 770 ten-second ECGs annotated per AHA standards (44 diagnoses + 15 modifiers), of which 46 % contain abnormalities and 14 % carry multiple labels . External validation utilized the PTB-XL dataset [29] (Germany), consisting of 21,799 ten-second ECGs annotated with SCP-ECG codes mapped to 71 AHA classes.

Table 1: AHA Code Descriptions

AHA Code	Freq.	Description
1	13 907	Normal ECG
22	2 712	Sinus bradycardia
147	2 044	T-wave abnormality
23	1 553	Sinus arrhythmia
105	1 259	Incomplete right bundle-branch block
146	1 063	ST deviation with T-wave change
145	770	ST deviation
21	725	Sinus tachycardia
106	710	Right bundle-branch block
50	460	Atrial fibrillation
60	427	Ventricular premature complex(es)
125	322	Low voltage
82	238	Prolonged PR interval
142	209	Left ventricular hypertrophy
30	220	Atrial premature complex(es)
120	161	Right-axis deviation
101	154	Left anterior fascicular block
121	138	Left-axis deviation
51	99	Atrial flutter
104	84	Left bundle-branch block
153	88	ST-T change due to ventricular hypertrophy
36	64	Junctional premature complex(es)

Table 2: Severity Label Mappings

AHA Code	Freq.	3-Class	4-Class
1	13 907	0 (Benign)	0 (Normal)
23	1 553		
147	2 044		
21	725	1 (Non-Urgent)	1 (Minor)
22	2 712		
30	220		
36	64		2 (Major)
82	238		
101	154		
105	1 259		
120	161		
121	138		
125	322		
145	770		
146	1 063		
50	460	2 (Urgent)	3 (Critical)
51	99		
60	427		
104	84		
106	710		
142	209		

34 unique AHA codes appear, 15 of which overlap with SPH and 19 of which are novel, allowing assessment of generalization to unseen conditions.Pronounced class imbalance is observed in both datasets.

3.3 Model Architecture and Training

Our model is based on a one-dimensional ResNet (ResNet1D) that processes 12-lead ECG signals. We selected ResNet for its residual connections, which facilitate the network's ability to learn both short events—such as sharp QRS spikes—and longer patterns, like overall heart rhythms, by preserving information as depth increases. To address noisy recordings and the rarity of certain heart conditions, we designed a comprehensive training pipeline covering data preparation, network structure, and optimization.

The core of the network consists of three main components. First, an initial convolutional layer (kernel size = 15, stride = 2) transforms the 12-channel input into 64 feature maps; this is followed by batch normalization (which stabilizes and speeds up training), a ReLU activation (which introduces nonlinearity by zeroing negative values), and max-pooling (kernel size = 3, stride = 2) to quickly reduce sequence length while preserving key waveforms. Next, four residual stages each contain two residual blocks. In the first stage, we maintain full temporal resolution at 64 channels to capture sharp, transient events. In stages two through four, we downsample by a factor of two (halving sequence length) and double the channel count (64 → 128 → 256 → 512) so that deeper layers can learn broader rhythm patterns. Within each block, a Squeeze-and-Excitation (SE) module—an attention mechanism that "squeezes" global channel statistics into a summary vector and then "excites" (re-weights) each channel—guides the model to focus on the most informative leads or features. Finally, after the last stage, we apply global average pooling to condense each channel into a single value, add dropout (p = 0.5) for regularization, and use a linear layer to produce class scores.

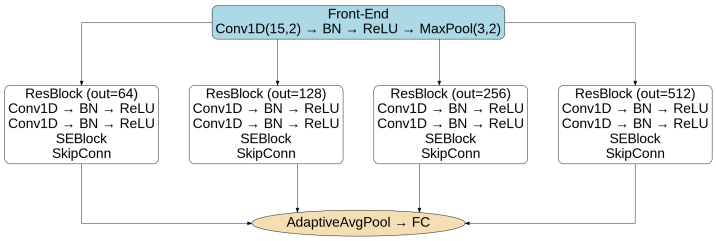


Figure 1: Optimized ResNet1D—Front-End, four SE-enhanced ResBlocks with downsampling, and Classification Head.

For training, we developed an ECGTrainer that uses the AdamW optimizer (weight decay = 1×10^{-5}) together with a OneCycleLR scheduler (peak learning rate = 1×10^{-3}) over up to 100 epochs. We select the loss function based on the task: weighted cross-entropy for single-label problems and binary cross-entropy with logits for multi-label situations, both incorporating class weights from the training set to mitigate imbalance. To prevent overfitting, we include dropout in the final layer and implement early stopping if the validation F1 score does not improve for 10 consecutive epochs. Additionally, we enable automatic mixed precision (AMP) to accelerate computation and reduce memory usage, making GPU training more efficient and scalable.

3.4 Evaluation and Measures

(1) **Validation Strategy**
   (a) *Internal validation:* We evaluated the model using a two-tier approach. First, internal validation on our in-house dataset employed stratified splits to preserve class frequencies and measure performance on known ECG patterns.
   (b) *External validation:* Second, external validation on PTB-XL tested both pathologies present in training and held-out conditions grouped by clinical severity, assessing the model's ability to generalize to new cohorts and unseen abnormalities.

