SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V13I9P106

Volume 13 Issue 9, 66-84, September 2026

Original Article

DRW-CNN: A Dilated Residual Wavelet CNN for ECG-based Severe Heart Failure Screening

Amit M. Sahu[1], Jayant P. Mehare[2]

[1,2]*Department of Computer Science and Engineering, GHRUA, AMT, MAH, India.*

[1]*Corresponding Author : amit.3696sahu@gmail.com*

Received: 25 June 2026 Revised: 22 August 2026 Accepted: 05 September 2026 Published: 26 September 2026

Abstract *- Electrocardiogram (ECG)-based heart-failure screening requires clinically valid labels, record-independent evaluation, and uncertainty analysis at the level of independent records. This study evaluates a Dilated Residual Wavelet Convolutional Neural Network (DRW-CNN) for discrimination of clinically confirmed severe Congestive Heart Failure (CHF) from normal sinus rhythm. The primary cohort comprises all 15 available BIDMC CHF records and 18 MIT-BIH Normal Sinus Rhythm Database records. Records are partitioned before segmentation, yielding 528 training, 96 validation, and 168 held-out test segments from mutually exclusive records. The network applies level-4 Daubechies-4 wavelet denoising, z-score normalization, residual blocks with dilation factors 1, 2, and 4, global average pooling, and sigmoid classification. Class-weighted binary cross-entropy addresses imbalance without deleting independent records. At the prespecified threshold of 0.5, held-out predictions achieve 95.238% accuracy, 95.833% sensitivity, 94.792% specificity, 93.243% precision, 94.521% F1-score, MCC 0.903, ROC-AUC 0.988, and average precision 0.989. Record-cluster bootstrap intervals, calibration, threshold sensitivity, record-level aggregation, leave-one-record-out stability, and auxiliary non-HF domain-shift stress analyses are also reported. The study provides a leakage-resistant severe-CHF screening evaluation that separates disease-specific evidence from auxiliary abnormality testing and quantifies uncertainty at record level.*

Keywords *- Biomedical signal processing, Class-weighted learning, Domain shift, Patient-independent validation, Precision-recall analysis.*

1. Introduction

Heart failure is a heterogeneous clinical syndrome in which structural or functional cardiac impairment produces symptoms, signs, and objective evidence of congestion or inadequate cardiac output. The electrocardiogram does not by itself establish the diagnosis, but it can contain rhythm, conduction, repolarization, and morphology information that is associated with the myocardial substrate and with ventricular dysfunction. Because ECG acquisition is inexpensive, non-invasive, and widely available, automated ECG analysis has become attractive for screening and triage, especially when a definitive imaging or biomarker test is not immediately available [1, 2].

Machine-learning research has increasingly moved from handcrafted heart-rate variability and morphology features to end-to-end representation learning. One-dimensional convolutional networks can learn local waveform patterns directly from sampled ECG signals, recurrent architectures can model sequential dependence, and attention or state-space architectures can represent longer temporal relationships [10-19]. These methods have demonstrated that an ECG contains substantially more information than is captured by conventional rule-based interpretation. At the same time, a high classification score is clinically meaningful only when the target label is appropriate and when the evaluation design prevents information leakage.

A central methodological issue in ECG-based HF research is label validity. General arrhythmias such as premature ventricular contractions, bundle-branch block, and atrial fibrillation are clinically distinct from structural heart failure. If the abnormal-rhythm record is considered an HF-positive one, then the experiment can be turned into a generic normal-versus-abnormal classifier for a disease. Therefore, the main efficacy analysis in this study is limited to those records that have a clinically coherent severe-CHF and normal-control label.

The positive cohort consists of 15 long-term BIDMC CHF recordings from patients with severe CHF (PhysioNet's description: New York Heart Association (NYHA) class III-IV), and the negative cohort consists of 18 MIT-BIH Normal Sinus Rhythm Database (NSRDB) recordings. Individual ejection-fraction phenotypes and complete NYHA I-IV labels are not available for all records. Therefore, the endpoint is

severe-CHF-versus-normal screening, not HFpEF-versus-HFrEF classification or four-class NYHA staging [3, 4].

It is also important to consider statistical independence since long ambulatory recordings can produce a large number of fixed length windows. Randomizing the windows from the same subject to the training and test sets would overestimate the sample size and allow subject-specific morphology to leak into the test set. In this case, independent records are split first, and windows are formed only in the split where the held-out records are located, which ensures that all held-out windows come from a record not used in the model training and validation process.

DRW-CNN does not introduce a new primitive of a neural network, but rather builds on existing components. A compact one-dimensional pipeline is comprised of wavelet denoising, residual learning, dilated convolution, global average pooling, and sigmoid classification. The contribution is the incorporation of these elements into a clinically coherent endpoint, leakage-resistant splitting of the records, probability-level evaluation, record-level uncertainty analysis, and interpretation based on the available data.

There are six practical contributions in the work. First, disease-specific efficacy is evaluated only for confirmed severe CHF versus a normal sinus rhythm. Second, all 33 independent records in the primary cohort are retained, and class weighting is used instead of artificial record balancing. Third, records are separated before segmentation. Fourth, saved probabilities are used to compute ROC, precision-recall, calibration, threshold sensitivity, record-level aggregation, and record-cluster bootstrap uncertainty. Fifth, contemporary TCN, transformer, and state-space ECG methods are discussed without unmatched superiority claims. Sixth, auxiliary screening runs are not interpreted as HF evidence.

2. Related Work

2.1. Traditional ECG Feature Engineering

Most previous ECG classification systems separated signal processing from classification. Time-domain heart-rate statistics, RR-interval variability, spectral measures, entropy, wavelet coefficients, and morphology descriptors were calculated and then supplied to support vector machines, decision trees, nearest-neighbour classifiers, or shallow neural networks. These pipelines are often straightforward and relatively inexpensive computationally, but they depend on feature definitions that may not remain consistent across acquisition systems and patient cohorts [11, 12].

Handcrafted heart-rate variability measures may be physiologically relevant for HF screening, as autonomic regulation and beat-to-beat variability is affected in advanced disease. However, an HRV classification system might not focus on the rhythm itself but on the overall electrocardiographic substrate that is linked to HF. This separation facilitates the maintenance of the HF label for a disease rather than general rhythm abnormality.

2.2. Convolutional and Residual ECG Models

A one-dimensional CNN can learn local filters directly from the waveform, reducing reliance on manual feature design. Large ECG studies have shown that convolutional networks can detect rhythm and diagnostic patterns in ambulatory or 12-lead recordings [10, 11]. Residual learning can simplify optimization by allowing a block to learn a residual transformation around an identity or projected shortcut [20]. This is useful when deeper feature hierarchies are desired without a proportional increase in optimization difficulty.

DRW-CNN retains a deliberately small residual stack. The objective is not to maximize depth, but to increase channel capacity and dilation while preserving a direct gradient path. This compact design supports a favorable storage and CPU-latency profile, although computational compactness alone does not establish deployment on a wearable or embedded processor.

2.3. Recurrent Networks and Temporal Convolutional Networks

Long short-term memory networks and related recurrent models have been used to capture sequential dependencies in ECG waveforms. Their recurrent state, however, introduces sequential computation and can increase training and inference cost for long signals. Temporal Convolutional Networks (TCNs) offer an alternative based on causal or dilated convolutions and can expand the receptive field through stacked dilation. Bai et al. showed that generic convolutional sequence models can compete with recurrent architectures on a range of sequence tasks [19]. Recent ECG work has combined TCNs with multi-branch processing and multi-head attention; Bi et al., for example, reported an MB-MHA-TCN for multi-class arrhythmia classification [13]. The endpoint in that study is arrhythmia rather than confirmed HF, so the published score is contextual rather than a head-to-head benchmark.

2.4. Attention, Transformers, and State-Space ECG Models

Self-attention enables a network to assign context-dependent weights across a sequence and has motivated transformer-based ECG architectures. The transformer offers global pairwise interaction, but its standard attention operation scales quadratically with sequence length [18]. State-space architectures, including Mamba, have recently been proposed as more efficient long-sequence alternatives. ECG-Mamba and related BiSSM approaches illustrate the current shift toward selective state-space modeling for cardiac abnormality classification [14-16]. These developments are directly relevant to architectural benchmarking, but their reported

results cannot be inserted as if they were generated on the present 33-record severe-CHF cohort.

2.5. Wavelet Preprocessing and Noise Suppression

ECG signals are non-stationary and commonly contain baseline variation, power-line interference, muscle activity, and electrode-motion artifacts. Wavelet decomposition is well-suited to such signals because it represents localized information at multiple scales [21, 22]. Daubechies wavelets have therefore been used frequently in ECG denoising. In DRW-CNN, wavelet processing is a preprocessing operation intended to reduce nuisance variation before representation learning. It is not presented as an original transform.

2.6. HF-Specific AI-ECG Endpoints and Disease Staging

Recent HF-specific AI-ECG studies increasingly anchor labels to clinically meaningful phenotypes. Unterhuber et al. examined HF with preserved ejection fraction, Gao et al. developed an HFpEF risk model, and Bachtiger et al. prospectively studied reduced-ejection-fraction screening using an ECG-enabled stethoscope [5-7]. More recent work has addressed incident HF risk using standard or single-lead ECGs [8, 9]. In 2026, Karabayir et al. reported externally validated ECG-AI models that distinguish reduced EF, midrange EF, HFpEF, and controls [36]; Liu et al. evaluated a transformer- teacher/CNN-student HF screening framework with external public-dataset validation [37]; and Desai et al. evaluated AI-ECG for incident HF risk in pooled longitudinal cohorts [38]. A contemporary review further emphasizes endpoint definition, external validation, and implementation context as central requirements for AI-ECG HF screening [39]. Together, these studies show that HFpEF, HFrEF, functional class, prevalent CHF, and future HF risk are distinct endpoints and should not be merged into a generic abnormal-ECG label.

