An Expert-Guided Multi-Source Evidence Fusion Method For Beat-by-Beat ECG Arrhythmia Classification

Hexiao Zhang*

School of Artificial Intelligence, Hebei University of Technology, Tianjin City, Tianjin, 300130, China

Corresponding author. E-mail: zhanghx415@126.com

Received: May 17, 2026; Accepted: Aug. 14, 2026

Beat-level ECG classification is difficult because morphologically similar beats can cross AAMI boundaries, minority classes are sparse, and complementary cues are often fused implicitly. ELBD-CDM preserves waveform, clinician-semantic and morphological, and statistical rhythm evidence as separate streams. MSER constructs beat-level evidence, and WQ-CMF uses the waveform query embedding to assign sample-specific weights rather than fixed concatenation. Under the adopted MIT-BIH beat-level stratified protocol, ELBD-CDM achieved 99.24% Accuracy, 95.51% Macro-F1, and 99.25% Weighted-F1. F1 scores for S, F, and Q were 94.00%, 90.00%, and 95.00%, respectively. Removing MSER and WQ-CMF reduced Macro-F1 by 4.96 and 4.05 percentage points. Under the strongest waveform-branch Gaussian-noise setting (40 dB), ELBD-CDM retained 83.6% Macro-F1 and outperformed the reproduced comparators. The test set contained 460 S, 141 F, and 1,193 Q beats; therefore, Macro-F1 was emphasized over Accuracy. These results support sample-specific evidence fusion under this protocol, although inter-patient and external validation remain necessary.

Keywords: ECG; arrhythmia; beat-level classification; multi-source evidence fusion; cross-modal fusion

http://dx.doi.org/10.6180/jase.202612_35.032

1. Introduction

Electrocardiography (ECG) is a non-invasive, cost-effective tool for detecting and monitoring cardiovascular abnormalities. Holter monitors, wearable devices, and remote-care systems can generate recordings that are too long for consistent manual review. Automated interpretation is therefore important for scalable cardiovascular monitoring [19]. Reliable algorithmic comparison also requires reproducible datasets, transparent label definitions, and clearly specified evaluation procedures [20, 21].

Robust beat-level classification remains challenging. Beats within the same AAMI class can vary across patients and recordings, while beats from different classes may share QRS morphology or rhythm patterns [19]. The S, F, and Q classes are particularly difficult because they are under-represented, boundary-sensitive, or heterogeneous. A classifier optimized mainly for overall Accuracy may therefore favor the abundant N class [12, 14]. Clinical interpretation also uses waveform shape, RR intervals, QRS duration, P-wave evidence, rhythm regularity, and signal-quality cues [13, 19]. These observations motivate beat-specific selection of the most relevant evidence source.

Earlier beat-level systems followed several directions. De Chazal et al. [1] combined morphology and heartbeat-interval features with conventional classifiers, showing the value of structured descriptors but retaining manual feature design. Kiranyaz et al. [2] used a patient-specific 1-D CNN to adapt waveform representations, whereas Acharya et al. [3] learned heartbeat representations with a deep CNN. Kachuee et al. [4] explored transferable representations across MIT-BIH and PTB, and Oh et al. [5] combined CNN and LSTM modules for variable-length beats. Sequence-to-sequence, residual CNN, and attention-

based temporal-convolutional models further improved waveform-centered temporal modeling [6–8]. However, these methods remained primarily waveform-driven and did not preserve clinician-semantic and statistical rhythm evidence as independently queryable sources.

Alkhawaldeh et al. [9] proposed a residual CNN-BiLSTM model and reported approximately 99.4% accuracy, precision, and recall on two ECG datasets after 20 epochs. Its residual and recurrent components improve long-range waveform modeling, but the attention and representation learning remain waveform-centered. Pham et al. [10] evaluated a deep heartbeat classifier on MIT-BIH, PTB, and PhysioNet Challenge 2017. They reported 98.5% accuracy on MIT-BIH, 98.28% on PTB, and an F1-score of approximately 86.71% on the Challenge dataset. These results show that performance depends on the task definition and data source. However, the architecture does not provide beat-specific control over separately interpretable evidence streams.

Bai et al. [11] combined CNN, BiGRU, and eight-head attention to capture local, temporal, and global waveform dependencies. They reported 99.41% accuracy and a 99.21% F1-score on MIT-BIH, together with 98.82% accuracy on the PTB Diagnostic ECG Database. Although this attention mechanism models waveform context effectively, it does not represent morphology, clinician-semantic, and statistical rhythm descriptors as independent evidence tokens. Zheng et al. [12] used dual-resolution STFT inputs, parallel CNN branches, squeeze-and-excitation blocks, and fixed element-wise fusion. They reported 99.13% accuracy and 94.46% Macro-F1 on MIT-BIH, and 95.84% accuracy and 95.91% Macro-F1 on SPH. Their results confirm the value of complementary inputs. However, interpolation and summation provide fixed fusion rather than fusion conditioned on the waveform state of each beat.

Karri and Annavarapu [13] combined DWT- and delta-sigma-based QRS detection, hybrid interval and morphology features, and an LSTM for patient-specific embedded classification. They reported 99.64% accuracy and a 98.18% F1-score on MIT-BIH. The system demonstrates the deployment value of structured features, but it uses a predetermined feature representation. Balakrishnan et al. [14] combined FBSE-TQWT, PCA, fuzzy encoding, and a DNN for the AAMI five-class MIT-BIH task. Their recalls of 0.72 for SVEB and 0.74 for F highlight the difficulty of minority classes. Because the time-frequency and fuzzy features are constructed before classification, the method cannot retrieve and reweight individual evidence sources according to the current waveform representation.

Another limitation due to data organization is revealed in dataset and benchmark studies. PTB-XL offers a public large 12-lead database of patients with recommended folds and comprehensive metadata [20]. The use of standardized benchmarks, transfer learning, uncertainty assessment, and interpretability analysis can significantly influence the conclusions drawn about ECG models, as demonstrated by Strodthoff et al. [21]. These studies are NOT direct beat level MIT-BIH competitors. Rather than that, they argue that the current results can only be applied to the adopted split and argue the need for future patient-independent/external validation.