(2) **Performance Metrics**
   (a) *Precision:* Of all true labels of a class the model predicts, the fraction that are correct—i.e., how often a positive prediction truly indicates the condition:

$$\text{Precision} = \frac{\sum_{j=1}^{C} TP_j}{\sum_{j=1}^{C} (TP_j + FP_j)}$$

   (b) *Recall:* Of all actual occurrences of each condition, the fraction the model correctly identifies—i.e., the sensitivity to real events in the ECG:

$$\text{Recall} = \frac{\sum_{j=1}^{C} TP_j}{\sum_{j=1}^{C} (TP_j + FN_j)}$$

   (c) F_1 *Score:* The harmonic mean of precision and recall, capturing the trade-off between false positives and false negatives in ECG classification:

$$F_1 = 2\,\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

   (d) *Macro-Averaged Metrics:* Compute each metric separately for each diagnosis and then average equally across classes:

$$\text{Macro-}m = \frac{1}{C}\sum_{j=1}^{C} m_j$$

   (e) *Weighted-Averaged Metrics:* Compute each metric per diagnosis and weight by the number of true examples n_j:

$$\text{Weighted-}m = \frac{\sum_{j=1}^{C} n_j\, m_j}{\sum_{j=1}^{C} n_j}$$

   (f) *Area Under the ROC Curve (AUC):* The probability that a randomly chosen pathological ECG is scored higher than a randomly chosen normal ECG across all thresholds:

$$\text{AUC} = \int_0^1 \text{TPR}(t)\, \mathrm{d}\big(\text{FPR}(t)\big)$$

   (g) *Confusion Matrix:* A $C \times C$ table M where entry M_{ij} counts ECGs of true class i predicted as class j, highlighting specific misclassification patterns.

(3) **Computational Performance Monitoring**
   (a) *Average Inference Latency:* Mean time (ms) to classify one 10-s ECG.

(b) *Training Time (minutes):* Total time required for the model to converge.
(c) *GPU Memory Usage (GB):* Peak VRAM consumption per epoch on an NVIDIA V100 GPU.
(d) *Epochs Trained:* Number of epochs completed before early stopping.

(4) **Interpretability Analyses**
(a) *Grad-CAM Heatmaps:* We generated Grad-CAM heatmaps over each ECG's time axis to highlight intervals (e.g., P-wave, QRS complex, ST-segment) most influential for the model's predictions.

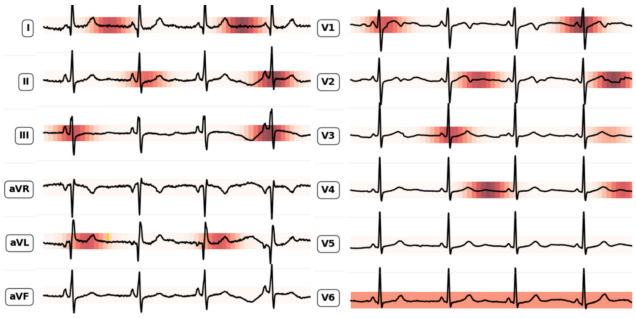
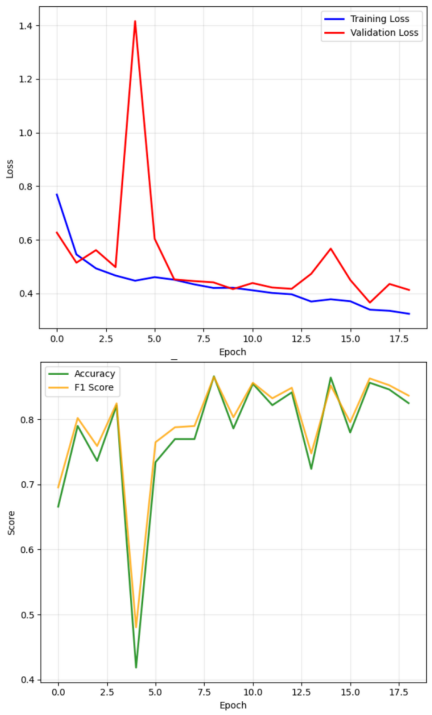


Figure 2: Grad-CAM Attention Visualization True Severity: Benign/Normal (Class 0) Predicted: Benign/Normal (Class 97.8% confidence)

Figure 3: Training and validation loss, accuracy, and F1-score for the ResNet1D three-class model over 19 epochs.



4 RESULTS

4.1 ResNet1D Three-Class Model Performance

The ResNet1D three-class model was trained for 19 epochs on a single GPU in approximately 45.5 minutes, reaching a peak memory usage of 0.35 GB. Figure 3 displays the training and validation loss curves alongside corresponding accuracy and F1-score metrics.

Over the course of training, the training loss decreased from 0.77 to 0.33, while the validation loss declined from 0.63 to 0.40. A transient rise in validation loss at epoch 4 coincides with the transition from the warm-up to the annealing phase in the OneCycleLR schedule. During this transition, the learning rate peaks, briefly destabilizing weight updates and causing accuracy and F1-score to dip to approximately 0.42 and 0.48 before recovering to the high 0.80s. After epoch 5, both loss curves continue to decrease in parallel with a residual gap of about 0.05, indicating true generalization rather than overfitting. A minor perturbation around epoch 15 likely reflects normal stochastic noise from a challenging mini-batch, as the loss promptly resumes its downward trend.

On the internal hold-out set (4,168 ECGs), the model attains an accuracy of 87% and a weighted F1-score of 0.87. Table 3 summarizes per-class performance: the Benign/Normal class achieves 0.91 precision, 0.92 recall, and an AUC of 0.912; the Serious/Treatment class records 0.82 precision, 0.91 recall, and an AUC of 0.984; the Moderate/Monitoring class lags with 0.68 precision, 0.65 recall, and an AUC of 0.890, reflecting inherent feature overlap.

Figure 4 shows ROC curves and a confusion matrix for internal validation. Although neither accuracy nor F1-score exceeds 0.90,

Table 3: Classification report for the ResNet1D three-class model on the internal hold-out set.