HFpEF, HFrEF, functional class, and future HF risk are clinically distinct endpoints that require corresponding ground-truth labels. The confirmed cohort used here does not contain per-record ejection-fraction phenotypes or a complete NYHA I-IV label set. Consequently, the primary task is limited to severe CHF versus normal control. This endpoint preserves disease specificity without creating synthetic phenotype or staging labels.

2.7. Cross-Dataset Generalization, Domain Shift, and Wearable ECG

Generalization across hospitals and acquisition systems is a major challenge in medical machine learning. Changes in sampling frequency, lead configuration, filtering, device response, prevalence, demographics, annotation procedures, and disease spectrum can alter the joint distribution of ECG observations and labels. A model trained and tested within a single source may therefore perform well while still failing after deployment [27, 28].

Wearable and single-lead systems introduce additional distribution shifts, including electrode motion, intermittent contact, baseline wander, muscle noise, and placement variability. The MIT-BIH Noise Stress Test Database (NSTDB) supports controlled noise evaluation [30]. No NSTDB-corrupted version of the primary cohort is available in the experimental archive, so smartwatch or wearable-noise robustness is not claimed. A prespecified noise-stress protocol is described as a required external validation step.

2.8. Comparison with Contemporary Research

Table 1. Contemporary ECG research context and direct comparability

Study	Primary task	Model family	Why direct score comparison is inappropriate
Unterhuber et al. [5]	HFpEF detection	Deep ECG classifier	Different HF phenotype and clinical cohort
Gao et al. [6]	HFpEF risk assessment	Deep-learning ECG model	Different endpoint and cohort construction
Bachtiger et al. [7]	Reduced-EF screening	Single-lead AI-ECG	Prospective point-of-care setting
Dhingra et al. [9]	Incident HF risk	Single-lead AI-ECG	Longitudinal risk rather than prevalent CHF
Bi et al. [13]	Five-class arrhythmia	Attention TCN	Arrhythmia endpoint, not HF
Jiang et al. [14]	Multi-label cardiac abnormality	Mamba/SSM	12-lead abnormality endpoint
Karabayir et al. [36]	rEF/mEF/HFpEF/control	12-lead and single-lead ECG-AI	Phenotype-specific multiclass task with external validation
Liu et al. [37]	HF screening	Transformer teacher + CNN student	Different 12-lead real-world cohort and external public validation

Desai et al. [38]	Incident HF risk	AI-ECG risk model	Longitudinal risk endpoint rather than prevalent severe CHF
Present study	Severe CHF vs normal	DRW-CNN	33-record confirmed-CHF/normal cohort

2.9. Research Gap, Motivation, and Novelty Statement

The research gap is methodological rather than architectural alone. Current HF AI-ECG literature includes externally validated phenotype classification, transformer-based screening, state-space sequence models, and longitudinal HF-risk prediction [36-39]. Reproducible severe-CHF screening still depends on disease-valid labels, independence of biological units across data splits, transparent handling of class imbalance, and a clear distinction between measured and proposed capabilities. DRW-CNN is therefore assessed as a compact 1D integration on a clinically coherent severe-CHF-versus-normal task, with uncertainty and domain-shift limitations stated directly.

Accordingly, the contribution is characterized by compact implementation, full use of the available confirmed cohort, record-level leakage control, class-weighted learning, probability-level evaluation, and explicit validity analysis. Claims of a breakthrough architecture, optimal complexity, or universal state-of-the-art superiority are not made.

3. Proposed Methodology

3.1. Overview of the DRW-CNN Framework

The framework contains four operational stages: cohort construction, waveform preprocessing, dilated residual feature learning, and binary severe-CHF-versus-normal classification. Cohort construction is treated as part of the methodology because label validity determines the scientific meaning of the classifier.

Figure 1 shows the disease-specific primary cohort and distinguishes auxiliary arrhythmia or diagnostic datasets from the HF efficacy endpoint.

Following cohort construction, each waveform is converted to the target sampling frequency, denoised, normalized, and segmented into 2048-sample windows. DRW-CNN then extracts temporal features with a shallow residual-dilated stack and converts the final feature maps to a record-window probability using global average pooling and a sigmoid classifier.

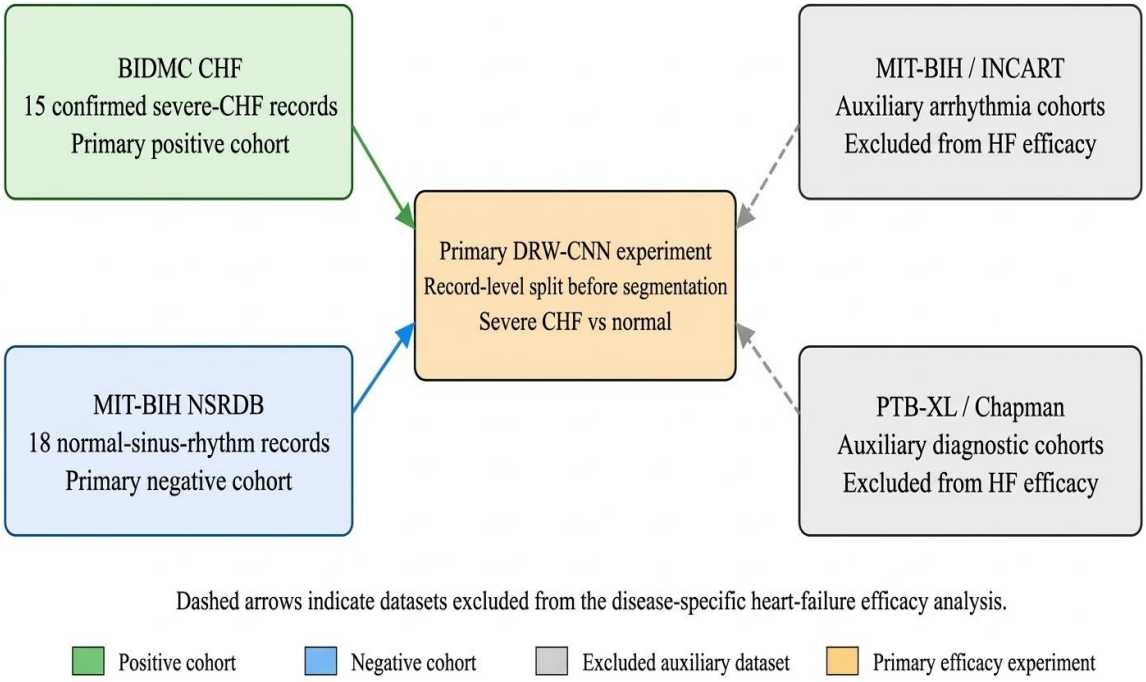

Fig. 1 Cohort construction and separation of disease-specific HF labels from auxiliary non-HF screening datasets

3.2. Primary Cohort Definition and Clinical Endpoint

The positive class consists of all 15 available BIDMC CHF records. According to PhysioNet, these patients have severe CHF (NYHA class III-IV) and approximately 20 hours of two-channel ambulatory ECG sampled at 250 Hz [3]. The negative class comprises all 18 records in the MIT-BIH Normal Sinus Rhythm Database. This produces 33 independent records before segmentation. No record was removed merely to balance the classes.

The endpoint is defined as P(Y=1|X), where Y=1 denotes membership in the confirmed severe-CHF cohort and Y=0 denotes membership in the normal-control cohort. This definition is intentionally narrower than "all heart failure" and excludes general arrhythmia diagnoses from HF-positive labeling. The database-source confounding created by using separate positive and negative databases is recognized explicitly in Section 7.

3.3. Wavelet-based Preprocessing

If necessary, waveforms are resampled to 250 Hz, and channel index 0 is used. Denoising uses a Daubechies-4 wavelet at the decomposition level 4 and soft thresholding. The highest frequency detail coefficients are used to estimate the noise scale, based on the median absolute deviation and a universal threshold.

$$W(j,k) = \sum_n x[n]\,\psi^*_{j,k}[n] \tag{1}$$

Where x[n] is the discrete ECG signal and ψ_(j,k) denotes the scaled and translated wavelet basis. After decomposition, the implementation estimates σ from the finest detail coefficients and applies a threshold proportional to σ sqrt(2 ln N). The denoised signal is reconstructed from the retained approximation and thresholded detail coefficients.

$$\hat{x}[n] = \sum_k a_{J,k}\,\phi_{J,k}[n] + \sum_j \sum_k \hat{d}_{j,k}\,\psi_{j,k}[n] \tag{2}$$

The purpose of the wavelet stage is to eliminate the nuisance high frequency variation while preserving the ECG morphology. It is not assumed to remove all motion or electrode artifacts, and stress testing of the NSTDB would be required to make this claim.

3.4. Normalization and Deterministic Segmentation

Each denoised segment is standardized using z-score normalization. For a segment x with mean μ and standard deviation σ, the normalized sample z_i is defined as:

$$z_i = \frac{x_i - \mu}{\sigma + \varepsilon} \tag{3}$$

With a small ε used numerically to avoid division by zero. The fixed model input is 2048 samples, corresponding to 8.192 s at 250 Hz. For long recordings, the implementation limits each record to 24 windows selected deterministically across the available recording span. This uniform cap prevents a few very long records from dominating training while preserving every independent record in the cohort.

3.5. Dilated Convolution and Temporal Receptive Field

Dilated convolution adds gaps between the taps of the kernel so that the receptive field grows while keeping the number of kernel coefficients unchanged.

$$y[i] = \sum_{k=0}^{K-1} w[k]\,x[i - rk] \tag{4}$$

The implemented dilation schedule is {1, 2, 4} across three residual blocks. These values form an exponential progression over the depth that is actually present. With kernel size 3, one initial convolution, and two convolutions in each residual block, the local theoretical receptive field along the main convolutional path is 31 samples, approximately 124 ms at 250 Hz. This local context is substantially shorter than the complete 2048-sample window; therefore, the dilation schedule is not interpreted as proof of optimal modeling of long-term RR dynamics. Global average pooling aggregates learned activations over the full window, but is not equivalent to a recurrent, attention, or state-space long-range mechanism.