Four related gaps are present in the literature. Fix or implicit fusion is complementary inputs that are often concatenated, summed or attended within a single representation, rather than retrieved from explicit evidence. Gap 2 is limited sensitivity to minority and boundary-sensitive beats, especially S and F, for which morphology alone may be ambiguous. Gap 3 is weak evidence-level interpretability because most attention weights do not directly show how clinician-semantic and statistical rhythm cues contribute to a decision. Gap 4 is protocol-dependent generalization: strong beat-level results do not establish patient-independent clinical validity.

ELBD-CDM addresses the first three gaps architecturally and treats the fourth as an evaluation boundary. Unlike multi-input attention models that learn channel or spatial importance within pre-fused representations, ELBD-CDM maintains waveform evidence, clinician-semantic and morphological evidence, and statistical rhythm evidence as distinct tokens. The waveform query embedding assigns sample-specific weights to the structured evidence tokens. Cross-modal alignment and CSS-Injection preserve the identity of each source during fusion. The main novelty is therefore beat-conditioned evidence selection, not merely multiple inputs or an additional attention layer. Here, clinician-semantic evidence denotes automatically computed, rule-inspired ECG descriptors rather than physician-provided labels or manual beat annotations.

The main contributions of this work are mapped directly to the four identified research gaps as follows:

1. **Decoupled Multi-Modal Evidence Representation (Gap 1):** To overcome the limitations of fixed or implicit feature fusion, the proposed MSER represents raw waveform signals, clinician-semantic and morphological priors, and statistical rhythm characteristics as distinct, alignable information streams rather than collapsing them into an undifferentiated feature vector.

2. **Query-Guided Adaptive Fusion for Ambiguous**

Beats (Gap 2): To enhance discrimination among minority and boundary-sensitive classes, the WQ-CMF mechanism dynamically evaluates structured evidence using waveform query embeddings. This sample-adaptive weighting prioritizes discriminative RR-interval metrics, morphological attributes, and signal-quality indicators for ambiguous ECG segments.

3. **Preserved Semantic Interpretability (Gap 3):** To ensure transparent attribution across modalities, CSS-Injection maintains clinician-semantic priors as dedicated tokens throughout the cross-modal attention layers, preventing the semantic degradation and dilution typical of conventional feature concatenation.

4. **Rigorous and Bounded Validation Protocol (Gap 4):** To eliminate overstated claims regarding generalization, evaluation is conducted strictly under a well-defined, stratified beat-level MIT-BIH protocol. Comprehensive performance metrics (Accuracy, Macro-F1, Weighted-F1), structural ablation studies, and waveform-perturbation analyses are reported, while explicitly scoping claims to prevent unjustified extrapolation to inter-patient or external cohorts.

2. Methods

2.1. Method Overview

ELBD-CDM is a two-stage framework that combines waveform learning with explicit structured evidence to stabilize local decision boundaries. As shown in Fig. 1, MSER constructs the evidence streams and WQ-CMF performs sample-specific cross-modal fusion.

MSER extracts three synchronized representations for each beat: a waveform query embedding, clinician-semantic and morphological evidence, and statistical rhythm evidence. The waveform query embedding guides evidence selection.

WQ-CMF projects the three evidence streams into a shared space. It uses the waveform query embedding to reweight structured evidence and then performs cross-modal alignment before five-class prediction. The contribution of each structured source therefore varies across beats.

MSER constructs waveform, clinician-semantic and morphological, and statistical rhythm evidence at the beat level. WQ-CMF uses the waveform query embedding to reweight and align the structured evidence through cross-modal attention and CSS-Injection before AAMI five-class prediction.

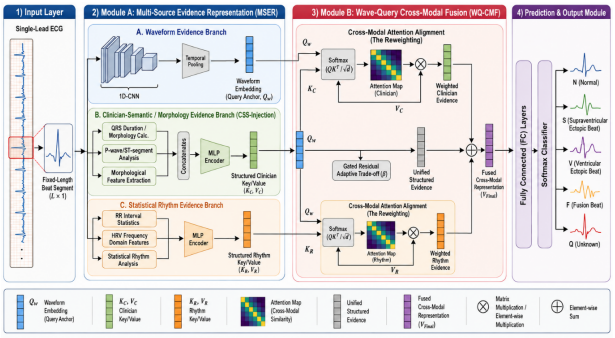

Fig. 1. Architecture of the proposed ELBD-CDM framework.

2.2. Multi-Source Evidence Representation Module (MSER)

MSER constructs synchronized beat-level representations while preserving the meaning of each evidence source. For each single-lead ECG record, it performs beat segmentation, structured evidence extraction, and waveform encoding. The resulting waveform, clinician-semantic and morphological, and statistical rhythm representations are passed to WQ-CMF as separate streams rather than one fixed feature vector.

2.2.1. *Preprocessing and beat-level sample construction*

Given a single-lead ECG record $x[n]$ sampled at $f_s = 360$ Hz in the MIT-BIH Arrhythmia Database, MSER applies morphology-oriented preprocessing to suppress non-pathological interference. A zero-phase 0.5-40 Hz Butterworth band-pass filter reduces baseline wander and high-frequency noise. A 60 Hz notch filter suppresses power-line interference when supported by the sampling rate. Each record is then normalized before beat segmentation to reduce amplitude-scale differences.