Class	Precision	Recall	F1-score	Support
0 (Benign/Normal)	0.91	0.92	0.92	3180
1 (Moderate/Monitoring)	0.68	0.65	0.66	818
2 (Serious/Treatment)	0.82	0.91	0.87	170
Macro avg	0.81	0.83	0.82	4168
Weighted avg	0.86	0.87	0.87	4168

and no loss reaches zero, this plateau likely arises from ambiguous waveform overlap or label noise rather than model capacity limits.

When applied without retraining to the external AHA-coded cohort, accuracy falls to 60.5% and micro-average AUC to 0.751 (Figure 5). The Benign and Serious classes maintain reasonable AUCs (0.850 and 0.745), whereas the Moderate class drops to an AUC of 0.659 and precision of 0.23. This degradation is attributed to distributional shifts in patient demographics and recording devices, which disproportionately affect the "Moderate" category defined by exclusion.

External evaluation on the Known AHA cohort yields an accuracy of 0.5701, an F1-score of 0.6059, across 16,812 samples, with an average inference time of 27.25 ms per ECG.

In the unseen-pathology evaluation, novel morphologies defaulted to the Moderate class. Recall for Benign and Serious classes dropped to 27% and 43% (micro-average AUC 0.467), underscoring uncertainty when encountering out-of-distribution examples. Additionally, the ResNet1D_3Class_Unseen evaluation yields an

Figure 4: ROC curves and confusion matrix for internal validation of the ResNet1D three-class model.

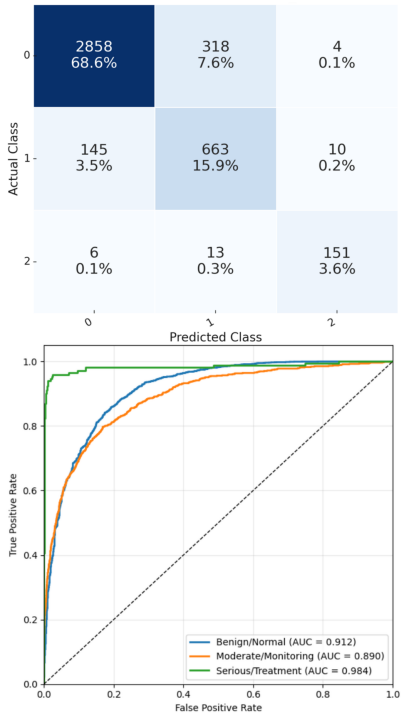
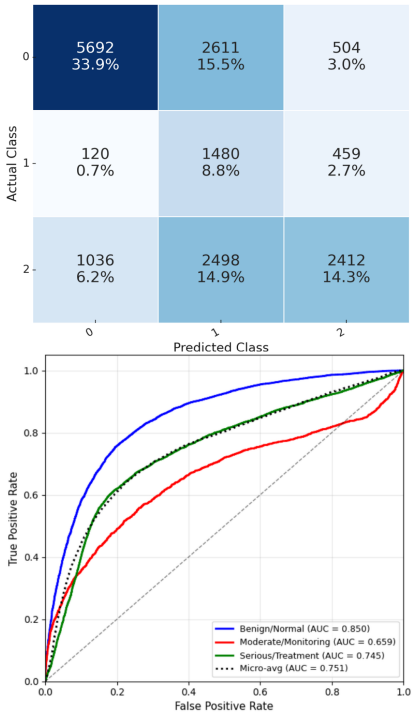


Figure 5: ROC curves and confusion matrix for the external AHA-coded cohort (three-class model).



accuracy of 0.2328, an F1-score of 0.3542, on 4, 312 samples, with an average inference time of 26.21 ms per ECG.

4.2 ResNet1D Four-Class Model Performance

The ResNet1D four-class model completed 15 epochs in 36.4 minutes, peaking at 0.41 GB of GPU memory. Figure 7 shows the loss, accuracy, and F1-score trajectories. Training loss declined from 1.02 to 0.42, and validation loss from 0.81 to 0.44. A brief fluctuation at epoch 2 corresponds to a momentary dip in validation accuracy (0.66) and F1-score (0.75), likely caused by a learning-rate transition or a difficult batch. From epoch 5 onward, both loss curves descend in near-parallel with a gap below 0.05, indicating stable convergence and feature internalization.

On the internal hold-out set, the model achieves 84.9% accuracy and a weighted F1-score of 0.85. Table 4 details per-class metrics: strong performance for Benign/Normal and Serious, with lower recall for Mild and Moderate due to feature overlap.

Table 4: Classification report for the ResNet1D four-class model on the internal hold-out set.

Class	Precision	Recall	F1-score	Support
0 (Benign/Normal)	0.91	0.92	0.92	3180
1 (Mild)	0.75	0.74	0.75	818
2 (Moderate)	0.65	0.64	0.64	170
3 (Serious)	0.82	0.89	0.85	170
Macro avg	0.78	0.80	0.79	4168
Weighted avg	0.85	0.85	0.85	4168

External evaluation on the Known AHA cohort shows an accuracy of 0.5566 and an F1-score of 0.6023, evaluated on 16, 812 samples with an average inference time of 27.20 ms per ECG. Figure 8 presents internal ROC curves and a confusion matrix. Without retraining on the external cohort, overall accuracy declines to 52.92% and micro-average AUC to 0.714 (Figure 9). The absence of mild-severity labels and distributional differences lead to AUCs of 0.56 (Benign), 0.45 (Moderate), and 0.63 (Serious). The F1-score on this cohort is 0.553.

