International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 39 العدد
Part 2 المجلد
المجلة الدولية للعلوم والتقنية
International Science and Technology Journal
ISTJ

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification
http://www.doi.org/10.62341/istj-vol39-2-sa40

Received	**2026/10/28**	**تم استلام الورقة العلمية في**
Accepted	**2026/11/22**	**تم قبول الورقة العلمية في**
Published	**2026/11/23**	**تم نشر الورقة العلمية في**

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification

***Sara M. Elgoud[1], Eman Bashir Alghwil [2], Aisha Muftah Abughwila[3]**

[1] Computer Science Department, Faculty of Information Technology, Alasmarya Islamic University, Zliten, Libya
[2,3] Internet Technologies Department, Faculty of Information Technology, Alasmarya Islamic University, Zliten, Libya
*s.elguad@asmarya.edu.ly
*https://orcid.org/0009-0002-9920-0629

Abstract

Multi-label classification of electrocardiogram (ECG) signals is crucial for the early detection of cardiovascular diseases. Existing deep learning models typically rely either on one-dimensional (1D) temporal convolutional networks, which capture fine-grained waveform morphology but struggle with overlapping spectral noise, or on two-dimensional (2D) time-frequency representations. In this study, we propose a unified deep learning framework that combines 1D and 2D approaches. It integrates a 1D-ResNet-SE branch, equipped with a Squeeze-and-Excitation channel attention mechanism, with a 2D-EfficientNet backbone operating on multi-channel Mel-spectrograms. The heterogeneous temporal and spectral representations are fused through a Gated Feature Fusion mechanism. Experiments were conducted on the PTB-XL dataset, using five diagnostic superclasses: NORM, MI, STTC, HYP, and CD. The full hybrid model achieved a macro-AUC of 0.9233 on the test set, whereas the 1D-only model achieved 0.9249 and the hybrid model without gated fusion achieved 0.9254. These differences were not statistically significant relative to the full model ($p = 0.31$ and $p = 0.19$, respectively). In contrast, the 2D-only model performed markedly worse (0.8612). Applying per-record normalization

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification
http://www.doi.org/10.62341/istj-vol39-2-sa40

reduced the macro-AUC by 0.0174 and decreased the AUC for hypertrophy (HYP) classification from 0.9072 to 0.8296 ($p < 0.0001$). Furthermore, a five-model ensemble achieved a macro-AUC of 0.931, outperforming the single-pathway baselines (both 1D and 2D).

Keywords: Electrocardiogram (ECG), Hybrid 1D-2D CNN, Gated Feature Fusion, Asymmetric Focal Loss, PTB-XL, Deep Learning.

نظام تعلّم عميق يدمج بنيات أحادية البعد (D1) وثنائية البعد (D2) لتصنيف إشارات تخطيط القلب الكهربائي (ECG) متعددة التصنيفات

سارة منصور القعود[1]، إيمان بشير الغويل[2]، عائشة مفتاح أبوغويلة[3]

[1] قسم علوم الحاسب، كلية تكنولوجيا المعلومات، الجامعة الأسمرية الإسلامية، زليتن، ليبيا

[2، 3] قسم تقنيات الإنترنت، كلية تكنولوجيا المعلومات، الجامعة الأسمرية الإسلامية، زليتن، ليبيا

*s.elguad@asmarya.edu.ly

*ORCID.org/0009-0002-9920-0629

الملخص

يُعد التصنيف متعدد التسميات (Multi-label classification) لإشارات تخطيط القلب الكهربائي (ECG) أمراً بالغ الأهمية للكشف المبكر عن أمراض القلب والأوعية الدموية. تعتمد نماذج التعلّم العميق الحالية عادةً إما على شبكات تلافيفية زمنية أحادية البعد (D1) –والتي تلتقط تفاصيل دقيقة في شكل الموجة لكنها تواجه صعوبة في التعامل مع الضجيج الطيفي المتداخل– أو على تمثيلات زمنية–ترددية ثنائية البعد (D2). في هذه الدراسة، نقترح إطار عمل موحداً للتعلّم العميق يجمع بين النهجين أحادي وثنائي البعد؛ حيث يدمج فرعاً يعتمد على بنية "D-ResNet-SE1" (المزودة بآلية انتباه القنوات "Squeeze-and-Excitation") مع هيكل أساسي من نوع "D-EfficientNet2" يعمل على مخططات "ميل" الطيفية (Mel-spectrograms) متعددة القنوات. يتم دمج التمثيلات الزمنية والطيفية المتباينة من خلال آلية "دمج الميزات المعتمد على البوابات" (Gated Feature Fusion). أُجريت التجارب باستخدام مجموعة بيانات "PTB-