R-peak localization is performed using a lightweight derivative-energy detector followed by local maximum refinement. The first-order difference and squared energy response are defined as follows:

$$d[n] = x_{hp}[n] - x_{hp}[n-1], \quad y[n] = d[n]^2 \tag{1}$$

Using a moving-average window of $L = \mathrm{floor}(0.12fs)$, the energy response enhances QRS-related activity. Candidate samples are selected with the adaptive threshold $0.3\max(y) + 1\text{e-}6$ and a minimum interval of 0.25 s to prevent duplicate detections. Each candidate is refined to the local maximum within a 0.06 s neighborhood. Beat segments extend 0.18 s before and 0.22 s after each refined R peak and are resampled to T = 300 points by linear interpolation. QRS onset and offset are identified within

the local QRS window using a derivative threshold of $0.05\max(|d|) + 1\text{e-}6$. Beats near record boundaries are padded symmetrically. A beat with missing or physiologically implausible adjacent RR intervals is retained only if its waveform segment and required structured features can be computed consistently.

To reduce the influence of gain and scale differences across records on local boundary discrimination, beat-wise standardization is performed for each segment:

$$\tilde{x}_i[t] = \frac{x_i[t] - \mu_i}{\sigma_i + \varepsilon} \tag{2}$$

$$\mu_i = \frac{1}{T}\sum_{t=1}^{T} x_i[t] \tag{3}$$

$$\sigma_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_i[t] - \mu_i)^2} \tag{4}$$

$$\varepsilon = 10^{-7} \tag{5}$$

This procedure converts each continuous ECG record into fixed-length, consistently scaled beat samples for subsequent structured evidence construction and waveform encoding.

2.2.2. *Structured evidence construction*

MSER maps clinically interpretable cues to each beat and organizes them into morphological, clinician-semantic, and statistical rhythm groups. Morphological cues include phase index, QRS duration, normalized P-wave energy, CP, and correlations with normal and ventricular templates. Clinician-semantic cues include alpha_star, Ashman-related rhythm evidence, baseline-wander and high-frequency signal-quality ratios, saturation ratio, and spike ratio. Statistical rhythm cues include local SDNN, RMSSD, SD1, SD2, spectral entropy, Hjorth descriptors, Teager-Kaiser energy, residual norm, QRS kurtosis, and ST slope. These cues are projected into evidence vectors for wave-query reweighting rather than appended as generic handcrafted inputs. Clinician-semantic evidence denotes automatically extracted, rule-inspired ECG descriptors, not manual clinician annotations.

$$\mathrm{QRS}_{ms} = (q_{off} - q_{on}) \cdot \frac{1000}{f_s} \tag{6}$$

To supplement diagnostic cues related to the P wave, local energy is further calculated within the P wave candidate window and normalized using median statistics, forming a structured description that reflects the presence and intensity variation of the P wave. In addition, to explicitly characterize morphological consistency and signal reliability, MSER also constructs correlation evidence based on the "normal template" and signal quality indicators based on the Welch power spectrum. Specifically, the point-wise median of all heartbeat segments is used to form the normal template t_N, and the normalized cross-correlation between the current sample and the template is calculated as Corr_N; meanwhile, the power spectral density is estimated by the Welch method as $P_{xx}(f)$, and signal quality indicators are constructed using energy ratios between different frequency bands:

$$P_{energy} = \sum_{n=q_{on}-\lfloor 0.25 f_s \rfloor}^{q_{on}-\lfloor 0.08 f_s \rfloor} x_{hp}[n]^2 \tag{7}$$

$$P_{energynorm} = \frac{P_{energy}}{\mathrm{median}_j(P_{energy,j}) + \delta} \tag{8}$$

$$\mathrm{Corr}_N = \frac{1}{T}\left\langle \frac{s_i - \mu(s_i)}{\sigma(s_i) + \epsilon}, \frac{t_N - \mu(t_N)}{\sigma(t_N) + \epsilon} \right\rangle \tag{9}$$

To improve stability under ambiguous boundaries and low signal-to-noise conditions, MSER computes local RR statistics, including SDNN and RMSSD, together with spectral entropy and frequency-band energy ratios. Missing, non-finite, and physiologically implausible values are handled consistently. Standardization parameters are estimated from the training set and applied unchanged to validation and test data to prevent information leakage.

$$\mathrm{SQI}(b_1, b_2 \mid b_3, b_4) = \frac{\int_{b_1}^{b_2} P_{xx}(f)df}{\int_{b_3}^{b_4} P_{xx}(f)df + \eta} \tag{10}$$

2.2.3. *Waveform evidence representation (waveform encoder)*

In addition to structured evidence, MSER uses a waveform encoder to transform each standardized beat waveform into the waveform query embedding required for cross-modal fusion. It is expressed as:

$$z_i^{(w)} = E_w(\tilde{x}_i) \tag{11}$$

$$\tilde{x}_i \in \mathbb{R}^T, \quad z_i^{(w)} \in \mathbb{R}^{d_w} \tag{12}$$

The waveform encoder contains four one-dimensional convolutional blocks with 16, 32, 64, and 128 filters and kernel sizes of 21, 23, 25, and 27, respectively. The first two blocks use max pooling with a pool size of 3 and a stride of 2; the third uses average pooling with the same settings. Channel attention is enabled, whereas spatial attention is disabled. Two LSTM layers with 64 and 32 units process the encoded sequence. A dense layer projects the waveform query embedding to $d = 128$, followed by dropout at 0.2. The classifier contains a 64-unit dense layer, dropout at 0.2, and a five-class softmax output. The structured input includes 6 morphological, 6 clinician-semantic, and

12 statistical rhythm features. The complete TensorFlow model has approximately 701,587 trainable parameters and is evaluated under the beat-level protocol in Section 3.1.

$$E_w = \text{Dense128}(\text{LSTM32}(\text{LSTM64}(\text{CNN\_blocks}(xw)))) \tag{13}$$

The waveform encoder converts local waveform differences into a compact query representation. Fusion therefore operates on a learned representation of the current beat rather than on potentially unstable raw samples.

MSER outputs three complementary representations: a waveform query embedding, a clinician-semantic and morphological evidence vector, and a statistical rhythm evidence vector. WQ-CMF uses them for evidence reweighting, cross-modal alignment, and adaptive fusion.