In the unseen-pathology evaluation (Figure 10), recall drops to 32% (Benign), 27% (Moderate), and 36% (Serious), with a micro-average AUC of 0.587. Furthermore, the ResNet1D_4Class_Unseen external evaluation yields an accuracy of 0.2001, an F1-score of 0.3106, on 4, 312 samples, and an average inference time of 26.35 ms per ECG.

4.3 Disease ResNet Model Performance

The multi-label Disease ResNet model was trained for 36 epochs in 84.7 minutes, reaching 0.47 GB peak GPU usage. Figure 11 shows training and validation dynamics. Training loss fell rapidly from 0.15 to 0.03 by epoch 10 and reached 0.01 by epoch 35. Validation loss decreased similarly until epoch 10 but began to increase after epoch 25 while training loss continued downward, indicating overfitting. Accuracy and weighted F1-score rose from 0.67 to 0.82 by the final epoch.

Table 5 summarizes per-class metrics for the top 20 AHA codes on the internal hold-out set. Frequent codes (e.g., AHA 1, 21) achieve

Figure 6: ROC curves and confusion matrix for unseen-pathology evaluation (three-class model).

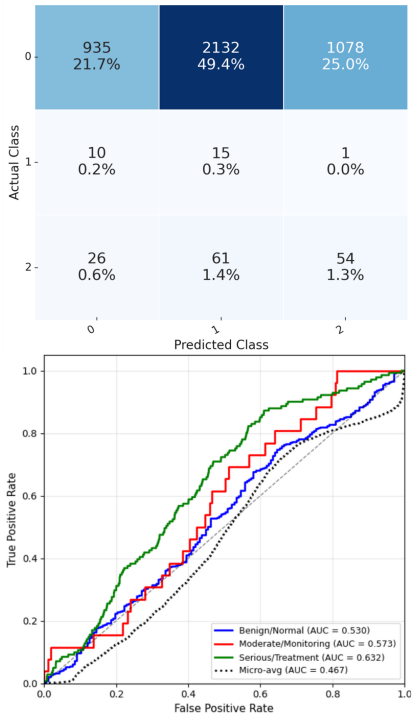
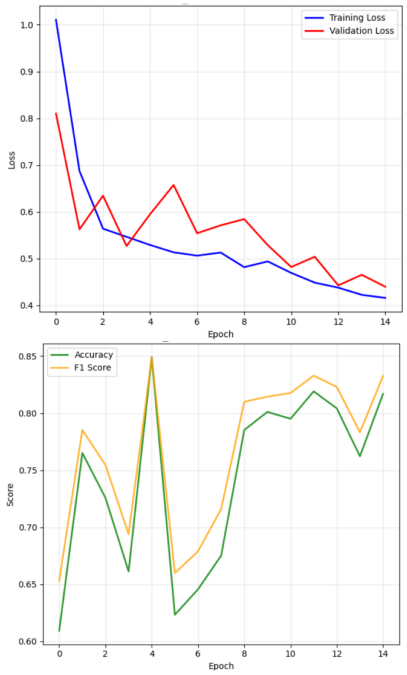


Figure 7: Training and validation loss, accuracy, and F1-score for the ResNet1D four-class model over 15 epochs.



precision and recall above 0.90, whereas mid-frequency and rare codes exhibit lower recall or collapse to zero.

External validation on a real-world cohort (Figure 12) shows robust generalization for common codes, yet simultaneously highlights the profound challenge of rare pathology transfer. This difficulty stems from the "long-tail problem," a common data characteristic where a few items are highly frequent (the "head"), while many others are extremely rare (the "tail"). Machine learning models, trained predominantly on the frequent examples, often struggle to generalize to these rare, out-of-distribution instances.

This struggle is directly reflected in the external evaluation of the ResNet1D_Disease model, which yielded an accuracy of 0 and an F1-score of 0 on 16,812 samples. These zero scores explicitly mean that for certain rare pathologies, the model made no correct predictions, underscoring the significant hurdle posed by the long-tail of diseases in clinical AI deployment.

4.4 Inference Timing Summary

Average inference time and throughput per model are shown in Table 6.

5 DISCUSSION

5.1 Model Performance and Label Granularity Analysis

Our results show that label granularity has a marked effect on classifier accuracy, stability, and clinical value. With a three-class scheme combining Mild and Moderate into “Non-serious”, the ResNet1D model achieves its best internal metrics (87% accuracy, $F_1 = 0.87$). By avoiding subtle waveform distinctions (often indistinguishable

Table 5: Per-class metrics for Disease ResNet on the internal hold-out set (top 20 AHA codes).

AHA Code	Precision	Recall	F1-Score	Support
1	0.88	0.95	0.92	2856
22	0.84	0.89	0.87	231
147	0.81	0.59	0.68	211
105	0.64	0.48	0.55	184
23	0.74	0.32	0.45	158
146	0.70	0.61	0.65	94
145	0.57	0.39	0.46	88
106	0.97	0.91	0.94	81
21	0.92	0.97	0.95	72
50	0.95	0.97	0.96	37
60	0.97	0.88	0.92	32
125	0.64	0.48	0.55	29
120	0.00	0.00	0.00	18
121	0.00	0.00	0.00	15
30	0.00	0.00	0.00	11
101	0.23	0.33	0.27	9
104	1.00	1.00	1.00	9
82	0.00	0.00	0.00	5
36	1.00	0.25	0.40	4
...				
Macro avg	0.59	0.50	0.53	4148
Weighted avg	0.84	0.84	0.84	4148

Figure 8: Training and validation loss, accuracy, and F1-score for the ResNet1D four-class model over 15 epochs.