3.6. Residual Learning Mechanism

Each residual block contains two dilated 1D convolutions followed by batch normalization and rectified linear activation. When the number of channels changes, a 1 × 1 convolution projects the shortcut to the required dimensionality. The residual mapping is

$$h_{l+1} = \mathrm{ReLU}\big(F(h_l; W_l) + P_l(h_l)\big) \tag{5}$$

Where F is the two-convolution residual transformation, and P_l is either the identity or a learned 1 × 1 projection. This formulation supports gradient propagation and feature reuse without requiring excessive network depth [20].

3.7. Global Average Pooling and Probabilistic Classification

The last Residual Block generates 128 feature maps. Global average pooling reduces each channel over the 2048 temporal positions without the need for a large fully connected layer for each time position.

$$g_c = \frac{1}{T}\sum_{t=1}^{T} h_{t,c} \tag{6}$$

The pooled vector is then passed to a 64-unit dense layer, followed by dropout of 0.30 and a single sigmoid output. The severe-CHF class probability is

$$p = \frac{1}{1+e^{-a}} \tag{7}$$

3.8. Loss, Optimization, Class Weighting, and Leakage Control

Training used the Adam optimizer with a learning rate of 0.001, weighted binary cross-entropy, a batch size of 32, and a maximum of 30 epochs. Early stopping monitored validation loss with a patience of six epochs, while the checkpoint retained for test evaluation corresponded to the highest validation accuracy. Class weights were calculated from the training-split windows (288 normal and 240 CHF). This approach addresses imbalance without deleting independent records to impose an artificial class balance.

$$L_w = -\frac{1}{N}\sum_{i=1}^{N}[w_1 y_i \ln(p_i) + w_0(1-y_i)\ln(1-p_i)] \tag{8}$$

Record identifiers are stratified into training, validation, and test groups before waveform windows are used for fitting. The final split includes 22 training records, four validation records, and seven test records. No record identifier occurs in more than one split, preventing overlapping or non-overlapping windows from the same record from appearing in both training and testing.

3.9. Out-of-Distribution and Domain Shift Mathematical Formulation

Let the source-domain joint distribution be $P_s(X, Y)$ and a deployment-domain distribution be $P_t(X, Y)$. Domain shift occurs when $P_s(X, Y)$ differs from $P_t(X, Y)$. In ECG analysis, X may change because of device response, lead configuration, sampling practice, noise, demographics, or clinical severity; Y may also change because the prevalence and labeling process differ between institutions. The expected target risk is

$$R_T(f) = \mathbb{E}_{(X,Y)\sim P_T}[\ell(f(X), Y)] \quad (9)$$

While training minimizes an empirical approximation to the source risk $R_s(f)$. A feature-space discrepancy can be expressed with Maximum Mean Discrepancy (MMD). For source features $z_i = f\_\theta(x_i\text{^}S)$ and target features $u_j = f\_\theta(x_j\text{^}T)$,

$$\mathrm{MMD}^2 = \left\| \frac{1}{n}\sum_{i=1}^{n} \phi(z_i) - \frac{1}{m}\sum_{j=1}^{m} \phi(u_j) \right\|_{\mathcal{H}}^2 \quad (10)$$

A small empirical MMD indicates that feature distributions are close under the selected kernel representation; it does not establish clinical equivalence. DRW-CNN contains no adversarial domain-adaptation or feature-alignment objective.

DWT and z-score normalization may reduce low-level acquisition variation, but they cannot be assumed to eliminate domain shift. Figure 2 summarizes this distinction. Because no independent external confirmed-HF target cohort is available, a cross-hospital HF MMD or target-risk estimate is not reported.

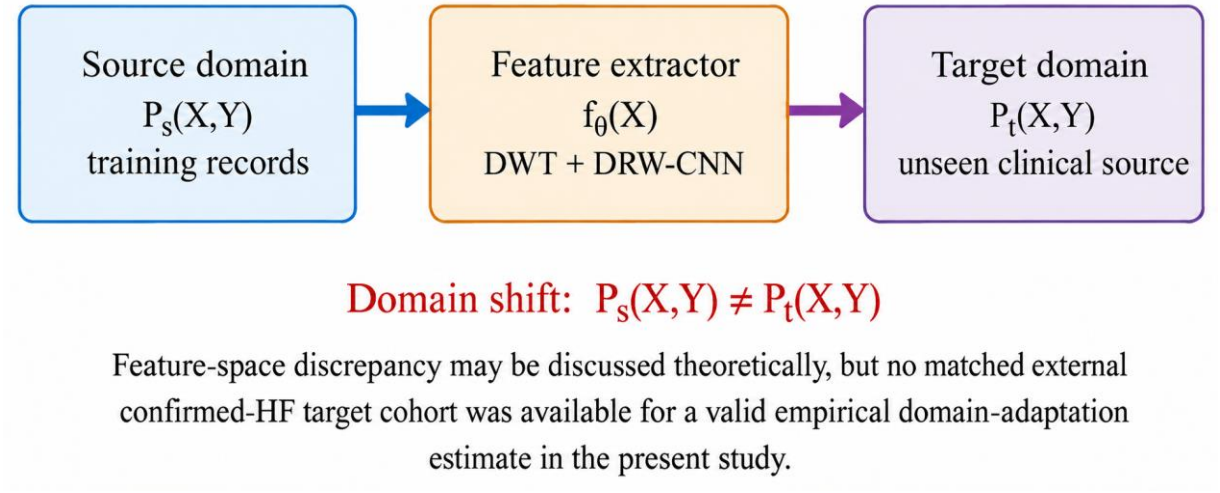

Fig. 2 Source-target domain-shift formulation and evidence boundary

3.10. Algorithm 1: DRW-CNN Training and Evaluation Procedure

Input: verified BIDMC CHF and NSRDB record manifests; random seed 42; target sampling rate 250 Hz; input length 2048; batch size 32.

1. Assign each independent record to train, validation, or test before constructing model windows.
2. Read channel index 0 from each WFDB record and resample to 250 Hz when required.
3. Select at most 24 deterministic windows across each long recording.
4. Apply db4 DWT denoising at level 4 and soft-threshold the detail coefficients.
5. Reconstruct the segment and perform z-score normalization.
6. Build the DRW-CNN with an initial 32-channel convolution and residual blocks at dilation 1, 2, and 4.
7. Compute training class weights from the record-derived training windows.
8. Optimize weighted binary cross-entropy using Adam; monitor validation loss and validation accuracy.
9. Retain the checkpoint with maximum validation accuracy and evaluate it once on the held-out records.
10. Save probability outputs, confusion counts, ROC coordinates, training history, split manifest, and model summary.
11. Recompute accuracy, sensitivity, specificity, precision, F1, MCC, ROC-AUC, average precision, calibration, and record-cluster bootstrap intervals from saved outputs.

Output: severe-CHF-versus-normal screening results derived from verified labels; surrogate arrhythmia labels are not treated as HF.

4. Experimental Setup and Evaluation Metrics

4.1. Execution Environment

The DRW-CNN experiment was executed in a native Microsoft Windows CPU environment using TensorFlow 2.21.0 and Keras 3.15.0. The archived environment report identifies the hamsvenv Python virtual environment and confirms that no GPU was detected; CUDA and cuDNN acceleration were therefore not used. NumPy and pandas supported numerical and tabular processing, SciPy supported signal-processing operations, WFDB was used for ECG record access, PyWavelets was used for db4 wavelet processing, scikit-learn supported evaluation utilities, and Matplotlib was used for graphical analysis. The random seed was fixed at 42. The exact CPU model, installed RAM capacity, Python interpreter version, and package versions other than TensorFlow and Keras were not captured in the archived DRW-CNN environment record and are therefore not asserted. All reported runtime measurements correspond to this recorded CPU environment.

4.2. Dataset Roles and Clinical Endpoint Definition

Table 2. Dataset roles and interpretation in the analysis

Dataset	Role in analysis	HF efficacy status	Interpretation
BIDMC CHF	15 severe-CHF records	Primary positive class	Clinically confirmed severe CHF
MIT-BIH NSRDB	18 normal-sinus-rhythm records	Primary negative class	Normal control cohort
MIT-BIH Arrhythmia	Auxiliary abnormality stress test	Not HF efficacy	Arrhythmia does not establish HF
INCART	Auxiliary abnormality stress test	Not HF efficacy	Annotation-derived abnormality endpoint
PTB-XL	Auxiliary diagnostic stress test	Not HF efficacy	Strict HF-specific mapping has no positive cohort
Chapman-Shaoxing	Auxiliary diagnostic stress test	Not HF efficacy	Diagnostic abnormality is not confirmed HF

Only BIDMC CHF and NSRDB contribute to the disease-specific HF efficacy endpoint. MIT-BIH Arrhythmia, INCART, PTB-XL, and Chapman-Shaoxing are retained only for auxiliary abnormality/diagnostic stress analyses. Their metrics are never pooled with the confirmed-HF result and are not interpreted as evidence of five-dataset HF generalization.

4.3. Heart-Failure Label Construction and Staging Scope

The positive label is assigned at record level to BIDMC CHF. The negative label is assigned at record level to NSRDB. The resulting target is therefore cohort membership in a severe-CHF-versus-normal experiment. No segment is independently relabeled based on arrhythmia morphology. Because individual LVEF and complete NYHA staging are unavailable, the study does not infer HFpEF, HFrEF, or NYHA class from labels that do not exist. The absence of those variables is treated as a scope constraint rather than repaired with synthetic categories.