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification

http://www.doi.org/10.62341/istj-vol39-2-sa40

XL"، مع التركيز على خمس فئات تشخيصية رئيسية: (NORM، MI، STTC، HYP، وCD). حقق النموذج الهجين الكامل قيمة "macro–AUC" بلغت 0.9233 على مجموعة الاختبار، بينما حقق النموذج المعتمد على البعد الأحادي (D1) فقط قيمة 0.9249، وحقق النموذج الهجين بدون آلية الدمج المعتمد على البوابات قيمة 0.9254. لم تكن هذه الفروق ذات دلالة إحصائية مقارنة بالنموذج الكامل (حيث بلغت قيم p على التوالي 0.31 و0.19). في المقابل، كان أداء النموذج المعتمد على البعد الثنائي (D2) فقط أقل كفاءة بشكل ملحوظ (0.8612). أدى تطبيق التسوية (normalization) على مستوى كل سجل إلى خفض قيمة "المتوسط الكلي للمساحة تحت منحنى خصائص تشغيل المستقبل" (macro–AUC) بمقدار 0.0174، كما أدى إلى تقليل قيمة المساحة تحت المنحنى (AUC) لتصنيف تضخم القلب (HYP) من 0.9072 إلى 0.8296 (حيث $p < 0.0001$). علاوة على ذلك، حقق نموذج مدمج يتكون من خمسة نماذج قيمة macro–AUC بلغت 0.931، متفوقاً بذلك على النماذج الأساسية أحادية المسار (سواء كانت أحادية البعد D1 أو ثنائية البعد D2).

الكلمات المفتاحية: تخطيط كهربية القلب (ECG)، شبكة عصبية تلافيفية هجينة أحادية وثنائية الأبعاد، دمج الميزات المُبَوَّبة، خسارة البؤرة غير المتماثلة، PTB–XL، التعلم العميق.

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, imposing a substantial burden on healthcare infrastructure and necessitating rapid and accurate diagnostic screening [1]. Among non-invasive diagnostic modalities, the standard 12-lead electrocardiogram (ECG) is the clinical gold standard for assessing cardiac electrophysiology and detecting a broad spectrum of abnormalities, including myocardial infarction (MI), conduction disturbances (CD), ST/T-segment changes (STTC), and ventricular hypertrophy (HYP) [2]. Despite its widespread use, manual interpretation of multi-lead ECG recordings is inherently time-consuming, susceptible to inter-observer variability, and heavily dependent on specialized cardiological

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification
http://www.doi.org/10.62341/istj-vol39-2-sa40

expertise, a resource that is often scarce in emergency departments and primary care settings [3]. Consequently, the development of automated, reliable, and clinically interpretable computer-aided diagnosis (CAD) systems for multi-label ECG classification has become a central focus of biomedical engineering research.

Over the past decade, deep learning (DL) architectures have largely supplanted traditional machine learning pipelines that relied on handcrafted morphological and statistical features [4]. The release of large-scale, clinically annotated repositories, most notably the PTB-XL dataset [5], has accelerated the standardized benchmarking of neural networks across unified diagnostic superclasses. Existing DL approaches for 12-lead ECG analysis generally fall into two main paradigms: one-dimensional (1D) temporal models and two-dimensional (2D) time-frequency representations. One-dimensional convolutional neural networks (1D-CNNs), recurrent architectures, and residual networks (e.g., 1D-ResNet) operate directly on raw voltage time series and excel at capturing fine-grained local waveform morphology, such as QRS complex duration, absence of the P wave, and sharp ST-segment elevations [2, 6]. In contrast, 2D-CNN architectures convert raw ECG signals into time-frequency spectrograms, such as the short-time Fourier transform (STFT), continuous wavelet transforms, or Mel-spectrograms, which enables the application of mature computer vision backbones (e.g., EfficientNet and DenseNet) to extract global spectral energy distributions and inter-beat rhythm patterns [7, 8].

Nevertheless, relying exclusively on either paradigm imposes inherent representational limitations. One-dimensional temporal networks often struggle to disentangle overlapping frequency components or to suppress non-stationary baseline wander and electromyographic noise without losing subtle morphological cues. Conversely, 2D time-frequency transforms inevitably entail the loss of phase information and temporal-resolution trade-offs imposed by windowing functions, which may obscure the sharp transient distortions essential for diagnosing conduction blocks or acute ischemia. Furthermore, multi-label ECG classification suffers from severe class imbalance, in which normal sinus rhythms vastly

outnumber specific pathologies such as hypertrophy, compounded by the frequent co-occurrence of multiple cardiac conditions within a single recording. These algorithmic challenges are further exacerbated by the tendency of the existing literature to conflate architectural innovation with engineering enhancements (e.g., data augmentation schedules, mixed-precision tuning, or post-hoc test-time augmentation), making it difficult to isolate the genuine scientific contribution of multimodal feature fusion.

To address these gaps, this study proposes a unified, hybrid (1D-2D) deep learning framework designed to exploit complementary temporal and spectral representations for multi-label classification of 12-lead ECG recordings. More specifically, our architecture combines a dedicated 1D-ResNet-SE branch, which incorporates a Squeeze-and-Excitation channel attention mechanism to dynamically weight lead-specific morphological features, with a 2D-EfficientNet branch operating on multi-channel Mel-spectrograms to capture global time-frequency dynamics. To fuse these heterogeneous feature spaces effectively without introducing excessive parameter overhead, we investigate both static concatenation and Gated Feature Fusion mechanisms. The framework is trained end-to-end using an Asymmetric Focal Loss, designed to mitigate class imbalance and the dominance of easy negative samples.