2.3. Wave-Query Cross-Modal Fusion Module (WQ-CMF)

WQ-CMF learns how waveform and structured evidence should interact for each beat. The waveform query embedding from MSER serves as the query anchor. The clinician-semantic and morphological vector and the statistical rhythm vector form two structured evidence sources. Shared-space projection, wave-query reweighting, symmetric cross-attention, and fusion readout then produce a representation for five-class prediction.

2.3.1. *Unified space alignment and wave-query evidence reweighting*

Because the three evidence streams differ in scale and distribution, WQ-CMF maps them to a shared fusion space of dimension $d = 128$. This operation yields the waveform query token q_w, the clinician-semantic and morphological token c, and the statistical rhythm token r:

$$\begin{aligned} a_i &= W_a z_i^{(w)} + b_a, \\ b_i &= W_b e_i^{(\text{clin})} + b_b, \\ c_i &= W_c e_i^{(\text{stat})} + b_c, \quad a_i, b_i, c_i \in \mathbb{R}^d \end{aligned} \tag{14}$$

Here, q_w denotes the waveform query token, c the clinician-semantic and morphological token, and r the statistical rhythm token. Token c remains independent so that its diagnostic meaning is not diluted by fixed concatenation.

WQ-CMF then applies wave-query evidence reweighting. Scaled dot-product attention matches q_w with c and r to obtain sample-specific weights. The key-value set comprises c and r, and the weighted evidence is computed as follows:

$$\alpha_i = \text{softmax}\left((1/\text{sqrt}(d))K_i a_i\right) \in R^2 \tag{15}$$

$$\widehat{a}_l = V_i^T \alpha_i = \alpha_{(i,1)} b_i + \alpha_{(i,2)} c_i \in R^d \tag{16}$$

These weights quantify how strongly the current beat depends on clinician-semantic and morphological evidence or statistical rhythm evidence. The model can emphasize RR or morphology cues when waveform boundaries are ambiguous and suppress a structured source that conflicts with the waveform. This sample-specific selection distinguishes WQ-CMF from fixed concatenation and element-wise fusion.

2.3.2. *Symmetric cross-attention and fusion readout*

To reduce bias from one-way waveform querying and strengthen agreement among evidence sources, WQ-CMF applies symmetric cross-attention to q_w, c, and r.

$$T_i = a_i, b_i, c_i \tag{17}$$

Each token retrieves complementary information from the other two tokens before being updated. The cross-modal attention operation is defined as follows:

$$\begin{aligned} \text{Attn}(t; T_i\{t\}) = \sum_{s \in T_i\{t\}} & \frac{\exp(\langle W_q t, W_k s\rangle / \text{sqrt}(d))}{\sum_{s' \in T_i\{t\}} \exp(\langle W_q t, W_k s'\rangle / \text{sqrt}(d))} \\ & \cdot (W_v s) \end{aligned} \tag{18}$$

This operation updates the waveform, clinician-semantic and morphological, and statistical rhythm tokens. Bidirectional alignment prevents reliance on a one-way waveform-to-structure path. The updated tokens are concatenated to form the joint representation:

$$u_i = [\hat{a}_i; \hat{b}_i; \hat{c}_i] \in R^{(3d)} \tag{19}$$

The final fused representation is generated through linear compression, nonlinear mapping, normalization, and dropout:

$$z_i^{(f)} = \text{Dropout}\left(\text{LayerNorm}\left(\text{ReLU}\left(W_f u_i + b_f\right)\right)\right) \in R^d \tag{20}$$

CSS-Injection operates in two stages. First, clinician-semantic evidence enters the attention key-value set as an independent token and directly participates in evidence matching and weight allocation. Second, symmetric cross-attention aligns this token with waveform and statistical rhythm evidence in the reverse direction. Clinician-semantic evidence therefore remains explicit throughout fusion rather than being appended only at the classifier input. The fused representation is then passed to the beat-level classification head.

3. Results and discussion

Experiments evaluate ELBD-CDM under the specified MIT-BIH beat-level protocol. Because no inter-patient experiment was conducted, the results apply only to this protocol and do not constitute patient-independent clinical validation.

3.1. Experimental Setup

3.1.1. Experimental platform, dataset, and partitioning

Experiments were implemented in TensorFlow on an NVIDIA T4 GPU. Heartbeat annotations from the MIT-BIH Arrhythmia Database [16, 17] were mapped to the five AAMI classes: N, S, V, F, and Q. Data were divided using an 8:2 stratified beat-level random train-test split, and 20% of the training portion was used for validation. This protocol supports controlled evaluation of the fusion architecture. However, beats from the same patient may appear in both training and test sets, which can overestimate patient-independent performance. The reported results are therefore limited to this beat-level protocol. Future work should include inter-patient and external-database validation.

3.1.2. Training hyperparameters

Each beat-level sample contains a 300-point waveform segment and its structured evidence vectors. Training used up to 40 epochs, a batch size of 128, an initial learning rate of 0.001, the Adam optimizer, and sparse categorical cross-entropy loss. Dropout was 0.2 in the general model layers and 0.1 in the fusion module. Early stopping reduced overfitting, and 20% of the training portion served as the validation set.

3.1.3. Evaluation metrics

Performance is evaluated using Accuracy, Macro-F1, and Weighted-F1. Macro-F1 is the primary metric because it gives equal weight to all five AAMI classes and is more informative than Accuracy for the imbalanced S, V, F, and Q classes.

3.1.4. Additional experimental settings

Random seeds were fixed to improve reproducibility. SMOTE was applied only to the training set after AAMI mapping and before model fitting; validation and test samples were not oversampled. This procedure reduces N-class dominance without changing the natural test distribution. Training and validation curves were saved to monitor convergence. Hyperparameter experiments varied learning rate, batch size, fusion dimension, and waveform encoder depth across repeated runs. The best validation-based configuration was reported to reduce single-run variability.

Table S1. AAMI class mapping and class distribution in the adopted MIT-BIH beat-level protocol.