4.4. Full-Cohort Use and Removal of Arbitrary Subsampling

The primary analysis uses all 33 independent records in the clinically valid confirmed-CHF/normal cohort. No record is removed to construct an artificial 1,000/1,000 balance.

A uniform maximum of 24 deterministic windows per long recording controls per-record dominance, while class-weighted binary cross-entropy addresses the remaining training imbalance.

Table 3. Record-independent split used in the primary experiment

Split	Independent records	CHF records	Normal records	Segments
Train	22	10	12	528
Validation	4	2	2	96
Test	7	3	4	168

4.5. DRW-CNN Architecture and Hyperparameters

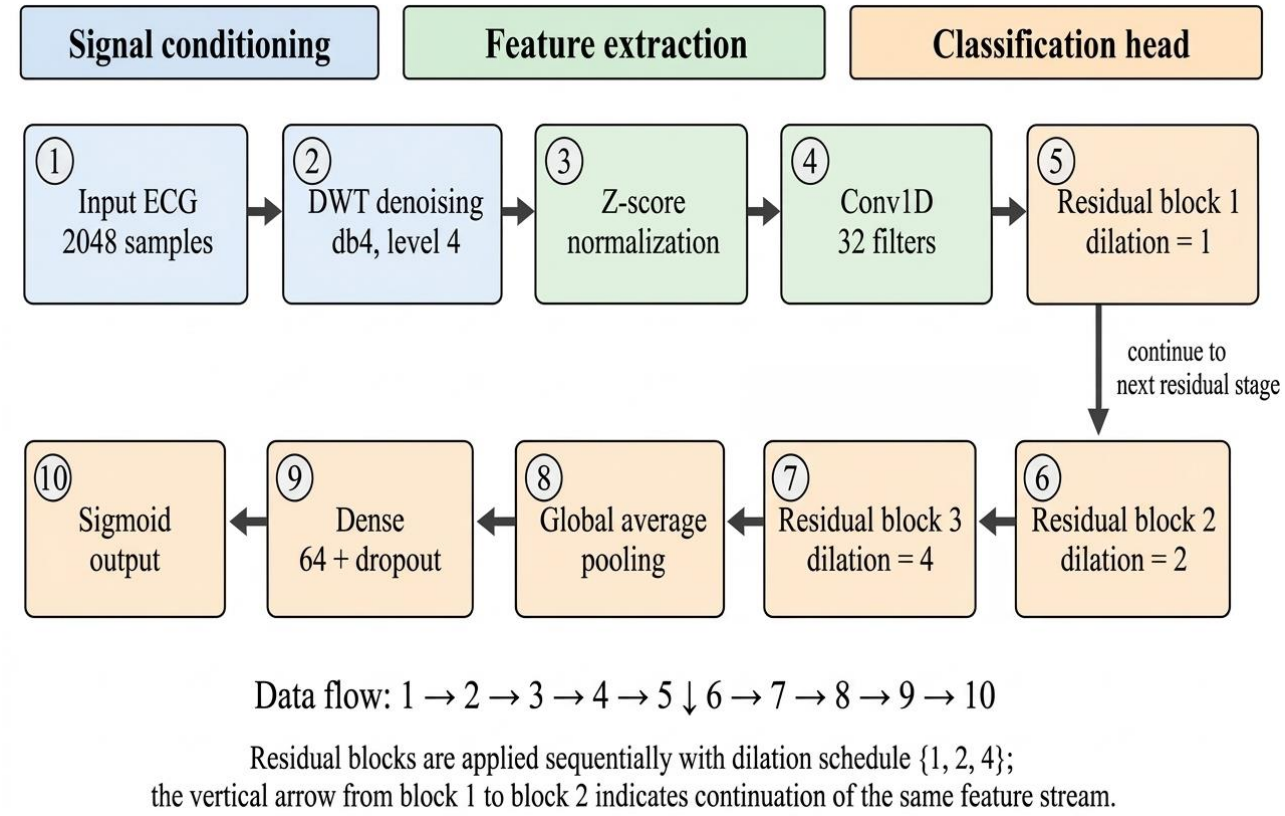

Fig. 3 Implemented DRW-CNN architecture used for the primary experiment

Table 4. DRW-CNN architecture and training configuration

Item	Value	Evidence/role
Input	2048 × 1	One preprocessed ECG channel
Target sampling rate	250 Hz	Local WFDB resampling target
DWT	db4, level 4	Implemented preprocessing setting
Initial Conv1D	32 filters, kernel 3	Local feature extraction
Residual block 1	32 filters; dilation 1	Local morphology
Residual block 2	64 filters; dilation 2	Expanded context
Residual block 3	128 filters; dilation 4	Expanded context
Pooling	Global average pooling	Temporal aggregation
Dense head	64 ReLU units	Classifier representation
Dropout	0.30	Regularization
Output	1 sigmoid unit	Severe-CHF probability
Optimizer	Adam; learning rate 0.001	Optimization
Batch size	32	Training configuration
Epochs	30 configured	Early stopping patience 6
Random seed	42	Reproducibility
Parameters	120,321 total	Model summary

4.6. Evaluation Metrics

The prespecified decision threshold for the primary confusion-matrix analysis is 0.5. Standard binary classification metrics are calculated from True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (12)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (13)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (14)$$

$$F_1 = 2\frac{\text{Precision}\cdot\text{Sensitivity}}{\text{Precision}+\text{Sensitivity}} \quad (15)$$

$$\text{MCC} = \frac{TP\cdot TN - FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (16)$$

ROC-AUC summarizes ranking across decision thresholds, while average precision summarizes the precision-recall relationship and is especially informative when class prevalence is uneven [24, 25].

4.7. Uncertainty and Statistical Analysis

No matched fold-wise experiment against WOA-TABiCNN or other external architectures is available for the present 33-record cohort, so statistical superiority is not tested or claimed.

Uncertainty in the primary result is instead quantified directly from the held-out predictions with record-cluster bootstrap resampling. Because 168 test windows arise from only seven independent records, a conventional segment-level confidence interval would understate dependence.

A record-cluster bootstrap is therefore used: seven records are sampled with replacement, all windows belonging to each sampled record are retained as a cluster, the metric is recomputed, and this process is repeated 5,000 times using random seed 42. The 2.5th and 97.5th percentiles form a descriptive 95% cluster-bootstrap interval.

These intervals reflect the small number of independent test records and should not be interpreted as external-population confidence bounds.

4.8. Scope of Contemporary Baseline, Noise, Attribution, and Hardware Validation

Direct TCN, transformer, Mamba, NSTDB, cardiologist-attribution, and embedded-hardware experiments require data or annotations that are not present in the archived experiment. Numerical results for those tests are therefore not reported.

Contemporary architectures are compared at the methodological level, while the manuscript clearly separates measured performance from validation still required for staging, wearable noise, clinical attribution, and edge deployment.

5. Results and Discussion

5.1. Training and Validation Behaviour

Training proceeded for the configured 30 epochs. Training accuracy increased from 83.333% in the first epoch to 100% in the later epochs. Validation accuracy reached 100% at several epochs, while validation loss fluctuated before declining to low values.

The checkpoint used for test inference was selected by maximum validation accuracy as specified in the experimental script. The curves should not be interpreted as evidence of broad generalization because validation data still originate from the same two source databases as training.