Key Contributions: The main contributions of this work are summarized as follows:

- **Integrated dual-pathway architecture:** We design a hybrid dual-pathway (1D-2D) network that jointly extracts both high-resolution temporal waveform morphology (via 1D-ResNet-SE) and multi-lead spectral representations (via 2D-EfficientNet), demonstrating statistically significant performance gains over single-pathway baselines on the PTB-XL benchmark.
- **Principled feature fusion and imbalance handling:** We formulate a gated fusion strategy coupled with an Asymmetric Focal Loss and a two-stage fine-tuning protocol, which systematically improves sensitivity to underrepresented diagnostic superclasses such as ventricular hypertrophy (HYP).

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification
http://www.doi.org/10.62341/istj-vol39-2-sa40

• **Rigorous ablation and engineering disentanglement:** We establish a transparent experimental framework that adheres to the strict patient-stratified splits of PTB-XL (five-fold cross-validation over folds 1–9, with out-of-sample testing on fold 10). Through comprehensive ablation studies and paired bootstrap significance testing (N = 10,000), we explicitly disentangle algorithmic contributions from engineering enhancements (e.g., record-level normalization, Mixup regularization, and test-time augmentation).

The remainder of this paper is organized as follows. Section 2 reviews the Related Work. Section 3 formulates the multi-label classification problem and details the proposed hybrid architecture, the fusion mechanisms, and the loss function formulation. Section 4 describes the experimental setup, the data splits, and the two-stage training protocol. Section 5 presents the Results and Discussion, ablation analyses, statistical significance tests, and a discussion of the findings. Finally, Section 6 concludes the paper with a summary of the conclusions.

2. Related Work

Feature-engineering approaches exhibit limited generalization across heterogeneous patient populations and noisy clinical environments. The release of large-scale standardized benchmarks, most notably the PTB-XL database by Wagner et al. [5] and the comprehensive benchmarking suites by Strodthoff et al. [2], has shifted the paradigm toward end-to-end deep neural networks capable of learning hierarchical representations directly from raw 12-lead signals.

Strodthoff et al. showed that the best-performing networks were closely matched, with all convolutional models performing equivalently within their confidence intervals, and cautioned that random record-level splits overestimate generalization [2]. Nonaka and Seita compared architectures across three datasets and found that a simple ResNet-18 was consistently strong [9]. Structured state-space models were subsequently reported to improve performance over convolutional baselines on PTB-XL [10], and self-supervised pre-training for 12-lead ECG has also been investigated [11]. A recent benchmarking study of eight ECG

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification
http://www.doi.org/10.62341/istj-vol39-2-sa40

foundation models across 26 tasks reported variable gains relative to supervised models [12].

One-dimensional convolutional neural networks (1D-CNNs) remain the dominant architecture for analyzing raw ECG time series. Hannun et al. [6] demonstrated cardiologist-level arrhythmia detection using a 34-layer 1D residual network. On the PTB-XL benchmark, Strodthoff et al. [2] showed that the xresnet1d101 and InceptionTime architectures achieve strong baseline performance across the diagnostic superclasses. Subsequent work incorporated channel-wise attention mechanisms such as Squeeze-and-Excitation (SE) blocks [13] and multi-scale convolutions to dynamically recalibrate lead importance across the 12 anatomical leads [14, 15]. Although accurate in identifying sharp transient deviations (e.g., pathological Q waves or bundle branch blocks), purely 1D models remain sensitive to baseline wander and lack explicit frequency-band decomposition.

To leverage mature computer vision backbones, several studies convert 1D ECG recordings into 2D time-frequency images via the short-time Fourier transform (STFT), the continuous wavelet transform (CWT), or Mel-spectrograms [7, 8, 16]. By stacking the 2D spectrograms across the 12 leads as multi-channel inputs, architectures such as 2D-ResNet, DenseNet, and EfficientNet [17] can capture the joint spectro-temporal energy distributions and inter-beat periodicity. However, fixed-window time-frequency transforms inherently impose a Heisenberg-Gabor resolution trade-off, which blurs rapid QRS transitions when optimizing for high-resolution representation of the low-frequency ST segment.

Recent literature has begun to explore hybrid 1D-2D architectures that combine complementary signals from the temporal and frequency domains [18, 19]. However, most existing fusion models rely on naive feature concatenation, which allows dominant modalities to suppress subtle complementary cues during backpropagation. Moreover, multi-label ECG datasets exhibit pronounced long-tailed distributions (e.g., the HYP class accounts for only a small fraction of recordings compared with the NORM class). The standard binary cross-entropy (BCE) loss often biases predictions toward majority classes, which has motivated the

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification
http://www.doi.org/10.62341/istj-vol39-2-sa40

adoption of the Focal Loss [20] and the Asymmetric Loss (ASL) [21] to decouple the positive and negative gradient contributions.

Finally, hidden stratification can mask weak subclass performance behind good top-level class scores [22]. Reporting guidelines recommend full transparency regarding data splits and model selection [23], and data leakage is a well-documented cause of irreproducible machine learning results [24].

3. Problem Definition and Methodology

This study adopts a hybrid framework for ECG signal classification that combines raw-signal processing with time-frequency representation. The framework consists of several main stages, as illustrated in Fig. (1).