AAMI class	Clinical meaning	Raw train	Raw test	Train after SMOTE
N	Normal and bundle branch related beats	56,677	14,170	56,677
S	Supraventricular ectopic beats	1,839	460	56,677
V	Ventricular ectopic beats	3,017	754	56,677
F	Fusion beats	566	141	56,677
Q	Unknown or unclassifiable beats	4,771	1,193	56,677

SMOTE is applied only to the training subset. The raw test distribution is kept unchanged and is not oversampled.

3.2. Comparative Experiments

3.2.1. Comparative experiment settings

Comparative experiments were conducted at two levels. First, CNN1D, FCN, BiLSTM, TCN, CNN-LSTM, ResNet1D, and InceptionTime [15] were trained as conventional baselines with the same preprocessing and evaluation pipeline. Second, six recent designs were approximately reimplemented under a common protocol to compare residual attention, fuzzy-feature learning, deep convolution, feature-based recurrence, CNN-recurrent attention, and multi-input attention.

The recent comparison set included SE-ResNet1D [10], Fuzzy-DNN [14], CNN-9 [18], LSTM-GQRS [13], CNN+BiGRU+MHA [11], and MI-CNN+SE [12]. Because source code and complete implementation details were unavailable for several methods, these models approximate the published descriptions rather than exactly reproduce them.

Comparison fairness is explicitly bounded. Classical backbones use the complete ELBD-CDM pipeline, whereas recent methods are reimplemented only to the extent allowed by their published descriptions. The results show relative behavior within the present pipeline and should not be interpreted as a universal ranking of the original systems.

3.2.2. Comparison with conventional models

Table 1 compares ELBD-CDM with conventional deep-learning backbones on the five-class task. ELBD-CDM achieved 99.24% Accuracy, 95.51% Macro-F1, and 99.25% Weighted-F1. Relative to ResNet1D, the strongest conventional baseline, these values are higher by 0.28, 1.46, and 0.28 percentage points, respectively.

Table 1. Comparison of baseline backbones against the ELBD-CDM on the five-class task

Backbone	Acc (%)	Macro-F1 (%)	Weighted-F1 (%)
ResNet1D	98.96	94.05	98.97
CNN-LSTM	98.04	90.57	98.11
InceptionTime	97.08	86.21	97.26
TCN	94.88	78.17	95.41
CNN1D	94.13	75.73	94.73
BiLSTM	93.29	76.04	94.14
FCN	92.13	76.74	92.84
ELBD-CDM	99.24	95.51	99.25

Legend: † best in column.

The larger gain in Macro-F1 than in Accuracy indicates that the improvement is not driven solely by the abundant N class. Single-path and simply hybridized baselines are less effective when low-support classes require both morphological and rhythm context.

3.2.3. Comparison with recent representative methods

Table 2 compares ELBD-CDM with recent representative designs under the same beat-level protocol. Relative to the implemented CNN-9 model, ELBD-CDM improved Accuracy, Macro-F1, and Weighted-F1 by 0.16, 1.25, and 0.16 percentage points. Relative to CNN+BiGRU+MHA, the gains were 0.41, 2.04, and 0.44 percentage

points. These results apply only to the present approximate reimplementations and do not establish superiority over the original systems under their native protocols.

Table 2. Comparison of recent representative methods against the ELBD-CDM on the five-class task.

Method	Acc (%)	Macro-F1 (%)	Weighted-F1 (%)
SE-ResNet1D [10]	96.62	85.81	96.87
Fuzzy-DNN [14]	95.84	84.30	96.13
CNN-9 [18]	99.08	94.26	99.09
LSTM-GQRS [13]	61.56	42.37	69.49
CNN+BiGRU +MHA [11]	98.83	93.47	98.81
MI-CNN+SE [12]	98.11	90.09	98.16
ELBD-CDM	99.24†	95.51†	99.25†

Legend: † best in column.

Within the unified pipeline, ELBD-CDM achieved the highest Macro-F1 among the implemented methods. Macro-F1 is emphasized because it prevents the majority N class from dominating the evaluation.

The approximate LSTM-GQRS implementation produced lower results than the other recent models: 61.56% Accuracy, 42.37% Macro-F1, and 69.49% Weighted-F1. The discrepancy may reflect differences from the original patient-specific feature pipeline and embedded-system assumptions, which could not be reproduced in the unified setting. These values are therefore reported only for the present approximation.

3.2.4. *Analysis of comparative results*

ELBD-CDM was competitive under the adopted five-class beat-level protocol, particularly in Macro-F1. The results support combining waveform and structured evidence. However, aggregate metrics do not demonstrate improvement for every class.

Conventional CNN, recurrent, temporal-convolutional, and simple hybrid baselines capture useful morphological or temporal context. However, they do not explicitly select clinician-semantic and statistical rhythm evidence for each beat. Their larger Macro-F1 gaps are therefore consistent with weaker performance on low-support and boundary-sensitive classes.

AAMI classes differ substantially in difficulty. N is abundant and morphologically stable. S often requires RR and P-wave context, V depends on ventricular morphology, F combines normal and ventricular activation patterns, and Q is heterogeneous. The test set contained 14,170 N, 460 S, 754 V, 141 F, and 1,193 Q beats. Macro-F1 is reported alongside Accuracy so that S and F receive the same weight as the more strongly represented classes.

The highest Macro-F1 among the implemented comparison models suggests that explicit evidence fusion can help when morphology alone is insufficient. However, the comparison uses a beat-level random split and approximate reimplementations, so the finding is limited to the present pipeline.

The comparative experiments support the proposed fusion strategy under the adopted protocol. Accuracy should be interpreted together with Macro-F1 and the stated partitioning limitation. Neither metric establishes patient-independent generalization.

3.3. Ablation Experiments

3.3.1. *Ablation experiment settings*

The ablation experiments examine module removal, module replacement, and hyperparameter sensitivity. Structural experiments remove MSER or WQ-CMF or replace them with simpler alternatives. Parameter experiments vary batch size, learning rate, fusion dimension, and waveform encoder depth.