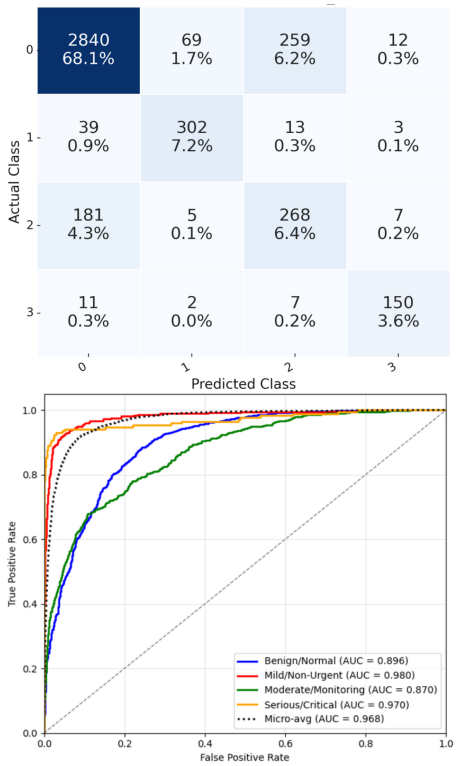


Figure 9: ROC curves and confusion matrix for external AHA-coded cohort (four-class model).

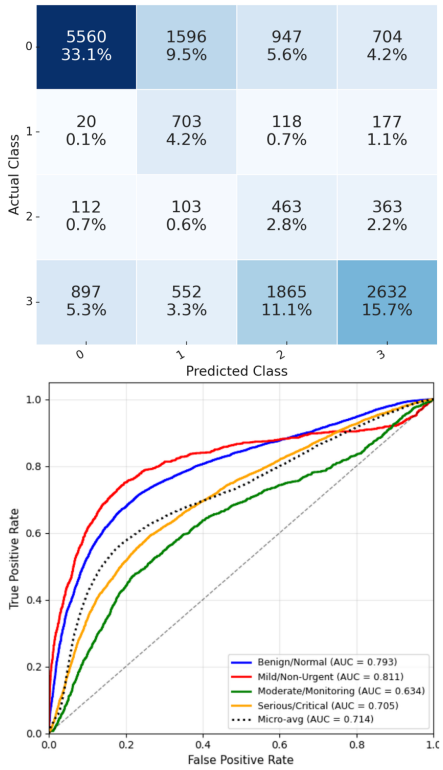


Table 6: Inference timing summary for ResNet1D models on AHA datasets.

Model	Samples	Avg Time(ms)	Throughput(samples/s)
3-Class	16,812	27.25	23.0
4-Class	16,812	27.20	23.3
Disease	16,812	27.22	23.7

from noise), this coarser scheme sharpens boundaries between benign and serious rhythms, yielding high recall (benign: 92%, AUC = 0.912; serious: 91%, AUC = 0.984) and reduced artifact sensitivity.

Reintroducing the Mild/Moderate split adds clinical nuance but reduces robustness. The four-class model drops to 84.9% accuracy and $F_1 = 0.85$, with most errors occurring between Mild and Moderate. These mistakes reflect both morphological overlap and class imbalance under uniform error weighting. On an external AHA-coded cohort, accuracy falls further to 52.9% (micro-AUC = 0.714) and AUCs for benign/serious decline to 0.56 and 0.63, respectively.

At the finest resolution of diagnostic codes, the model accurately identifies arrhythmias with clear signatures (e.g. atrial fibrillation, AUC > 0.90) but struggles with subtle features (minor PR prolongation, low-amplitude T-wave changes). Rare or mid-frequency codes perform near chance (AUC 0.45–0.46). Thus, broader categories aggregate noise-prone patterns into learnable groups and improve resilience to shifts in demographics, devices and annotation styles.

Technically, coarse labels constrain noise-driven feature fluctuations and balance recall, whereas fine labels require greater representational capacity and amplify noise sensitivity. Practically, fewer classes lower the risk of performance collapse under distribution shifts.

5.2 Clinical Workflows, Reliability and Ethical Considerations

A three-category system—"serious," "monitor" and "normal"—matches or exceeds more detailed schemes in real-world settings. In primary care, where ECGs are often noisy, the key decision is referral versus observation. A three-label output yields a clear action: serious rhythms trigger referral, normal rhythms reassure both clinician and patient, and monitor rhythms prompt follow-up. This reduces unnecessary specialist visits, mitigates patient anxiety and lightens cardiology workload.

We recommend a two-phase workflow. First, the GP's device applies the three-class model for an instantaneous recommendation. Then all recordings are uploaded to a cloud service for comprehensive analysis. Serious cases are prioritized for specialist review, while monitor and normal cases are queued by urgency. This ensures high-risk patients receive prompt attention without neglecting ambiguous tracings.

Ongoing collaboration with clinicians is essential. During development, we review Grad-CAM heatmaps with doctors to identify confusing ECG segments and adjust decision thresholds. We

Figure 10: ROC curves and confusion matrix for unseen-pathology evaluation (four-class model).