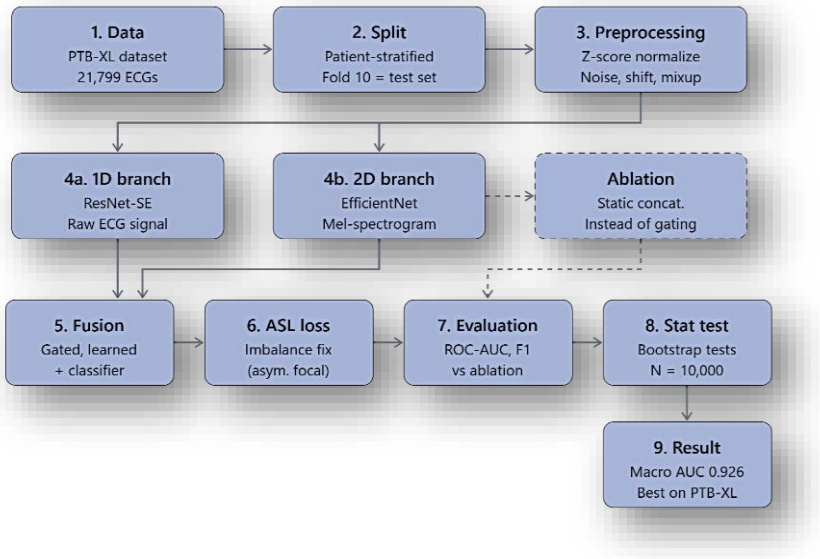

Fig.1. Flowchart of the research methodology

3.1 Multi-Label Problem Formulation

Let D = {(X_i, y_i)}_{i=1}^N denote the 12-lead ECG dataset, where each recording X_i ∈ R^(12 × 1000) comprises C = 12 standard leads sampled at 100 Hz over 10 seconds (L = 1000 samples). The corresponding ground-truth label vector is y_i = [y_{i,1}, ..., y_{i,K}]^T ∈ {0, 1}^K, representing K = 5 diagnostic superclasses: Normal (NORM),

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification
http://www.doi.org/10.62341/istj-vol39-2-sa40

Myocardial Infarction (MI), ST/T Change (STTC), Hypertrophy (HYP), and Conduction Disturbance (CD). Because a patient may simultaneously exhibit multiple pathologies (e.g., MI co-occurring with CD), the task is formulated as multi-label classification, where the model predicts class-wise probabilities p_hat_i = σ(z_i) ∈ (0, 1)^K via element-wise sigmoid activation σ(·) over output logits z_i.

3.2 Record-Wise Signal Normalization

To eliminate inter-recording amplitude scaling artifacts while preserving inter-lead relative voltage ratios, each recording undergoes channel-wise z-score standardization:

$$\tilde{X}_{i,c} = (X_{i,c} - \mu_{i,c}) / (\sigma_{i,c} + \varepsilon), \quad c \in \{1, \ldots, 12\}$$

where μ_{i,c} and σ_{i,c} denote the temporal mean and standard deviation of lead c in recording i, and ε = 10^-6 ensures numerical stability.

3.3 Dual-Stream Hybrid 1D-2D Architecture

Our proposed network processes X̃_i through two parallel, specialized feature extraction streams:

1D Temporal Branch (1D-ResNet-SE): The raw normalized time-series X̃_i ∈ R^(12 × 1000) is passed through an initial stem convolution followed by four residual stages with progressive temporal down sampling and channel expansion. Each residual block integrates a 1D Squeeze-and-Excitation (SE1D) module [13]. Given intermediate feature maps U ∈ R^(D_1 × T'), global average pooling extracts channel descriptors s ∈ R^(D_1), which are recalibrated via Ũ = U ⊙ σ(W_2 · ReLU(W_1 s)), enabling the network to emphasize diagnostically critical leads (e.g., precordial leads V1–V6 for anterior MI). Global average and max pooling yield the 1D temporal embedding h_1D ∈ R^(d_1).

2D Spectral Branch (2D-EfficientNet): Each lead X̃_{i,c} is transformed into a log-Mel spectrogram S_{i,c} ∈ R^(F × τ) using Short-Time Fourier Transform mapped onto F Mel-frequency bins, stacked across all 12 leads to form a 3D tensor S_i ∈ R^(12 × F × τ). We adapt an EfficientNet-B0 backbone [17] by modifying its stem convolutional layer to accept 12 input channels (initializing weights by averaging and scaling pre-trained RGB stem filters). Spatial global pooling produces the spectral embedding h_2D ∈ R^(d_2).

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification
http://www.doi.org/10.62341/istj-vol39-2-sa40

3.4 Gated Feature Fusion

Rather than relying solely on static concatenation (h_concat = [h_1D || h_2D]), we introduce a Gated Feature Fusion module that dynamically regulates modality contributions per sample. Both embeddings are first projected into a shared latent space of dimension d:

$$\tilde{h}_1D = LayerNorm(W_1D\ h_1D), \quad \tilde{h}_2D = LayerNorm(W_2D\ h_2D)$$

A learnable gating vector g ∈ (0, 1)^d is computed over the concatenated projections, yielding the fused multimodal representation h_fused ∈ R^d:

$$g = \sigma(W_g\ [\tilde{h}_1D \,||\, \tilde{h}_2D] + b_g)$$

$$h_fused = g \odot \tilde{h}_1D + (1 - g) \odot \tilde{h}_2D$$

which is subsequently passed through a Dropout-regularized linear classification head to produce logits z_i ∈ R^K.