3.3.2. *Non-redundancy verification*

Tables 3(a) and 3(b) show that both modules contribute at the aggregate and class levels. Removing MSER reduced Accuracy from 99.24% to 97.46%, Macro-F1 from 95.51% to 90.55%, and Weighted-F1 from 99.25% to 97.71%. These changes correspond to decreases of 1.78, 4.96, and 1.54 percentage points. Class-specific F1 decreased by 0.99, 8.00, 0.81, 10.00, and 5.00 percentage points for N, S, V, F, and Q, respectively. Removing WQ-CMF reduced Accuracy, Macro-F1, and Weighted-F1 by 0.78, 4.05, and 0.71 percentage points. It also reduced class-specific F1 by 0.14, 6.00, 3.11, 8.00, and 3.00 percentage points for N, S, V, F, and Q, respectively. F and S showed the largest losses under both ablations, while N remained comparatively stable. The reduction for Q indicates that structured evidence construction and sample-specific fusion also benefit heterogeneous beats. Together, these results support the complementary roles of MSER in constructing clinician-semantic and morphological evidence and statistical rhythm evidence and WQ-CMF in selecting and aligning that evidence for each beat.

Table 3(a). Aggregate ablation results for MSER and WQ-CMF on the five-class task.

Variant	MSER	WQ-CMF	Acc (%)	Macro-F1 (%)	Weighted-F1 (%)
w/o MSER	×	✓	97.46	90.55	97.71
w/o WQ-CMF	✓	×	98.46	91.46	98.54
ELBD-CDM	✓	✓	99.24†	95.51†	99.25†

Legend: ✓ enabled; × disabled; † best in column. Macro-F1 is the unweighted mean of per-class F1; Weighted-F1 is the support-weighted mean of per-class F1.

3.3.3. *Optimality verification*

Table 4 shows that simpler replacements for either module reduce performance. Replacing MSER decreased Macro-F1 by 2.07 percentage points, confirming the value of explicit clinician-semantic and morphological evidence and statistical rhythm evidence. Replacing WQ-CMF decreased Macro-F1 by 1.49 percentage points, supporting waveform-conditioned weighting over simpler fusion.

3.3.4. *Hyperparameter analysis*

The hyperparameter analysis tests whether performance depends on a narrow configuration. As shown in Fig. 2, a batch size of 128

Table 3(b). Class-specific F1 scores in the removal ablation study.

Variant	N F1 (%)	S F1 (%)	V F1 (%)	F F1 (%)	Q F1 (%)
w/o MSER	98.91	86.00	97.84	80.00	90.00
w/o WQ-CMF	99.76	88.00	95.54	82.00	92.00
ELBD-CDM	99.90†	94.00†	98.65†	90.00†	95.00†

Legend: † best in column. All values are percentages.

Table 4. Replacement ablation study against the ELBD-CDM on the five-class task.

Variant	Acc (%)	Macro-F1 (%)	Weighted-F1 (%)
Replace WQ-CMF	98.82	94.02	98.86
Replace MSER	98.83	93.44	98.87
ELBD-CDM	99.24†	95.51†	99.25†

Legend: † best in column.

produced the highest Macro-F1, although the difference from 256 was small.

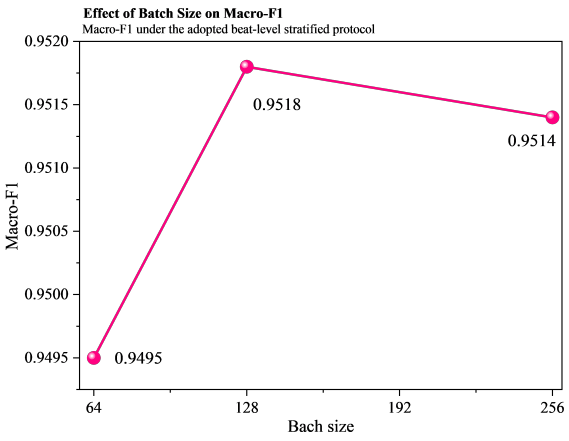

Fig. 2. Variation of model Macro-F1 under different batch sizes.

Fig. 3 shows stable performance across the evaluated learning rates. The highest Macro-F1 was obtained at 8×10^{-4}.

Figs. 4 and 5 show non-monotonic relationships between performance and fusion dimension or waveform encoder depth. Increasing model capacity therefore does not guarantee better performance. These settings should be selected by validation performance rather than model size.

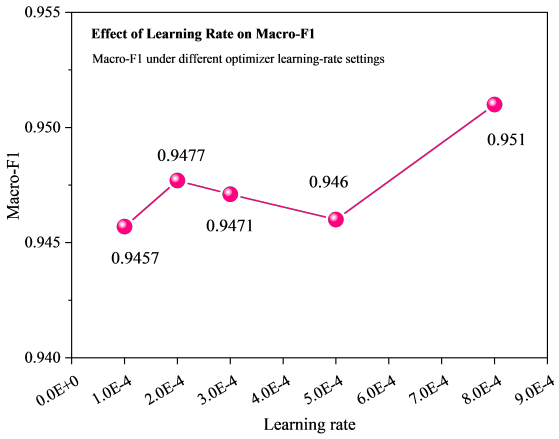

Fig. 3. Variation of model Macro-F1 under different learning rates.

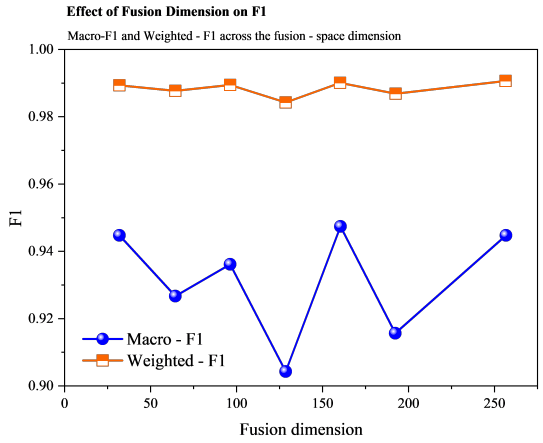

Fig. 4. Variation of model performance under different fusion dimensions.