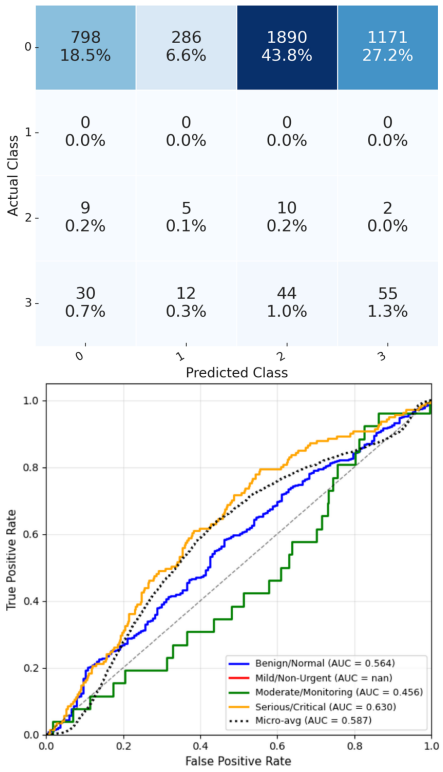
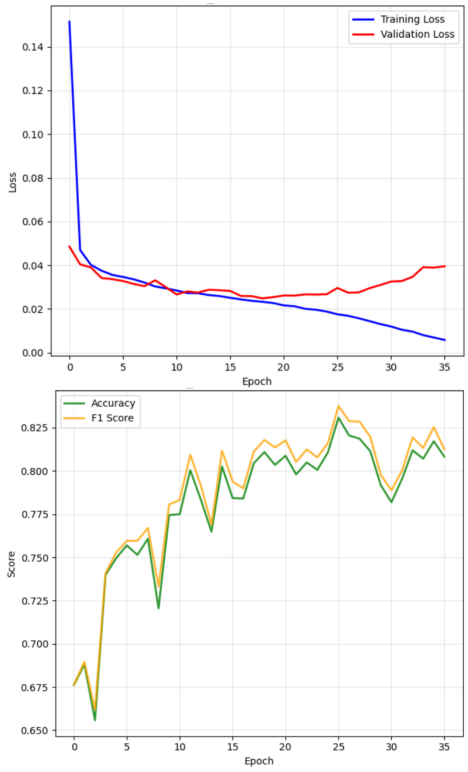


Figure 11: Training and validation loss, accuracy, and F1-score for the Disease ResNet model over 36 epochs.



recommend that, upon deployment, transparent performance metrics—accuracy, confidence distributions and alert volumes—be published so that GPs can determine when to rely on the algorithm and when to apply additional clinical judgment. Furthermore, integrating explainable AI outputs with clinician expertise should be prioritized to enhance patient safety and diagnostic accuracy without displacing human decision-making.

5.3 Limitations and Future Directions

This study highlights the influence of label granularity in ECG classification but also presents several limitations. The dataset's limited size and class diversity—particularly in underrepresented and borderline categories like Moderate—restricts generalizability and amplifies class imbalance issues. Although internal validation was promising, performance declines on external datasets indicate a need for improved robustness.

Future research should explore more targeted fine-tuning and optimization strategies for each model architecture, as this study focused primarily on isolating the effects of label granularity. The current results do not reflect the full performance potential of the examined labeling schemes. Further model-specific enhancements are necessary to better understand how these labeling strategies perform under optimized conditions and how well they generalize to real-world settings.

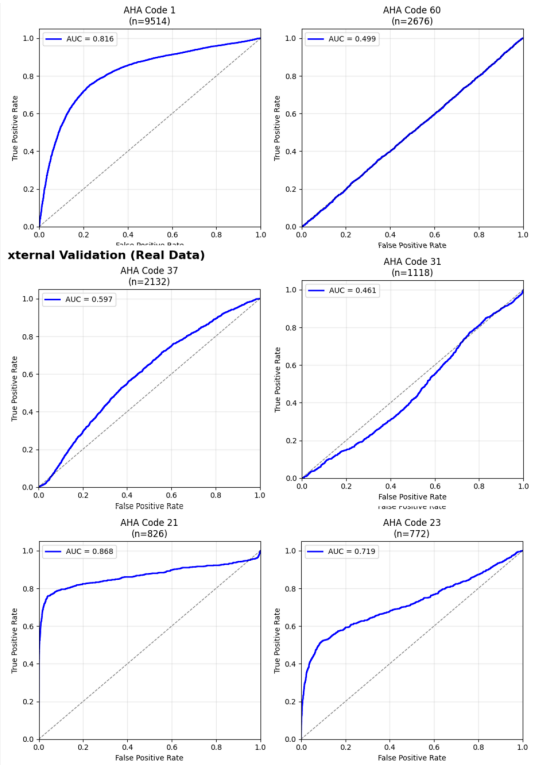
Revisiting the label taxonomy by employing unsupervised or semi-supervised methods could lead to more clinically relevant groupings and help reduce ambiguity. Techniques like focal loss and curriculum learning may effectively tackle class imbalance. Additionally, incorporating metadata such as age, sex, and symptoms could improve diagnostic accuracy, particularly in ambiguous cases.

6 CONCLUSION

This study examined the influence of label granularity on the performance of deep learning models for electrocardiogram (ECG) classification. The findings demonstrate that adopting severity-based, coarse-grained labels leads to substantial improvements in model robustness, generalizability, and clinical applicability, particularly in the context of primary care triage. The three-class ResNet1D model achieved an internal validation accuracy of 87% and a weighted F1-score of 0.87, significantly outperforming models trained on fine-grained diagnostic codes.

Despite this strong internal performance, the model's accuracy decreased to 60.5% on external data and to 57% when tested on previously unseen pathologies. Nevertheless, these outcomes represent a considerable improvement over the four-class diagnostic model, which attained only 52.92% accuracy on known external data and 20.01% on unknown pathologies. Furthermore, the fine-grained model occasionally failed to produce any correct predictions for certain rare diseases, underscoring the limitations of high-resolution diagnostic labels in generalization tasks.

Figure 12: ROC curves for eight representative AHA codes in external validation of the Disease ResNet model.



Overall, the results of this thesis highlight the practical and strategic advantages of label aggregation for developing reliable and interpretable AI systems in healthcare. These insights directly support the ongoing development of Diplora's AI-enhanced Holter devices, which aim to assist general practitioners by facilitating timely and accurate cardiac risk stratification, improving referral decisions, and optimizing clinical workflows.