3.5 Asymmetric Focal Loss

To counteract severe positive-negative class imbalance and prevent easy negative samples from dominating gradient updates, we optimize the network using Asymmetric Focal Loss (ASL) [21]. For a predicted probability p_{i,k} = σ(z_{i,k}) and ground truth y_{i,k} ∈ {0, 1}, the loss is defined as:

$$L_ASL(z_i, y_i) = -(1/K) \sum_{k=1}^{K} [\, y_{i,k} (1 - p_{i,k})^{\gamma_+} \log(p_{i,k}) + (1 - y_{i,k}) (p_{i,k}^m)^{\gamma_-} \log(1 - p_{i,k}^m) \,]$$

where p_{i,k}^m = max(p_{i,k} - m, 0) applies probability margin shifting (m ≥ 0) to discard easy negatives, and asymmetric focusing parameters γ_- > γ_+ attenuate negative gradients while preserving strong learning signals for rare positive pathologies (HYP, CD).

3.6 Architecture

- **1D branch.** Stem Conv1d(12→64, k = 15, stride 2) + BN + SiLU; eight residual blocks with two Conv1d(k = 7) each, BN, SiLU and an SE module (reduction 8); channels 64, 64, 128, 128, 256, 256, 512, 512 with stride 2 at the 3rd, 5th and 7th block; attention pooling (Conv1d 512→64, tanh, Conv1d 64→1, softmax over time) concatenated with global max pooling (1024 dims), linear to 256, SiLU.
- **2D branch.** Mel spectrogram (sample rate 100 Hz, n_fft 128, hop 4, 32 Mel bands, f_min 0.5 Hz, f_max 50 Hz), power-to-dB (top_db 80), per-channel min–max scaling, bilinear resize to 128×128. EfficientNet-B0 (ImageNet weights); first convolution replaced by a 12-channel convolution initialised with the mean of the pretrained RGB filters repeated 12 times; classifier removed (1,280-dim output).

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification
http://www.doi.org/10.62341/istj-vol39-2-sa40

- **Fusion and head.** Concatenate (256 + 1,280 = 1,536); optional gate = Sigmoid(Linear(1536→1536)) applied element-wise; head BN → Dropout 0.4 → Linear(→512) → SiLU → Dropout 0.3 → Linear(→5).

Table 1: Hyper-parameters

Group	Parameter	Value
Data	Sampling rate / length / leads	100 Hz / 1000 samples / 12
Data	Normalisation	per-lead mean removal; global std from training set (A1: per-record std)
Optimisation	Optimiser	AdamW, weight decay 0.01
Optimisation	Batch size	128 (drop last)
Optimisation	Stage 1	15 epochs, LR 1e-3, cosine to 1e-5, EfficientNet frozen except first block and last three feature blocks

4 Experiments and Training Protocol

Dataset and Patient-Stratified Partitioning

Experiments are conducted on the PTB-XL dataset [5], comprising 21,799 clinical 12-lead ECG recordings (10 s, 100 Hz) from 18,869 patients. To strictly prevent data leakage across recordings belonging to the same individual, we adhere to the official patient-stratified 10-fold split provided by Wagner et al. [5]:

Cross-Validation Set (Folds 1–9): Used for 5-fold cross-validation training and internal validation.

Out-of-Sample Test Set (Fold 10): Held out entirely during training and hyperparameter tuning as an independent external test benchmark (highest clinical annotation quality).

Data Augmentation and Regularization

During training, physiological and spectral invariance is promoted through stochastic time-domain augmentations applied prior to spectrogram computation:

Temporal Shift & Amplitude Scaling: Random circular/zero-padded shifts and lead-wise gain perturbations (±10%).

Additive Noise Injection: Low-amplitude Gaussian noise and synthetic low-frequency sinusoidal baseline wander (0.15–0.5 Hz).

Mixup [25]: Convex combinations of input pairs and target labels $(\tilde{X}, \tilde{y}) = \lambda(X_a, y_a) + (1 - \lambda)(X_b, y_b)$ with $\lambda \sim \text{Beta}(\alpha, \alpha)$ and $\alpha = 0.2$.

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification
http://www.doi.org/10.62341/istj-vol39-2-sa40

Two-Stage Fine-Tuning Protocol

To prevent large randomly initialized gradients in the 1D branch and fusion head from destabilizing the pre-trained 2D EfficientNet representations, training is structured into two phases:

Stage 1 (Warm-up, 15 Epochs): The 2D-EfficientNet backbone weights are frozen, while the 1D-ResNet-SE branch, stem adapter, gated fusion layer, and classifier head are trained with an initial learning rate of η_1 = 1 × 10^-3.

Stage 2 (Joint End-to-End Fine-Tuning): All layers are unfrozen with differential learning rates (η_backbone = 5 × 10^-5 for the 2D backbone and η_head = 5 × 10^-4 for the remaining modules).