3.3.5. *Analysis of ablation and hyperparameter results*

The ablations show complementary roles for the two modules. MSER constructs explicit structured evidence, while WQ-CMF selects and aligns that evidence for each beat. The hyperparameter analysis also shows that the gain is not a monotonic consequence of increasing model capacity.

3.4. Robustness Experiments

3.4.1. *Robustness experiment settings*

Robustness was evaluated by adding Gaussian noise only to the waveform branch during testing while keeping the structured evidence unchanged. Macro-F1 was the primary metric. This waveform-branch stress test evaluates whether structured evidence supports classification when local waveform morphology is perturbed. It does not simulate simultaneous degradation of all evidence sources or all clinical acquisition artifacts.

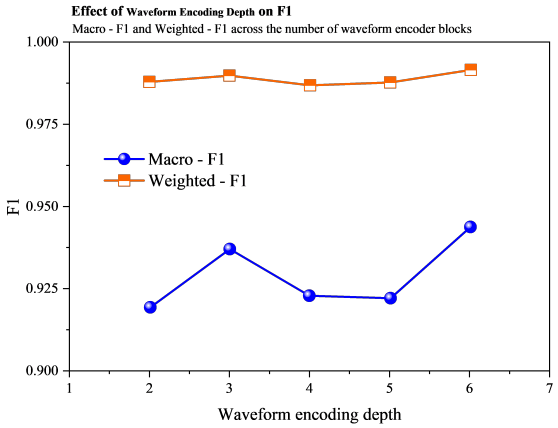

Fig. 5. Variation of model performance under different waveform encoder depths.

3.4.2. Robustness results

Fig. 6 shows that all models degraded as waveform-branch noise increased, but ELBD-CDM retained the highest Macro-F1 across the evaluated range. Its Macro-F1 decreased from 95.2% at 0 dB to 83.6% at 40 dB, a decline of 11.6 percentage points. At 40 dB, the reproduced CNN+BiGRU+MHA, MI-CNN+SE, and LSTM-GQRS models achieved 16.6%, 18.0%, and 14.1%, respectively. The smaller decline of ELBD-CDM is consistent with partial compensation from the structured evidence streams.

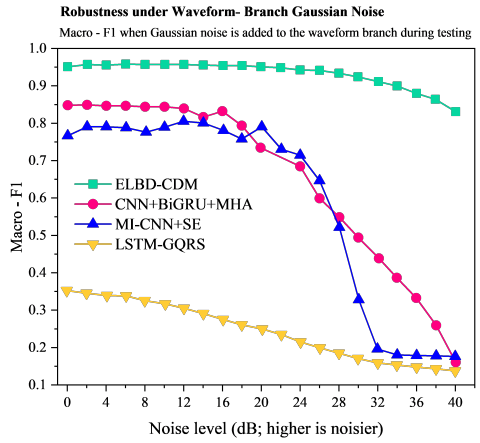

Fig. 6. Macro-F1 curves of different models under waveform-branch Gaussian noise perturbation.

3.4.3. Analysis of robustness results

The results support robustness to waveform-branch degradation, not universal robustness. When waveform quality decreases, ELBD-CDM can use clinician-semantic and morphological evidence together with statistical rhythm evidence. However, the experiment does not perturb all evidence sources or reproduce baseline wander, electrode motion, lead reversal, and other acquisition artifacts.

4. Conclusions

- ELBD-CDM is an expert-guided evidence-fusion framework for beat-level ECG arrhythmia classification. It jointly models waveform evidence, clinician-semantic and morphological evidence, and statistical rhythm evidence.
- MSER constructs structured evidence at the beat level, allowing the classifier to use interpretable morphological, rhythm, P-wave-related, and signal-quality descriptors.
- WQ-CMF uses wave-query evidence reweighting and CSS-Injection to select and align structured evidence according to the waveform query embedding rather than using fixed concatenation.
- Under the adopted MIT-BIH beat-level stratified protocol, ELBD-CDM achieved 99.24% Accuracy, 95.51% Macro-F1, and 99.25% Weighted-F1. These results show strong performance within the unified pipeline but do not establish patient-independent validity.
- Structural ablations and waveform-branch perturbation tests support the complementary roles of MSER and WQ-CMF. Removing these modules reduced Macro-F1 by 4.96 and 4.05 percentage points, respectively. ELBD-CDM retained 83.6% Macro-F1 under the strongest evaluated perturbation of 40 dB.

Limitations and future work

This study has four main limitations. First, it uses one public database and a beat-level stratified random split, so beats from the same patient can appear in both training and test sets. Second, the class-specific ablation results apply only to the adopted split and do not quantify variability across repeated seeds or patient-independent partitions. Third, external validation on independent datasets is unavailable. Fourth, the framework has not been evaluated with multi-lead recordings, real-time deployment, or prospective clinical workflows. Future work will prioritize repeated class-resolved ablations with uncertainty estimates, inter-patient evaluation, external-database validation, multi-lead extension, and deployment testing under realistic acquisition artifacts.

Conflicts of interest

The author declares no conflicts of interest.

Authorship contribution statement

Hexiao Zhang: Conceptualization, Project administration, Writing-Original draft preparation.

Data availability

The data analyzed in this study were obtained from the publicly available MIT-BIH Arrhythmia Database, version 1.0.0, hosted by PhysioNet at https://physionet.org/content/mitdb/1.0.0/. The database was accessed and used in accordance with its applicable access policy and license.

Author statement

The author has read and approved the manuscript, met the stated authorship requirements, and believes that the manuscript represents honest work.

Funding

Inapplicable

Ethics statement

This study used the publicly available, de-identified MIT-BIH Arrhythmia Database hosted by PhysioNet. It involved no new participant recruitment, intervention, or collection of identifiable information. The database was used in accordance with its access policy and license.