REFERENCES

[1] Abdallah Abdellatif, Hamdan Abdellatef, Jeevan Kanesan, Chee-Onn Chow, Joon Huang Chuah, and Hassan Muwafaq Gheni. 2022. An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods. *IEEE Access* 10 (2022), 79974–79985. https://doi.org/10.1109/ACCESS.2022.3191669

[2] U. Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Muhammad Adam, Arkadiusz Gertych, and Tan Ru San. 2017. A Deep Convolutional Neural Network Model to Classify Heartbeats. *Computers in Biology and Medicine* 89, 1 (Oct. 2017), 389–396. https://doi.org/10.1016/j.compbiomed.2017.08.022 Epub 24 August 2017.

[3] N. E. Almansouri, M. Awe, S. Rajavelu, K. Jahnavi, R. Shastry, A. Hasan, H. Hasan, M. Lakkimsetti, R. K. AlAbbasi, B. C. Gutiérrez, and A. Haider. 2024. Early diagnosis of cardiovascular diseases in the era of artificial intelligence: An in-depth review. *Cureus* 16, 3 (March 2024), e55869. https://doi.org/10.7759/cureus.55869

[4] American Heart Association. 2022. Heart Disease and Stroke Statistics – At-a-Glance: 2022 Heart Disease and Stroke Statistics Update Fact Sheet. PDF file. https://www.heart.org/-/media/PHD-Files-2/Science-News/2/2022-Heart-and-Stroke-Stat-Update/2022-Stat-Update-At-a-Glance.pdf © 2022 American Heart Association, Inc.; accessed via AHA website.

[5] Yaqoob Ansari, Omar Mourad, Khalid Qaraqe, and Erchin Serpedin. 2023. Deep learning for ECG Arrhythmia detection and classification: an overview of progress for period 2017–2023. *Frontiers in Physiology* 14 (Sep 2023), 1246746. https://doi.org/10.3389/fphys.2023.1246746 Open access (CC BY 4.0).

Figure 13: ROC curves for eight representative AHA codes in external validation of the Disease ResNet model.

[6] Bilal Ashraf, Husan Ali, Muhammad Aseer Khan, and Fahad R. Albogamy. 2025. EffNet: an efficient one-dimensional convolutional neural networks for efficient classification of long-term ECG fragments. *Biomedical Physics & Engineering Express* 11, 2 (2025), 025041. https://doi.org/10.1088/2057-1976/adb58a

[7] A. Bhanu Sri, R. Saivineela, K. Devaki Jyothirmayee, A. R. N. V. S. Vineesha, V. Sai Dinesh, and V. Yuga Nivas. 2023. Heart Disease Detection and Severity Prediction Using ML Techniques. *International Research Journal of Modernization in Engineering, Technology and Science* 5, 4 (2023), 7496–7502. https://www.irjmets.com/uploadedfiles/paper/issue_4_april_2023/35852/final/fin_irjmets1687355866.pdf Peer-Reviewed, Open Access, Fully Refereed.

[8] Lennert Bontinck, Karel Fonteyn, Tom Dhaene, and Dirk Deschrijver. 2024. ECGencode: Compact and computationally efficient deep learning feature encoder for ECG signals. *Expert Systems with Applications* 255 (2024), 124775. https://doi.org/10.1016/j.eswa.2024.124775

[9] Sahar Boulkaboul, Samira Bouchama, Syfax Kasser, and Belkacem Ait Si Ali. 2024. D-Resnet: Deep Resnet-based approach for ECG classification. *Applied Informatics* 26, 1 (2024), 64–71.

[10] Wenjuan Cai, Yundai Chen, Jun Guo, Baoshi Han, Yajun Shi, Lei Ji, Jinliang Wang, Guanglei Zhang, and Jianwen Luo. 2020. Accurate Detection of Atrial Fibrillation from 12-Lead ECG Using Deep Neural Network. *Computers in Biology and Medicine* 116 (2020), 103378. https://doi.org/10.1016/j.compbiomed.2019.103378 Epub 2019 Aug 2.

[11] Chen Chen, Zhengchun Hua, Ruiqi Zhang, Guangyuan Liu, and Junye Pharmaceutical Ltd. 2020. Automated arrhythmia classification based on a combination network of CNN and LSTM. *Biomedical Signal Processing and Control* 57 (March 2020), 101819. https://doi.org/10.1016/j.bspc.2019.101819

[12] Zhuo Chen, Ruizhou Ding, Ting-Wu Chin, and Diana Marculescu. 2018. Understanding the Impact of Label Granularity on CNN-Based Image Classification. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. 895–904. https://doi.org/10.1109/ICDMW.2018.00131

[13] S. S. Diware, S. Dash, A. B. Gebregiorgis, R. V. Joshi, C. Strydis, S. Hamdioui, and R. K. Bishnoi. 2023. Severity-based hierarchical ECG classification using neural

networks. *IEEE Transactions on Biomedical Circuits and Systems* 17, 1 (2023), 77–91. https://doi.org/10.1109/TBCAS.2023.3242683

[14] Wei Fan, Yujuan Si, Weiyi Yang, and Meiqi Sun. 2022. Imbalanced ECG data classification using a novel model based on active training subset selection and modified broad learning system. *Measurement* 195 (2022), 111412. https://doi.org/10.1016/j.measurement.2022.111412

[15] Marc Goettling, Alexander Hammer, Hagen Malberg, and Martin Schmidt. 2024. xECGArch: a trustworthy deep learning architecture for interpretable ECG analysis considering short-term and long-term features. *Scientific Reports* 14, 1 (Jun 2024), 13122. https://doi.org/10.1038/s41598-024-63656-x