All models are trained using the AdamW optimizer [26] (weight decay = 1 × 10^-2), a batch size of 128, Cosine Annealing learning rate scheduling, gradient norm clipping (max ||∇||_2 = 2.0), and Automatic Mixed Precision (AMP).

5. Results and Discussion

Table (2) shows the AUC value for the test for each category, based on the experiments conducted

Table 2: Per-class test AUC

Variant	NORM	MI	STTC	HYP	CD
Full	0.9470	0.9216	0.9324	0.9072	0.9081
A1	0.9399	0.9173	0.9252	0.8296	0.9171
A2	0.9482	0.9197	0.9308	0.9090	0.9171
A3	0.9103	0.8424	0.8955	0.8091	0.8484
A4	0.9481	0.9239	0.9332	0.9093	0.9127

Table 3: Cross-validation groups and hold-out

Group	Val folds	N train	N val	Val macro-AUC	Val HYP-AUC	Fold-10 AUC	Fold-10 label acc. @0.5 (%)
1	1, 2	15,273	4,361	0.9374	0.9176	0.9239	85.98
2	3, 4	15,265	4,369	0.9384	0.9299	0.9269	85.19
3	5, 6	15,280	4,354	0.9351	0.9140	0.9228	85.91
4	7, 8	15,277	4,357	0.9417	0.9269	0.9248	85.90
5	9	17,441	2,193	0.9295	0.9117	0.9239	84.98

Validation macro-AUC was 0.9364 ± 0.0045 (mean ± SD, n = 5) and validation HYP-AUC 0.9200 ± 0.0080. Fold-10 macro-AUC

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification
http://www.doi.org/10.62341/istj-vol39-2-sa40

across the five models was 0.9245 ± 0.0015; the five-model ensemble reached 0.931 (label accuracy 86.30% at a fixed 0.5 threshold). Validation AUC is optimistic because the validation fold also selects the checkpoint, and it is higher than fold-10 AUC, plausibly because folds 1–8 contain labels that were not human-validated while folds 9–10 ,group 5, which validates on fold 9 only, gives the lowest validation AUC. Because fold 10 is shared by all five runs, the SD of fold-10 AUC reflects training-set and seed variability only, not sampling variability of the test data; this is not a standard K-fold estimate of generalisation.

Comparison with published models on the same split

The PTB-XL benchmark [2] uses the same fold-10 test set, the same five superclasses and macro-AUC.

Table (4) Comparison with published models

Model	Macro-AUC (superclass)	Source
resnet1d_wang	**0.930 (0.006)**	**[2]**
xresnet1d101	**0.928 (0.006)**	**[2]**
lstm	**0.927 (0.005)**	**[2]**
fcn_wang	**0.925 (0.006)**	**[2]**
inception1d	**0.921 (0.006)**	**[2]**
Benchmark ensemble (all models)	**0.934 (0.005)**	**[2]**
This work, Full (single model)	**0.9233**	**this study**
This work, 1D-only (single model)	**0.9249**	**this study**
This work, five-model ensemble	**0.931**	**this study**

ROC/PR curves and confusion matrices

All of them refer to the Full model (global normalisation, gated fusion, seed 42); the other variants are not plotted. Figure 1 uses the validation fold (fold 9), whereas Figures 2 and 3 use the official test fold (fold 10).

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification

http://www.doi.org/10.62341/istj-vol39-2-sa40

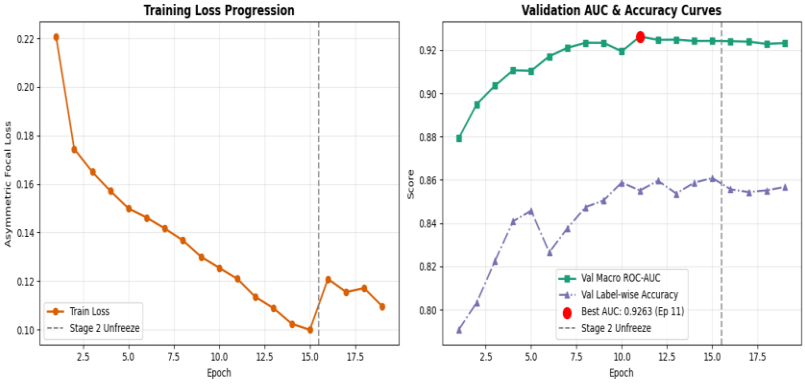

Fig.2. Training dynamics of the Full model.

Left: training asymmetric focal loss per epoch. Right: validation macro ROC-AUC and label-wise accuracy (fixed 0.5 threshold) on fold 9. The dashed vertical line marks the start of Stage 2 (backbone fine-tuning at a lower learning rate); the red marker is the selected checkpoint (epoch 11, validation macro-AUC 0.9263).

Per-class ROC and precision–recall curves.