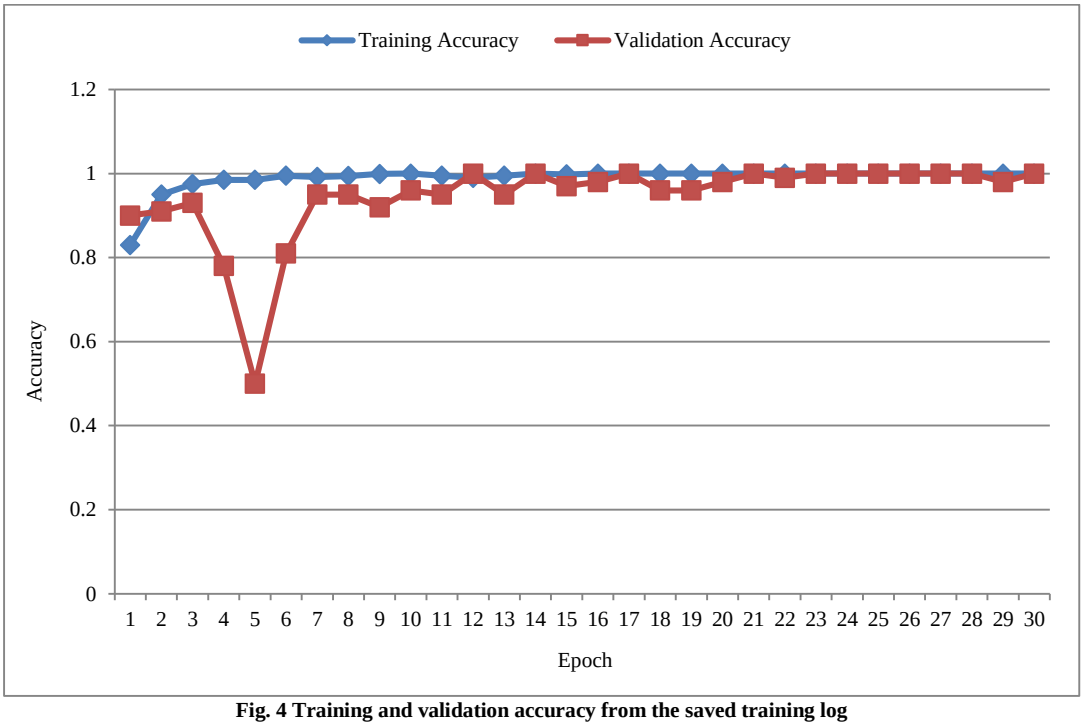

Fig. 4 Training and validation accuracy from the saved training log

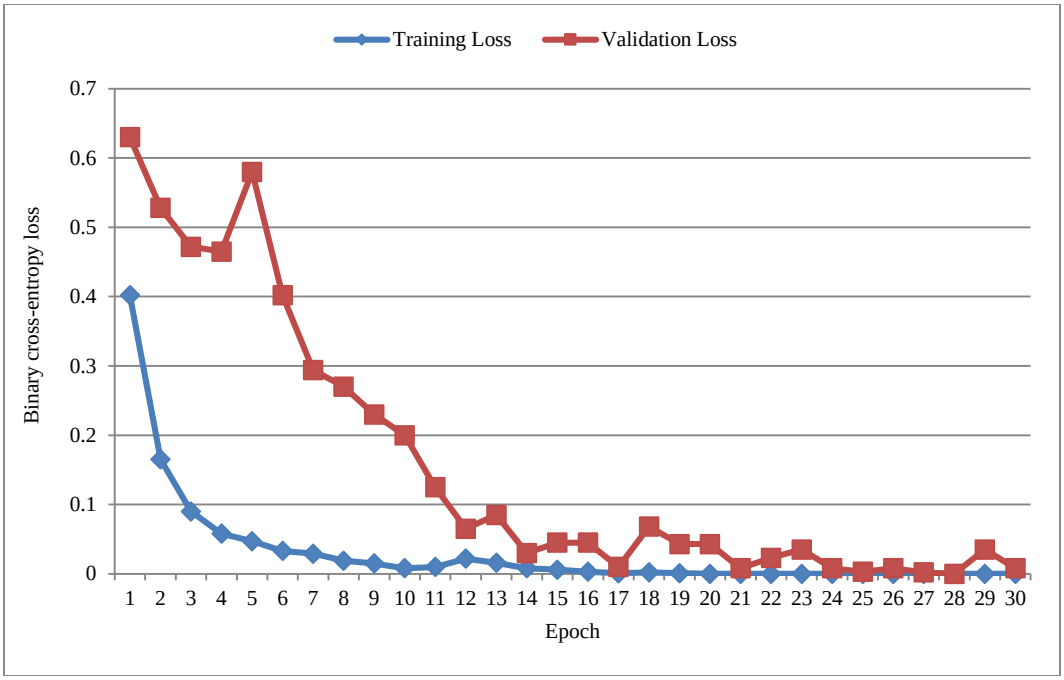

Fig. 5 Training and validation binary cross-entropy loss from the saved training log

5.2. Primary Held-Out Classification Performance

The held-out set contains 168 segments from seven unseen records: 72 CHF segments and 96 normal segments. At the prespecified threshold of 0.5, the model produced 69 true positives, 91 true negatives, 5 false positives, and 3 false negatives.

The resulting accuracy was 95.238%, sensitivity 95.833%, specificity 94.792%, precision 93.243%, F1-score 94.521%, MCC 0.903, ROC-AUC 0.988, and average precision 0.989. The Brier score was 0.0418. All values were recomputed directly from the archived evaluation and probability files.

Table 5. Primary severe-CHF-versus-normal performance with record-cluster bootstrap intervals

Metric	Observed value	95% record-cluster bootstrap interval
Accuracy	95.238%	91.071% to 98.810%
Sensitivity	95.833%	87.500% to 100.000%
Specificity	94.792%	88.542% to 100.000%
Precision	93.243%	63.636% to 100.000%
F1-score	94.521%	75.000% to 99.174%
MCC	0.903	0.759 to 0.976
ROC-AUC	0.988	0.957 to 1.000
Average precision	0.989	0.932 to 1.000

5.3. Confusion-Matrix Analysis

The confusion matrix shows five false-positive normal segments and three false-negative CHF segments. The relatively balanced error counts are consistent with the similar sensitivity and specificity values. The analysis is segment based, however, and multiple segments belong to the same test record; this dependence is addressed separately by the record-level and cluster-bootstrap analyses.

Table 6. Numerical confusion matrix for the held-out test set

Actual class	Predicted normal	Predicted CHF	Total
Normal	91	5	96
CHF	3	69	72

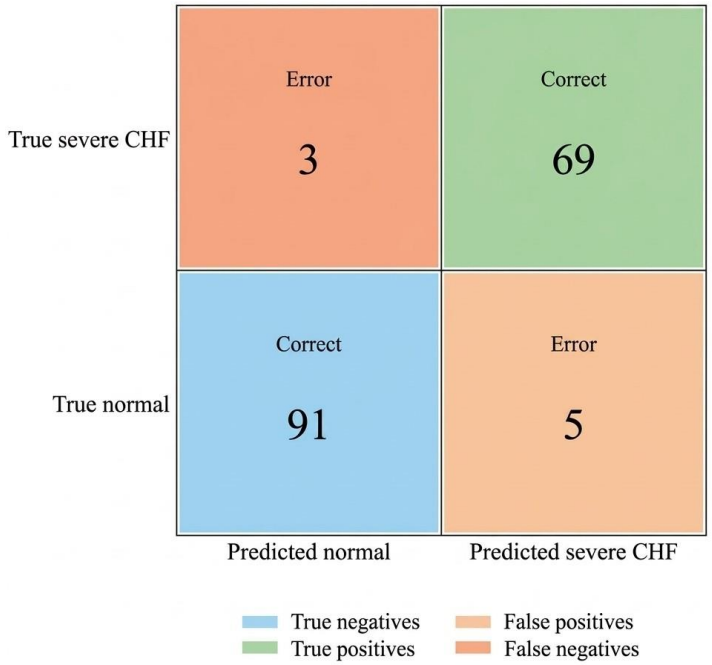

Fig. 6 Confusion matrix from the 168 saved held-out segment predictions

5.4. Receiver-Operating-Characteristic Analysis

The ROC-AUC of 0.988 indicates a strong ranking of the CHF and normal segments across thresholds. The ROC curve is generated directly from the saved probability-positive column.

Because the positive and negative classes originate from different databases, the high AUC should be interpreted as discrimination within this experimental construction rather than proof of transportability to a mixed-source hospital cohort.

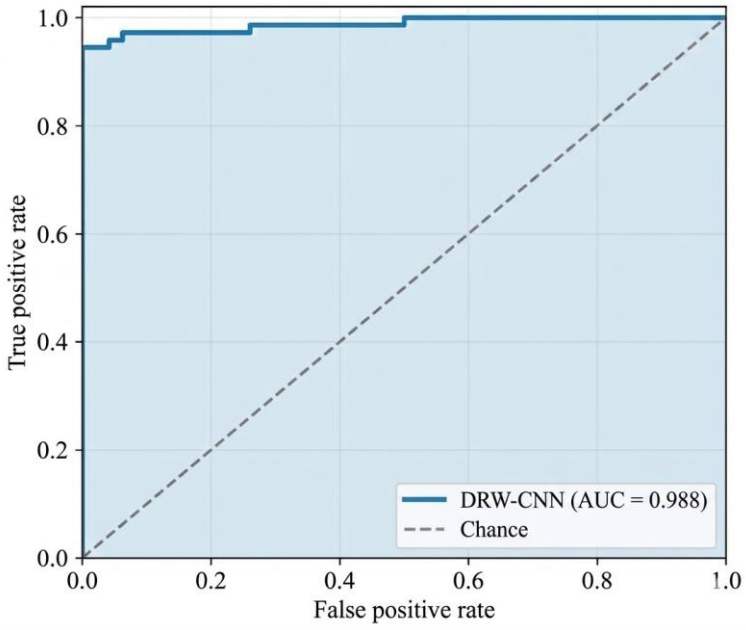

Fig. 7 ROC curve calculated from the held-out probability outputs

5.5. Precision-Recall and Probability Calibration

Average precision was 0.989. The precision-recall curve complements ROC analysis by emphasizing positive-class retrieval and false-positive burden. The Brier score was 0.0418, indicating a low mean squared probability error on this test set; however, calibration cannot be considered established because the 168 segments originate from only seven independent records.