[16] V. Jahmunah, E. Y. K. Ng, Ru-San Tan, Shu Lih Oh, and U. Rajendra Acharya. 2022. Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals. *Computers in Biology and Medicine* 146 (Jul 2022), 105550. https://doi.org/10.1016/j.compbiomed.2022.105550

[17] Mudassar Khalid, Charnchai Pluempitiwiriyawej, Somkiat Wangsiripitak, Ghulam Murtaza, and Abdulkadhem Abdulkadhem. 2024. The Applications of Deep Learning in ECG Classification for Disease Diagnosis: A Systematic Review and Meta-Data Analysis. *Engineering Journal* 28 (08 2024), 45–77. https://doi.org/10.4186/ej.2024.28.8.45

[18] Jiewei Lai, Huixin Tan, Jinliang Wang, Lei Ji, Jun Guo, Baoshi Han, Yajun Shi, Qianjin Feng, and Wei Yang. 2023. Practical intelligent diagnostic algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset. *Nature Communications* 14, 1 (June 2023), 3741. https://doi.org/10.1038/s41467-023-39472-8 Open access; Published 23 June 2023.

[19] Zheng-Xuan Li, Ying-Shao Hsu, Po-Yung Chou, and Cheng-Hung Lin. 2025. ECG Signal Classification Using 1D ResNet-18 with Integrated CBAM and Auxiliary Classifier. (2025), 1–4. https://doi.org/10.1109/ICCT-Pacific63901.2025.11012832

[20] Hui Liu, Dan Chen, Da Chen, Xiyu Zhang, Huijie Li, Lipan Bian, Minglei Shu, and Yinglong Wang. 2022. A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. *Scientific Data* 9 (2022), 272. https://doi.org/10.1038/s41597-022-01403-5 Data Descriptor; Open access; Published: 07 June 2022.

[21] Seth S. Martin, Aaron W. Aday, Zaid I. Almarzooq, Cheryl A. M. Anderson, Pankaj Arora, Christy L. Avery, Carissa M. Baker-Smith, Bethany B. Barone Gibbs, Andrea Z. Beaton, Amelia K. Boehme, and et al. 2024. Heart Disease and Stroke Statistics—2024 Update: A Report of U.S. and Global Data From the American Heart Association. *Circulation* 149, 8 (2024), e347–e913. https://doi.org/10.1161/CIR.0000000000001209 https://doi.org/10.1161/CIR.0000000000001209.

[22] M. Martínez-Sellés and M. Marina-Breysse. 2023. Current and future use of artificial intelligence in electrocardiography. *Journal of Cardiovascular Development and Disease* 10, 4 (2023), 175. https://doi.org/10.3390/jcdd10040175

[23] Hemaxi Narotamo, Mariana Dias, Ricardo Santos, André V. Carreiro, Hugo Gamboa, and Margarida Silveira. 2024. Deep learning for ECG classification: A comparative study of 1D and 2D representations and multimodal fusion approaches. *Biomedical Signal Processing and Control* 93 (Jul 2024), 106141. https://doi.org/10.1016/j.bspc.2024.106141

[24] Jianhui Peng, Ao Ran, Chenjin Yu, and Huafeng Liu. 2024. EGCNet: a hierarchical graph convolutional neural network for improved classification of electrocardiograms. *EURASIP Journal on Advances in Signal Processing* 2024, 1 (2024), Article 93. https://doi.org/10.1186/s13634-024-01187-3

[25] Eedara Prabhakararao and Samarendra Dandapat. 2020. Myocardial Infarction Severity Stages Classification From ECG Signals Using Attentional Recurrent Neural Network. *IEEE Sensors Journal* 20, 15 (2020), 8711–8720. https://doi.org/10.1109/JSEN.2020.2984493

[26] Tariq Sadad, Mejdl Safran, Inayat Khan, Sultan Alfarhood, Razaullah Khan, and Imran Ashraf. 2023. Efficient Classification of ECG Images Using a Lightweight CNN with Attention Module and IoT. *Sensors* 23, 18 (2023), 7697. https://doi.org/10.3390/s23187697

[27] Nizar Sakli, Haifa Ghabri, Soufiene Othman, Faris Almalki, Hedi Sakli, Obaid Ali, and Mustapha Najjari. 2022. ResNet-50 for 12-Lead Electrocardiogram Automated Diagnosis. *Computational Intelligence and Neuroscience* 2022 (Apr 2022), 7617551. https://doi.org/10.1155/2022/7617551

[28] Sanshiro Togo, Yuki Sugiura, Sayumi Suzuki, Kazuto Ohno, Keitaro Akita, Kenichiro Suwa, Shin-ichi Shibata, Michio Kimura, and Yuichiro Maekawa. 2023. Model for classification of heart failure severity in patients with hypertrophic cardiomyopathy using a deep neural network algorithm with a 12-lead electrocardiogram. *Open Heart* 10, 2 (Dec. 6 2023), e002414. https://doi.org/10.1136/openhrt-2023-002414

[29] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Felix I. Lunze, Wojciech Samek, and Tobias Schaeffter. 2020. PTB-XL: A Large Publicly Available ECG Dataset. *Scientific Data* 7, 1 (2020), 154. https://doi.org/10.1038/s41597-020-0495-6

[30] Guanhua Yang, Shuxin Zou, Huaiqin Qin, Yuhan Cao, Zhen Zhang, and Xiaoming Deng. 2025. Robust 12-Lead ECG Classification with Lightweight ResNet: An Adaptive Second-Order Learning Rate Optimization Approach. *Electronics* 14, 10 (Oct 2025), 1941. https://doi.org/10.3390/electronics14101941