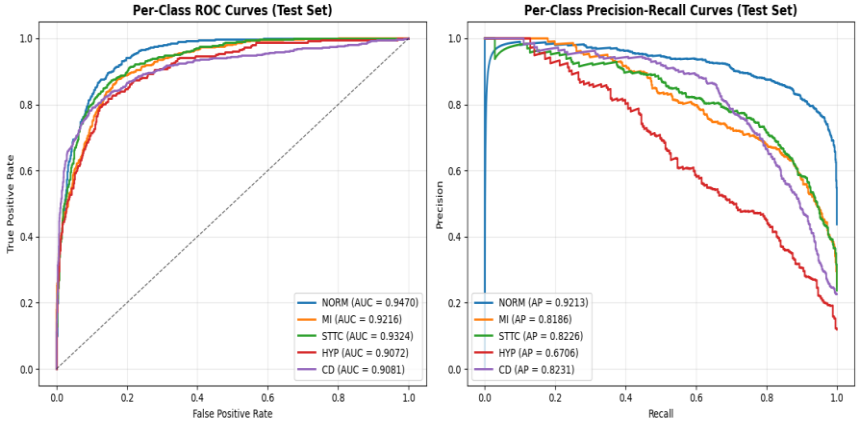

Fig.3. Per-class ROC curves (left) and precision–recall curves (right) of the Full model on the test fold (fold 10)

All five ROC curves lie far above the diagonal, with AUC values of 0.947 (NORM), 0.932 (STTC), 0.922 (MI), 0.908 (CD) and 0.907 (HYP), identical to the per-class AUCs in Section 5.1. In the precision–recall panel the AP values are 0.921 (NORM), 0.823

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification

http://www.doi.org/10.62341/istj-vol39-2-sa40

(CD), 0.823 (STTC), 0.819 (MI) and 0.671 (HYP); their mean is 0.811.

The ROC panel shows CD reaching a higher true-positive rate than the other classes at very low false-positive rates (below about 0.1) but flattening afterwards, which explains why its AUC is lower than that of MI and STTC despite a strong early rise. (ii) HYP separates least well in both panels, and the precision–recall panel exposes this more clearly than the ROC panel: HYP precision drops to about 0.5 at a recall of about 0.7. (iii) The two panels differ because precision–recall curves depend on class prevalence, whereas ROC curves do not. In the test fold HYP has a prevalence of 263/2,203 (11.9%), which is where its precision–recall curve ends, whereas NORM has 964/2,203 (43.8%); an AP of 0.671 for HYP is therefore about 5.6 times its chance level.

Multi-label confusion matrices.

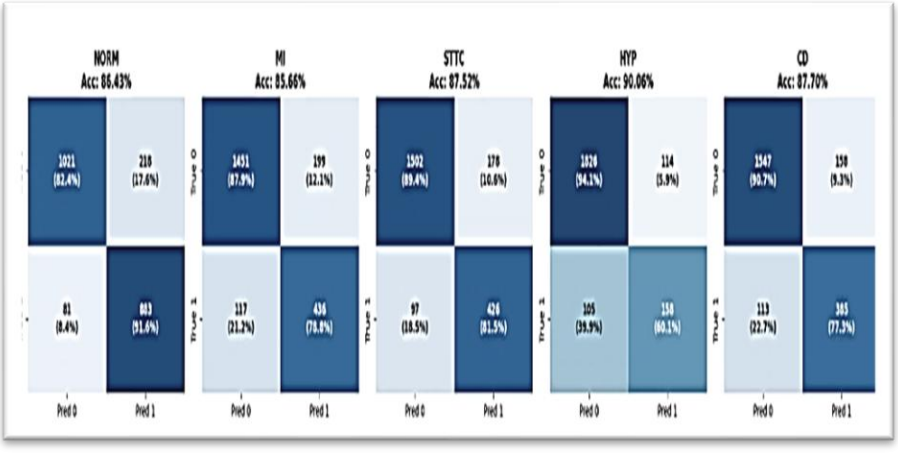

Figure 4. One-versus-rest (per-label) confusion matrices of the Full model on the test fold (fold 10).

using the per-class thresholds tuned on the validation fold. Each cell shows the count and the percentage of its row (true class). The panel title gives the label-wise accuracy of that class.

Because PTB-XL superclasses are not mutually exclusive the confusion matrix is computed separately for each label as a 2×2 table (label absent or present), not as a single 5×5 matrix. Each

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification
http://www.doi.org/10.62341/istj-vol39-2-sa40

matrix sums to 2,203 records, which is the size of the test fold as read from the matrices.

Table (4). Per-class classification performance (confusion matrix counts, sensitivity, specificity, and precision) of the proposed hybrid model on the PTB-XL test set.

Class	Support	TN	FP	FN	TP	Sensitivity (%)	Specificity (%)	Precision
NORM	964	1021	218	81	883	91.6	82.4	0.802
MI	553	1451	199	117	436	78.8	87.9	0.687
STTC	523	1502	178	97	426	81.5	89.4	0.705
HYP	263	1826	114	105	158	60.1	94.1	0.581
CD	498	1547	158	113	385	77.3	90.7	0.709

The supports (964, 553, 523, 263, 498) and the precision and recall values agree with the per-class figures reported in Section 5.1, and the mean of the five panel accuracies (86.43, 85.66, 87.52, 90.06 and 87.70%) equals the label-wise accuracy of 87.47%. (ii) For every class recall is higher than precision, as expected when thresholds maximise F1 on the validation set: the model accepts more false positives in order to miss fewer true cases. (iii) NORM has the highest sensitivity (91.6%) but also the lowest specificity (82.4%); 218 records without the NORM label were predicted as normal, which would be the more concerning error type in a screening setting, although the figure does not show which pathologies these records carry. (iv) HYP has the highest label-wise accuracy (90.06%) but the lowest sensitivity (60.1%): 105 of 263 hypertrophy records are missed. The high accuracy is a consequence of the many negatives (1,940 of 2,203 records, 88.1%), so a model that never predicted HYP would already reach 88.1%; this is the inflation of label-wise accuracy.