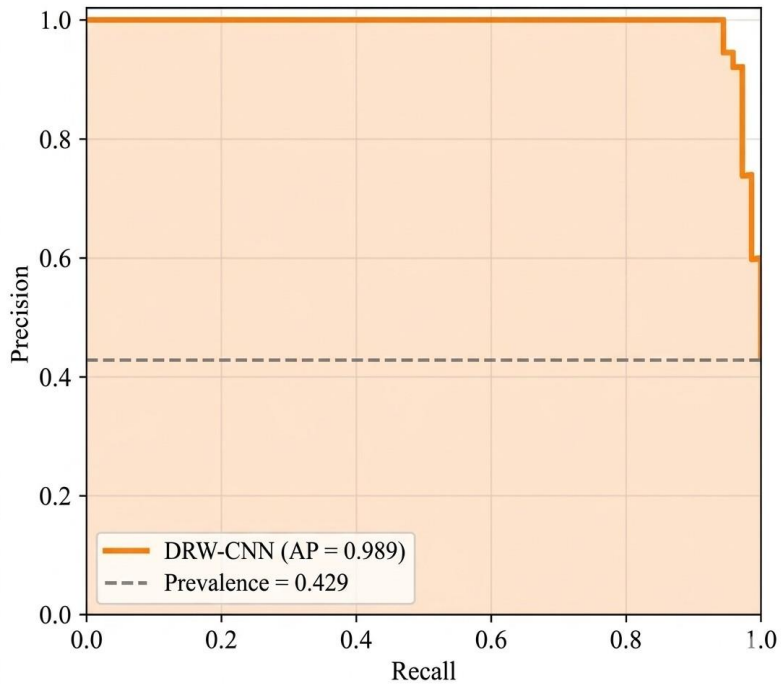

Fig. 8 Precision-recall curve calculated from the held-out probabilities

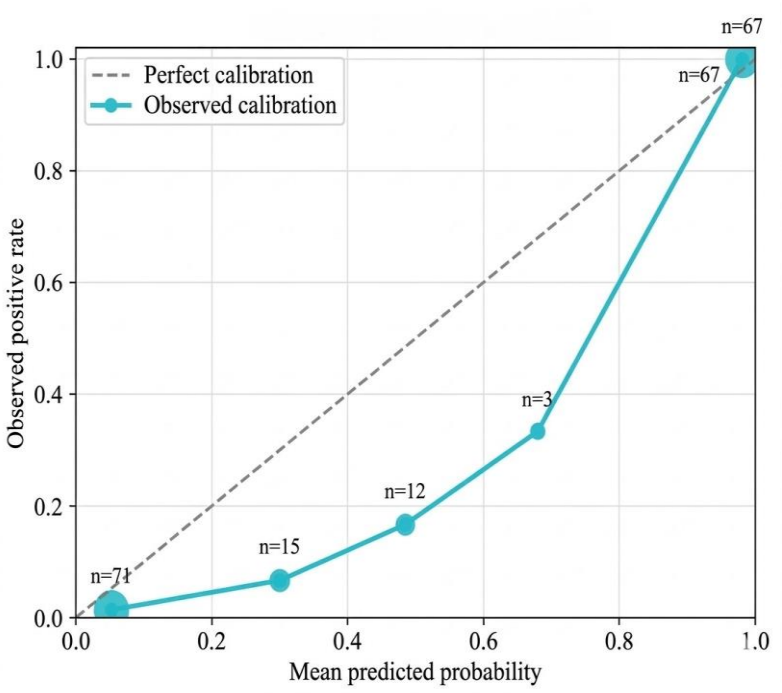

Fig. 9 Five-bin descriptive calibration plot for the held-out predictions

Table 7. Descriptive calibration bins from held-out probabilities

Mean predicted probability	Observed CHF fraction	Segments
0.0522	0.0141	71
0.3001	0.0667	15
0.4848	0.1667	12
0.6798	0.3333	3
0.9807	1.0000	67

5.6. Record-Cluster Bootstrap Uncertainty

Resampling whole records rather than individual windows produced a 95% cluster-bootstrap interval of 91.071% to 98.810% for accuracy and 0.957 to 1.000 for ROC-AUC.

The precision interval is notably wider because false-positive windows are concentrated within a small number of normal test records. These intervals make the effective sample-size limitation visible rather than treating 168 correlated segments as 168 independent patients.

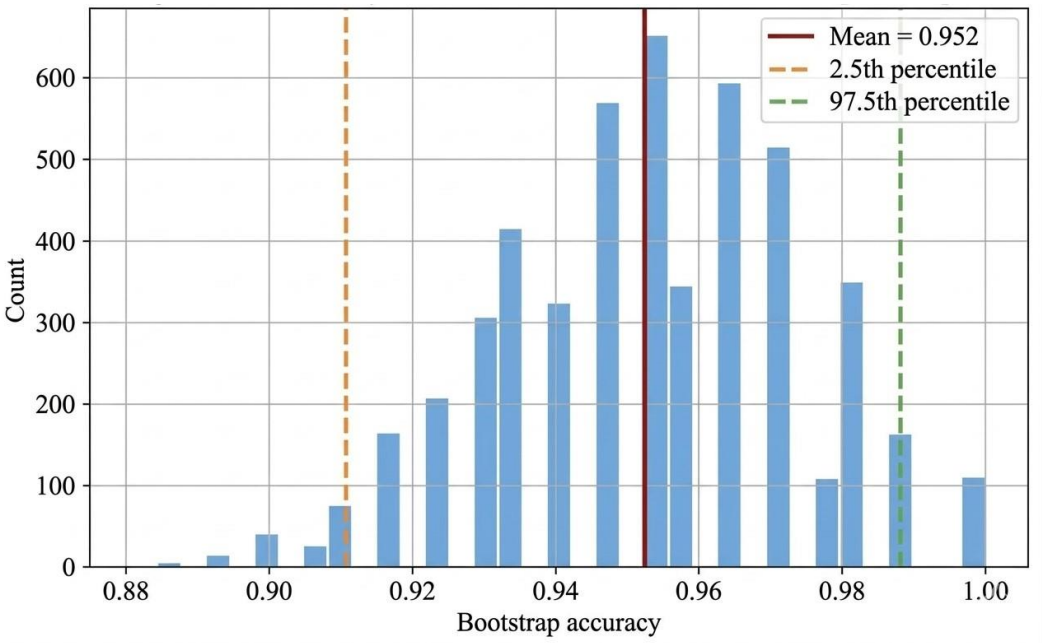

Fig. 10 Distribution of accuracy under 5,000 record-cluster bootstrap resamples

5.7. Decision-Threshold Sensitivity

The primary analysis retains the conventional 0.5 threshold. A descriptive threshold analysis is included to show how sensitivity and specificity trade off without claiming post-hoc optimization. Increasing the threshold from 0.3 to 0.7 reduces false positives but increases false negatives. Because these calculations use the held-out test set, no alternative threshold is selected from this table for deployment; threshold selection should be performed on a separate validation cohort matched to the intended clinical use.

Table 8. Descriptive threshold sensitivity on the held-out predictions

Threshold	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 (%)	FP	FN
0.30	88.095	97.222	81.250	79.545	87.500	18	2
0.40	91.667	97.222	87.500	85.366	90.909	12	2
0.50	95.238	95.833	94.792	93.243	94.521	5	3
0.60	96.429	94.444	97.917	97.143	95.775	2	4
0.70	97.619	94.444	100.000	100.000	97.143	0	4

5.8. Record-Level Sensitivity Analysis

Segment probabilities were averaged within each of the seven test records. At threshold 0.5, all seven-record means were on the correct side of the decision boundary. Because only three CHF and four normal records are present, this observation is reported as a sensitivity check rather than as a precise 100% patient-level accuracy estimate; the exact two-sided 95% binomial confidence interval for 7/7 correct records is approximately 59.0% to 100%. The analysis verifies that segment-level performance is not produced solely by isolated windows while a record mean is misclassified.

Table 9. Mean severe-CHF probability for each held-out record

Record ID	True class	Mean probability	Record-level decision
16773	Normal	0.2026	Normal
16786	Normal	0.0734	Normal
17453	Normal	0.2781	Normal
19090	Normal	0.0300	Normal

chf02	CHF	0.8715	CHF
chf05	CHF	0.9587	CHF
chf08	CHF	0.9930	CHF

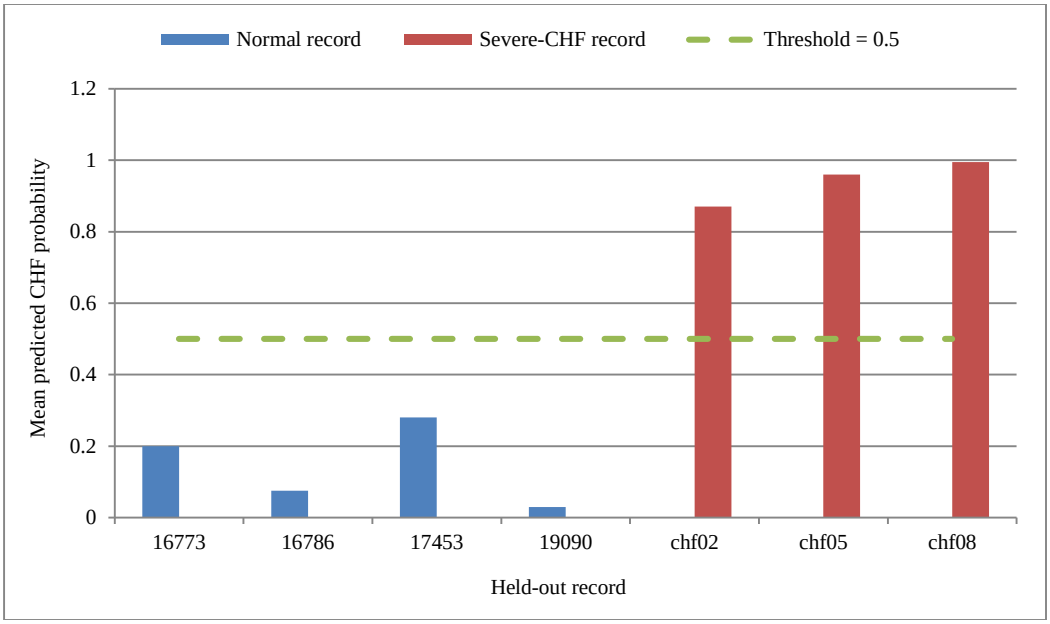

Fig. 11 Mean severe-CHF probability for each of the seven held-out records

5.8.1. Leave-One-Record-Out Stability Analysis

A fixed-test-set leave-one-record-out sensitivity analysis was performed to determine whether the aggregate segment result was driven by any single held-out record. Each of the seven test records was removed in turn, and all metrics were recomputed from the remaining archived probabilities without retraining or threshold adjustment. Accuracy remained between 94.444% and 96.528%, ROC-AUC between 0.982 and 1.000, average precision between 0.978 and 1.000, and MCC between 0.877 and 0.931. Four test records contained no segment errors; the eight observed errors were concentrated in two normal records (three and two false-positive windows) and one CHF record (three false-negative windows). Thus, no single record is solely responsible for the reported discrimination. This is a robustness analysis of one locked test split, not a substitute for repeated group cross-validation or external validation.