References

[1] P. De Chazal, M. O'Dwyer, and R. B. Reilly, (2004) *"Automatic classification of heartbeats using ECG morphology and heartbeat interval features"* **IEEE transactions on biomedical engineering *51***(7): 1196–1206. DOI: https://doi.org/10.1109/TBME.2004.827359.

[2] S. Kiranyaz, T. Ince, and M. Gabbouj, (2015) *"Real-time patient-specific ECG classification by 1-D convolutional neural networks"* **IEEE transactions on biomedical engineering *63***(3): 664–675. DOI: https://doi.org/10.1109/TBME.2015.2468589.

[3] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. San Tan, (2017) *"A deep convolutional neural network model to classify heartbeats"* **Computers in biology and medicine *89***: 389–396. DOI: https://doi.org/10.1016/j.compbiomed.2017.08.022.

[4] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, (2018) *"Ecg heartbeat classification: A deep transferable representation"* **arXiv preprint arXiv:1805.00794**: DOI: https://doi.org/10.1109/ICHI.2018.00092.

[5] S. L. Oh, E. Y. Ng, R. San Tan, and U. R. Acharya, (2018) *"Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats"* **Computers in biology and medicine *102***: 278–287. DOI: https://doi.org/10.1016/j.compbiomed.2018.06.002.

[6] S. Mousavi and F. Afghah. "Inter-and intra-patient ecg heartbeat classification for arrhythmia detection: a sequence to sequence deep learning approach". In: *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2019, 1308–1312. DOI: https://doi.org/10.1109/ICASSP.2019.8683140.

[7] F. Khan, X. Yu, Z. Yuan, and A. U. Rehman, (2023) *"ECG classification using 1-D convolutional deep residual neural network"* **Plos one *18***(4): e0284791. DOI: https://doi.org/10.1371/journal.pone.0284791.

[8] Y. Zhao, J. Ren, B. Zhang, J. Wu, and Y. Lyu, (2023) *"An explainable attention-based TCN heartbeats classification model for arrhythmia detection"* **Biomedical Signal Processing and Control *80***: 104337. DOI: https://doi.org/10.1016/j.bspc.2022.104337.

[9] R. S. Alkhawaldeh, B. Al-Ahmad, A. Ksibi, N. Ghatasheh, E. M. Abu-Taieh, G. Aldehim, M. Ayadi, and S. M. Alkhawaldeh, (2023) *"Convolution neural network bidirectional long short-term memory for heartbeat arrhythmia classification"* **International Journal of Computational Intelligence Systems *16***(1): 197. DOI: https://doi.org/10.1007/s44196-023-00374-8.

[10] B.-T. Pham, P. T. Le, T.-C. Tai, Y.-C. Hsu, Y.-H. Li, and J.-C. Wang, (2023) *"Electrocardiogram heartbeat classification for arrhythmias and myocardial infarction"* **Sensors *23***(6): 2993. DOI: https://doi.org/10.3390/s23062993.

[11] X. Bai, X. Dong, Y. Li, R. Liu, and H. Zhang, (2024) *"A hybrid deep learning network for automatic diagnosis of cardiac arrhythmia based on 12-lead ECG"* **Scientific Reports *14***(1): 24441. DOI: https://doi.org/10.1038/s41598-024-75531-w.

[12] B. Zheng, W. Luo, M. Zhang, and H. Jin, (2025) *"Arrhythmia classification based on multi-input convolutional neural network with attention mechanism"* **Plos one *20***(6): e0326079. DOI: https://doi.org/10.1371/journal.pone.0326079.

[13] M. Karri and C. S. R. Annavarapu, (2023) *"A real-time embedded system to detect QRS-complex and arrhythmia classification using LSTM through hybridized features"* **Expert Systems with Applications *214***: 119221. DOI: https://doi.org/10.1016/j.eswa.2022.119221.

[14] K. Balakrishnan, D. Velusamy, K. Ramasamy, and L. Pruinelli, (2025) *"ECG-based cardiac arrhythmia classification using fuzzy encoded features and deep neural networks"* **Biomedical Engineering Advances *9***: 100167. DOI: https://doi.org/10.1016/j.bea.2025.100167.

[15] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, (2020) *"Inceptiontime: Finding alexnet for time series classification"* **Data mining and knowledge discovery *34***(6): 1936–1962. DOI: https://doi.org/10.1007/s10618-020-00710-y.

[16] G. B. Moody and R. G. Mark, (2001) *"The impact of the MIT-BIH arrhythmia database"* **IEEE engineering in medicine and biology magazine *20***(3): 45–50. DOI: https://doi.org/10.1109/51.932724.

[17] G. Moody and R. Mark. *MIT-BIH Arrhythmia Database (version 1.0. 0). PhysioNet*. 2005. DOI: https://doi.org/10.13026/C2F305.

[18] M. A. Raza, M. Anwar, K. Nisar, A. A. A. Ibrahim, U. A. Raza, S. A. Khan, and F. Ahmad, (2023) *"Classification of electrocardiogram signals for arrhythmia detection using convolutional neural network"* **Computers, Materials and Continua *77***(3): 3817–3834. DOI: https://doi.org/10.32604/cmc.2023.032275.

[19] E. J. d. S. Luz, W. R. Schwartz, G. Cámara-Chávez, and D. Menotti, (2016) *"ECG-based heartbeat classification for arrhythmia detection: A survey"* **Computer methods and programs in biomedicine *127***: 144–164. DOI: https://doi.org/10.1016/j.cmpb.2015.12.008.

[20] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, (2020) *"PTB-XL, a large publicly available electrocardiography dataset"* **Scientific data *7***(1): 154. DOI: https://doi.org/10.6084/m9.figshare.12098055.

[21] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, (2020) *"Deep learning for ECG analysis: Benchmarks and insights from PTB-XL"* **IEEE journal of biomedical and health informatics *25***(5): 1519–1528. DOI: https://doi.org/10.1109/JBHI.2020.3022989.