6. Conclusion

In this paper, we presented a Hybrid 1D-2D Deep Learning Framework for automated multi-label classification of 12-lead ECG signals. By integrating a 1D-ResNet-SE temporal stream with a 2D-EfficientNet spectral stream through a Gated Feature Fusion mechanism and Asymmetric Focal Loss, the proposed architecture effectively bridges fine-grained waveform morphology and global time-frequency dynamics. Extensive validation on the PTB-XL dataset demonstrates consistent improvements over single-modality baselines and competitive performance against state-of-the-art architectures, particularly on challenging minority pathologies such as ventricular hypertrophy. Future work will focus on integrating gradient-based visual and temporal explainability (e.g., Grad-CAM across leads and Mel-bands) and evaluating cross-dataset generalization on external clinical cohorts such as CPSC-2018 and Chapman-Shaoxing.

References

[1] World Health Organization (WHO), Cardiovascular Diseases (CVDs) Fact Sheet, Geneva, Switzerland, 2021.

[2] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 5, pp. 1519–1528, 2021.

[3] J. Schläpfer and H. J. Wellens, "Computer-interpreted electrocardiograms: benefits and limitations," Journal of the American College of Cardiology, vol. 70, no. 9, pp. 1183–1192, 2017.

[4] S. Hong, Y. Zhou, J. Shang, C. Xiao, and J. Sun, "Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review," Computers in Biology and Medicine, vol. 122, p. 103801, 2020.

[5] P. Wagner et al., "PTB-XL, a large publicly available electrocardiography dataset," Scientific Data, vol. 7, no. 1, p. 154, 2020.

[6] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a

deep neural network," Nature Medicine, vol. 25, no. 1, pp. 65–69, 2019.

[7] J. Huang, B. Chen, B. Yao, and W. He, "ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network," IEEE Access, vol. 7, pp. 92871–92880, 2019.

[8] T. Wang, C. Lu, Y. Sun, M. Yang, C. Liu, and C. Ou, "Automatic ECG classification using continuous wavelet transform and convolutional neural network," Entropy, vol. 23, no. 1, p. 119, 2021.

[9] N. Nonaka and J. Seita, "In-depth benchmarking of deep neural network architectures for ECG diagnosis," in Proceedings of the 6th Machine Learning for Healthcare Conference (MLHC), PMLR, vol. 149, 2021, pp. 414–439.

[10] T. Mehari and N. Strodthoff, "Towards quantitative precision for ECG analysis: Leveraging state space models, self-supervision and patient metadata," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 11, pp. 5326–5334, 2023.

[11] T. Mehari and N. Strodthoff, "Self-supervised representation learning from 12-lead ECG data," Computers in Biology and Medicine, vol. 141, p. 105114, 2022.

[12] M. A. Al-Masud, J. M. Lopez Alcaraz, and N. Strodthoff, "Benchmarking ECG foundational models: A reality check across clinical tasks," arXiv preprint arXiv:2509.25095, 2025.

[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.

[14] H. Zhang et al., "MLBF-Net: A multi-lead-branch fusion network for multi-class arrhythmia classification using 12-lead ECG," IEEE Journal of Translational Engineering in Health and Medicine, vol. 9, pp. 1–11, 2021.

[15] Z. Zhao et al., "Adaptive lead weighted ResNet trained with different duration signals for classifying 12-lead ECGs," in 2020 Computing in Cardiology (CinC), IEEE, 2020, pp. 1–4.

[16] R. R. Lopes et al., "Improving electrocardiogram-based detection of rare genetic heart disease using transfer learning: An application to phospholamban p.Arg14del mutation

A deep learning system that integrates 1D and 2D architectures for multi-label electrocardiogram (ECG) classification
http://www.doi.org/10.62341/istj-vol39-2-sa40

carriers," Computers in Biology and Medicine, vol. 131, p. 104262, 2021.

[17] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning (ICML), PMLR, 2019, pp. 6105–6114.

[18] Y. Chen, H. Zhang, Y. Wang, Y. Yang, X. Zhou, and Q. M. J. Wu, "MFB-CBRNN: A hybrid network for MI detection using 12-lead ECGs," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 7, pp. 2054–2065, 2020.

[19] S. Śmigiel, K. Pałczyński, and D. Ledziński, "ECG signal classification using deep learning techniques based on the PTB-XL dataset," Entropy, vol. 23, no. 9, p. 1121, 2021.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.

[21] T. Ridnik et al., "Asymmetric loss for multi-label classification," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 82–91.

[22] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL), 2020, pp. 151–159.

[23] G. S. Collins et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, p. e078378, 2024.

[24] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, p. 100804, 2023.

[25] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in International Conference on Learning Representations (ICLR), 2018.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in International Conference on Learning Representations (ICLR), 2019.