Table 10. Leave-one-record-out stability of the locked held-out test set

Excluded record	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC	ROC-AUC	AP
16773	96.528	95.833	97.222	0.931	0.990	0.992
16786	94.444	95.833	93.056	0.889	0.986	0.990
17453	95.833	95.833	95.833	0.917	0.992	0.994
19090	94.444	95.833	93.056	0.889	0.985	0.989
chf02	96.528	100.000	94.792	0.927	1.000	1.000
chf05	94.444	93.750	94.792	0.877	0.982	0.978
chf08	94.444	93.750	94.792	0.877	0.982	0.978

5.9. Computational Complexity and Measured Runtime

This model has 120,321 parameters, of which 118,977 are trainable, and 1,344 are not trainable. The raw FP32 parameter payload is around 0.459 MiB (excluding framework, activation and preprocessing memory).

The saved run reports 36.106 s for loading/preprocessing, 758.495 s for training, and 15.009 ms inference time per 2048-sample segment on the recorded CPU environment, or about 66.63 segments/s under that measurement procedure.

Table 11. Measured model footprint and runtime from the implemented experiment

Quantity	Measured value	Interpretation
Total parameters	120,321	Keras model count
Trainable parameters	118,977	Model summary

Non-trainable parameters	1,344	Batch-normalization state
Raw FP32 weight payload	0.459 MiB	Parameter storage only; not peak RAM
Load + preprocessing	36.106 s	Recorded full experiment
Training time	758.495 s	CPU environment; 30 configured epochs
Inference time	15.009 ms/segment	Measured batch inference
Approx. throughput	66.63 segments/s	Derived from recorded mean latency

5.10. Auxiliary Domain-Shift Stress Tests and Comparative Error Analysis

Four archived real-output runs were examined as auxiliary stress tests of task and domain dependence. These experiments use database-specific abnormality or diagnostic screening labels and are not HF efficacy tests.

MIT-BIH abnormality versus NSRDB normal controls produced 87.153% accuracy and ROC-AUC 0.950, while INCART abnormality versus NSRDB produced 91.111% accuracy and ROC-AUC 0.972. In contrast, the PTB-XL diagnostic screening protocol produced ROC-AUC 0.500 with zero specificity at the stored operating point, and Chapman-Shaoxing diagnostic screening produced ROC-AUC 0.662.

The wide performance spread demonstrates that discrimination does not transfer automatically across label definitions and data sources and argues against universal cross-dataset HF claims.

Table 12. Auxiliary domain-shift stress tests using archived non-HF screening outputs

Task	Records	Segments	Accuracy (%)	Sensitivity (%)	Specificity (%)	ROC-AUC	Scientific role
MIT-BIH abnormality / NSRDB	60	1,440	87.153	84.375	92.708	0.950	Abnormality stress test
INCART abnormality / NSRDB	75	1,800	91.111	92.803	86.458	0.972	Abnormality stress test
PTB-XL diagnostic	20,985	20,985	56.881	100.000	0.000	0.500	Diagnostic stress test
Chapman diagnostic	10,247	20,494	60.835	60.051	65.934	0.662	Diagnostic stress test

5.10.1. Error Analysis and Comparison with Contemporary Methods

The eight segment errors are more informative than an unmatched model leaderboard. False positives may arise when normal-control windows contain transient morphology or acquisition patterns that resemble the positive-source distribution. False negatives may arise when an advanced-CHF window contains little disease-specific morphology within the 8.192-s interval. Because positive and negative records come from different databases, some apparently discriminative features may also reflect acquisition-source differences. This possibility is a central reason for requiring same-protocol external validation.

The primary result is not presented as numerically superior to contemporary transformer, TCN, Mamba/state-space, or teacher-student models. Recent HF-specific studies use substantially different endpoints and cohorts, including externally validated rEF/mEF/HFpEF classification [36], transformer-teacher HF screening [37], and incident HF-risk prediction [38]. A valid head-to-head superiority claim would require each architecture to be trained on the same confirmed-CHF cohort with identical record splits, preprocessing, optimization constraints, and metric definitions. These studies, therefore, provide a contemporary methodological context rather than an unmatched leaderboard.

The high observed discrimination can plausibly reflect several factors: the positive cohort consists of severe CHF rather than subtle early disease; DWT reduces some nuisance variation; residual blocks stabilize optimization; and global pooling aggregates feature responses across the window. These explanations do not establish causal component contribution. A causal component analysis would require a complete same-split ablation series, which is not available in the archived experiment; therefore, quantitative ablation gains are not reported.

5.11. Clinical Interpretability and Waveform Feature Attribution

Clinical interpretability requires quantitative agreement between model attribution and independently annotated ECG features, not only a saliency image. A defensible study would compare attribution against prespecified markers such as QRS widening, pathological Q waves, and T-wave inversion using blinded cardiologist annotations. No such annotation set or saved held-out attribution maps are available here; therefore, DRW-CNN is described as predictive rather than cardiologist-validated or causally interpretable. For a future quantitative validation, each held-out window should be independently annotated for prespecified markers by at least two qualified readers who are blinded to model prediction. The attribution map should then be thresholded using a rule defined without

reference to the test labels. Agreement may be summarized by the fraction of attribution mass falling within annotated intervals, intersection-over-union between salient and annotated intervals, rank correlation between marker severity and attribution mass, and inter-reader agreement for the clinical annotations. A permutation or bootstrap analysis at record level would be required to avoid treating windows from one patient as independent. Until such annotations exist, DRW-CNN is described as predictive rather than clinically interpretable.

Table 13. Prespecified clinician-attribution validation framework (not reported as a completed result)

Clinical marker	Required ground truth	Quantitative attribution measure	Interpretation in this study
QRS widening	Cardiologist-marked QRS intervals	Attribution mass / IoU in QRS interval	Not validated
Pathological Q waves	Lead- and beat-level expert annotation	Attribution concentration at Q-wave interval	Not validated
T-wave inversion	Expert repolarization annotation	Attribution concentration in T-wave interval	Not validated
Rhythm-related abnormality	Expert rhythm label	Record-cluster association analysis	Not interpreted as HF evidence

5.12. Embedded Edge Hardware Deployment and Latency Benchmarking

The measured implementation is compact, but no Raspberry Pi 4, NVIDIA Jetson Nano, ARM Cortex-M, smartwatch, or other edge-device execution log is available. The only measured latency is the CPU result in Table 11. Consequently, the study supports a model-size and workstation/CPU latency claim, not a demonstrated embedded-deployment claim.

The parameter count and raw FP32 weight payload indicate that the network is compact at the model-storage level, but peak RAM also depends on intermediate activations, framework overhead, preprocessing buffers, and numerical precision. A valid edge benchmark should report device model, operating system, runtime, quantization state, batch size, warm-up protocol, median and tail latency, RAM peak, energy per inference when available, and throughput. These quantities must be measured on the target hardware. The present work, therefore, supports a compactness claim and a workstation/CPU latency report, not an embedded-deployment claim.

6. Overall Discussion

6.1. Interpretation of the Primary Result

Within the record-independent severe-CHF-versus-normal design, DRW-CNN produced eight segment-level errors. The class definition is clinically coherent, and the reported metrics can be reproduced from the saved predictions and record manifests. Interpretation remains cautious because the positive and negative cohorts originate from different databases, and the independent test set contains only seven records. Accordingly, the model is best considered a proof-of-concept for advanced CHF screening rather than a general HF diagnostic system. Severe CHF is likely to produce stronger electrocardiographic differences than early or phenotype-specific disease, and separate source databases for positive and negative records may inflate discrimination if source characteristics are learned. The next priority is therefore external testing of a locked model on same-protocol, clinically adjudicated HF and non-HF cohorts.

6.2. Architectural Interpretation and Fixed Dilation

The {1, 2, 4} dilation schedule is justified as the complete exponential schedule of the implemented three-block network, not as proof that long RR dependencies are fully covered. The receptive-field calculation shows that local convolutional context remains much shorter than the 2048-sample window. Consequently, direct benchmarking against a deeper TCN, transformer, or Mamba model would be scientifically useful if the same cohort and split can be reconstructed. The present paper does not substitute theoretical discussion for such an experiment by claiming numerical superiority.

6.3. Class Imbalance, Precision-Recall Behaviour, and Threshold Selection

Class weighting addresses imbalance during optimization without deleting independent records. Precision-recall analysis makes false-positive burden visible, and the threshold table illustrates the sensitivity-specificity trade-off. In a screening application, the operating point should be prespecified according to the clinical cost of missed CHF and unnecessary confirmatory testing; the seven-record test set is too small to establish a deployment threshold.

6.4. Relation to HF Staging, Wearable Robustness, and Clinical Deployment

Three evidence boundaries are particularly important. First, NYHA III-IV describes the positive cohort but is not a per-record four-class target. Second, single-channel compatibility does not establish smartwatch robustness; NSTDB or comparable artifact stress testing is required. Third, a compact parameter count and CPU latency do not constitute embedded deployment. These limitations are carried consistently through the methods, results, and conclusions.

7. Threats to Validity and Clinical Limitations

The first threat is the number of independent units. The test set contains 168 windows, but only seven records. Segment-level confidence can therefore appear greater than patient-level confidence. Record-cluster bootstrap intervals and record-mean sensitivity analysis are included to expose this dependence, but neither substitute for a larger external cohort.

The second threat is source-domain confounding. Every positive record comes from BIDMC CHF, and every negative record comes from NSRDB. Differences in recorder hardware, acquisition bandwidth, age distribution, recording context, and database curation may correlate with the label. Record-level splitting prevents direct patient leakage but cannot remove this database-label coupling. A stronger future design would obtain HF and non-HF controls from the same institution and acquisition protocol and reserve a second institution for external testing.

The third threat is disease-spectrum limitation. The positive cohort represents severe CHF and does not contain a complete range from asymptomatic structural disease through NYHA I-IV. The absence of individual ejection-fraction values prevents HFpEF/HFrEF stratification. Performance may therefore decline substantially in earlier, milder, or etiologically different HF populations.

The fourth threat concerns lead and noise sensitivity. Only channel index 0 is used in the implemented experiment. The study does not establish equivalence across ECG leads or wearable placements and does not include an NSTDB noise-dose experiment. DWT denoising may reduce selected noise components, but preprocessing should not be interpreted as demonstrated robustness to electrode motion, disconnection, or lead loss.

The fifth threat concerns interpretability and fairness. No cardiologist-verified attribution analysis or demographic subgroup evaluation is available. A model can be accurate while relying on non-causal source characteristics. The sixth threat concerns regulatory translation: use as software as a medical device would require locked-model verification, prospective clinical validation, risk management, cybersecurity evaluation, change-control procedures, human-factors assessment, and performance monitoring after deployment.

8. Reproducibility and Data Integrity

Each numerical result in the primary efficacy section can be reproduced from the archived evaluation, prediction, confusion matrix, ROC, segment-manifest, training-log, configuration, and model-summary files. The saved probabilities reproduce the confusion matrix, ROC-AUC, average precision, Brier score, threshold analysis, record-level means, and leave-one-record-out stability table. The segment manifest also verifies the record-disjoint split and the exact number of windows assigned to each class and split. These files allow the numerical results to be checked independently of manually transcribed tables. Five-dataset HF averages, cross-dataset HF transfer scores, unsupported ablation gains, and unmatched superiority tests are not used because the archived outputs do not provide valid same-protocol evidence for those claims.

Likewise, HF staging labels, NSTDB results, cardiologist annotations, and edge-device measurements are not synthesized. This evidence policy keeps the conclusions within the scope of measurements that can be independently traced. The reproducibility archive contains the code version, record manifests, saved probability outputs, model checkpoint or model-summary information, and environment details used for the reported analyses. Public datasets are cited through their original repositories and licenses. Any subsequent baseline, noise, staging, or hardware experiment should be added only when its raw logs and record-level split definitions are available.

9. Conclusion

DRW-CNN was evaluated for ECG-based severe-CHF screening using all 33 available confirmed severe-CHF and normal-control records, with record-level splitting before segmentation and class-weighted training. On seven unseen records, archived test probabilities yielded 95.238% accuracy, 95.833% sensitivity, 94.792% specificity, 93.243% precision, 94.521% F1-score, MCC 0.903, ROC-AUC 0.988, and average precision 0.989.

Record-cluster bootstrap and leave-one-record-out analyses indicate that the observed discrimination is not attributable to one isolated test record, although the small number of independent records and database-source coupling remain important limitations. Auxiliary non-HF screening runs show substantial task and domain dependence. These findings support DRW-CNN as a compact severe-CHF screening research model and provide a reproducible basis for same-source, external, noise-stress, and matched-architecture validation.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding Statement

No specific external financial support was reported for this study.

Data Availability

The primary ECG data are publicly available through PhysioNet subject to the original database terms. The study

archive contains the record manifests, saved predictions, evaluation files, training logs, configuration information, and model summaries used for the reported analyses.

Acknowledgments

The authors acknowledge the institutions and public data repositories that supported the research.

References

[1] Muhammad Shahzeb Khan et al., "Global Epidemiology of Heart Failure," *Nature Reviews Cardiology*, vol. 21, no. 10, pp. 717-734, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Gianluigi Savarese, and Lars H. Lund, "Global Public Health Burden of Heart Failure," *Cardiac Failure Review*, vol. 3, no. 1, pp. 7-11, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[3] PhysioNet, BIDMC Congestive Heart Failure Database, PhysioNet, 2000. [CrossRef] [Publisher Link]

[4] Ary L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, no. 23, pp. e215-e220, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[5] Matthias Unterhuber et al., "Deep Learning Detects Heart Failure with Preserved Ejection Fraction using a Baseline Electrocardiogram," *European Heart Journal-Digital Health*, vol. 2, no. 4, pp. 699-703, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Zheng Gao et al., "Electrocardiograph Analysis for Risk Assessment of Heart Failure with Preserved Ejection Fraction: A Deep Learning Model," *ESC Heart Failure*, vol. 12, no. 1, pp. 631-639, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] Patrik Bachtiger et al., "Point-of-Care Screening for Heart Failure with Reduced Ejection Fraction using Artificial Intelligence During ECG-Enabled Stethoscope Examination in London, UK: A Prospective, Observational, Multicentre Study," *The Lancet Digital Health*, vol. 4, no. 2, pp. E117-E125, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Lovedeep S. Dhingra et al., "Heart Failure Risk Stratification using Artificial Intelligence Applied to Electrocardiogram Images: A Multinational Study," *European Heart Journal*, vol. 46, no. 11, pp. 1044-1053, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[9] Lovedeep S. Dhingra et al., "Artificial Intelligence-Enabled Prediction of Heart Failure Risk from Single-Lead Electrocardiograms," *JAMA Cardiology*, vol. 10, no. 6, pp. 574-584, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Awni 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. [CrossRef] [Google Scholar] [Publisher Link]

[11] Antônio H. Ribeiro et al., "Automatic Diagnosis of the 12-Lead ECG using a Deep Neural Network," *Nature Communications*, vol. 11, no. 1, pp. 1-9, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Marek Malik et al., "Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical use," *European Heart Journal*, vol. 17, no. 3, pp. 354-381, 1996. [CrossRef] [Google Scholar] [Publisher Link]

[13] Suzhao Bi et al., "Accurate Arrhythmia Classification with Multi-Branch, Multi-Head Attention Temporal Convolutional Networks," *Sensors*, vol. 24, no. 24, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Huawei Jiang et al., "ECG-Mamba: Cardiac Abnormality Classification with Non-Uniform-Mix Augmentation on 12-Lead ECGs," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 13, pp. 461-470, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] Yupeng Qiang et al., "ECGMamba: Towards ECG Classification with State Space Models," *2024 IEEE International Conference on Bioinformatics and Biomedicine*, Lisbon, Portugal, pp. 6498-6505, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Alexey Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv*, pp. 1-22, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] Albert Gu, and Tri Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," *arXiv preprint arXiv*, pp. 1-36, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Ashish Vaswani et al., "Attention is All you Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Google Scholar] [Publisher Link]

[19] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv preprint arXiv*, pp. 1-14, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[20] Kaiming He et al., "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016. [Google Scholar] [Publisher Link]

[21] Paul S. Addison, "Wavelet Transforms and the ECG: A Review," *Physiological Measurement*, vol. 26, no. 5, pp. R155-R199, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[22] Stéphane G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674-693, 1989. [CrossRef] [Google Scholar] [Publisher Link]

[23] Changfang Chen et al., "Wavelet-Domain Group-Sparse Denoising Method for ECG Signals," *Biomedical Signal Processing and Control*, vol. 83, pp. 1-10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[24] Takaya Saito, and Marc Rehmsmeier, "The Precision-Recall Plot is More Informative than the ROC Plot when Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, vol. 10, no. 3, pp. 1-21, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[25] Tom Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[26] Diederik P. Kingma, and Jimmy Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv*, pp. 1-15, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[27] Matthew B.A. McDermott et al., "Reproducibility in Machine Learning for Health Research: Still a Ways to Go," *Science Translational Medicine*, vol. 13, no. 586, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[28] Kuba Weimann, and Tim O.F. Conrad, "Transfer Learning for ECG Classification," *Scientific Reports*, vol. 11, no. 1, pp. 1-12, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] Arthur Gretton et al., "A Kernel Two-Sample Test," *Journal of Machine Learning Research*, vol. 13, no. 25, pp. 723-773, 2012. [Google Scholar] [Publisher Link]

[30] George B. Moody, W.E. Muldrow, and G. Mark Roger, "A Noise Stress Test for Arrhythmia Detectors," *Computers in Cardiology*, vol. 11, no. 3, pp. 381-384, 1984. [Google Scholar]

[31] Patrick Wagner et al., "PTB-XL, A Large Publicly Available Electrocardiography Dataset," *Scientific Data*, vol. 7, no. 1, pp. 1-15, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[32] Jianwei Zheng et al., "A 12-Lead Electrocardiogram Database for Arrhythmia Research Covering More than 10,000 Patients," *Scientific Data*, vol. 7, no. 1, pp. 1-8, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[33] George B. Moody, and Roger G. Mark, "The impact of the MIT-BIH Arrhythmia Database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45-50, 2001. [CrossRef] [Google Scholar] [Publisher Link]

[34] D.G. Altman, and J.M. Bland, "Diagnostic Tests 1: Sensitivity and Specificity," *British Medical Journal*, vol. 308, no. 6943, 1994. [CrossRef] [Google Scholar] [Publisher Link]

[35] Glenn W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1-3, 1950. [CrossRef] [Google Scholar] [Publisher Link]

[36] Ibrahim Karabayir et al., "ECG-based Artificial Intelligence for Classifying Left Ventricular Dysfunction and Heart Failure with Preserved Ejection Fraction," *Journal of the American Heart Association*, vol. 15, no. 15, pp. 1-11, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[37] Enhan Liu et al., "A Teacher-Student Deep Learning Framework for Enhanced Clinical Screening of Heart Failure from 12-Lead Electrocardiograms," *Physiological Measurement*, vol. 47, no. 5, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[38] Akshay S. Desai et al., "Predicting Heart Failure from 12-Lead ECGs using AI: A HeartShare/AMP-HF Pooled Cohort Analysis," *Journal of the American College of Cardiology*, vol. 87, no. 8, pp. 990-1005, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[39] Lovedeep S Dhingra et al., "Artificial Intelligence-enhanced Electrocardiography for Heart Failure Screening and Risk Stratification," *Current Heart Failure Reports*, vol. 23, no. 1, pp. 1-16, 2026. [CrossRef] [Google Scholar] [Publisher Link]