

A Query-Centric Diagnostic Attention BLSTM Network for Short-Window ECG-Based Atrial Fibrillation Detection

Yang Li

Anhui Science and Technology University

Jianguo Chen

Anhui Science and Technology University

Manhong Shi

shimh@ahstu.edu.cn

Anhui Science and Technology University

Research Article

Keywords: short-window ECG, atrial fibrillation, bidirectional long short-term memory, query-centric diagnostic attention, deep learning

Posted Date: August 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10613107/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

A Query-Centric Diagnostic Attention BLSTM Network for Short-Window ECG-Based Atrial Fibrillation Detection

Yang Li¹, Jianguo Chen¹, Manhong Shi^{1*}

¹ College of Information and Network Engineering, Anhui Science and Technology University, Bengbu 233000, China

* **Correspondence:**

shimh@ahstu.edu.cn

Background

The development of short-window electrocardiogram (ECG)-based methods for detecting atrial fibrillation (AF), which is associated with increased risks of stroke, heart failure, and cardiovascular death, has been gaining significant attention for wearable monitoring devices and large-scale screening. Currently, there are CNN-BLSTM-based approaches that use either global pooling or attention methods to consolidate the temporal information within each ECG window. This approach, however, could limit the ability of the developed model to use important local information in short ECG segments. This study examines the performance of the Query-Centric Diagnostic Attention BLSTM Network (QDA-BLSTM), which can be employed for detection of AF from short ECG windows.

Methods

The proposed architecture utilizes a four-scale 1D CNN encoder that extracts morphological features from ECG signals at different scales; a channel recalibration module that enhances the quality of representation; a BLSTM encoder that encodes the temporal information in both directions; and a query-centric diagnostic attention mechanism in which five learnable diagnostic queries provide complementary temporal information, reducing reliance upon a single global aggregation operation. Internal evaluation was performed using the subject-independent internal dataset derived from both AFDB and NSRDB, while the model was externally evaluated on CPSC2018 without additional training. The area under the receiver operating characteristic curve (AUC) on the validation set was used to select checkpoints, and prediction thresholds were selected by maximizing balanced accuracy on the validation set. The proposed methodology was compared with CNN-BLSTM using an exact two-sided McNemar test.

Results

On the internal test set, QDA-BLSTM achieved 99.42% accuracy, 99.70% sensitivity, and 99.15% specificity, compared with 98.17%, 99.90%, and 96.43%, respectively, for CNN-BLSTM. The difference in prediction errors was statistically significant ($P = 7.86 \times 10^{-23}$). On the external dataset, QDA-BLSTM achieved 91.01% accuracy, 85.67% sensitivity, and 97.91% specificity, compared with 89.71%, 84.49%, and 96.45%, respectively, for CNN-BLSTM.

Conclusions

The use of QDA-BLSTM improved prediction accuracy and specificity over CNN-BLSTM on both the internal and external datasets. These results indicate that learnable diagnostic queries can provide complementary temporal information for detecting AF from short ECG windows.

Keywords: short-window ECG; atrial fibrillation; bidirectional long short-term memory; query-centric diagnostic attention; deep learning

1. Introduction

Atrial fibrillation (AF) is the most frequently encountered of all sustained cardiac arrhythmias, and is associated with stroke, heart failure and coronary heart disease, as well as increased cardiovascular morbidity. Early recognition of AF is therefore important for early intervention and planning for long-term risk management. The electrocardiogram (ECG) is the most commonly used noninvasive tool for AF diagnosis since the ECG corresponds to the heart rhythm and electrical pattern. In practice, AF is diagnosed through visual inspection of the ECG signals, characterized by irregularity of the rhythm strip, changes in the P waves, and changes in atrial activity [1].

ECG interpretation, however, can be labor-intensive, reliant on physicians, and can take a long time. As the popularity of wearables and other forms of remote monitoring technology increases throughout the world, the ability to automatically detect AF using short-window ECG technology is rising in popularity for its screening potential and capability for remote patient monitoring [2]. Existing ECG-based AF detection techniques can be split into approaches by handcrafted feature extraction and deep learning approaches. Classical techniques usually use R-R interval variability, heart rate variability, wavelet descriptors, nonlinear nonparametric measures, entropy measures, and more, possibly concatenating them with other multi-domain statistical features, then combine that information with classical classifiers [3-9]. For example, AF will be classified using single lead ECG as demonstrated by testing completed during the PhysioNet/Computing in Cardiology Challenge 2017, while algorithms based upon random forests, wavelet transforms, entropy, R-R interval, and ensemble learning have been developed for classifying AF recorded during short/continuous ECG (wearable) recordings. Although these methods are relatively computationally efficient and interpretable, their performance depends on factors such as accurate detection of R-peaks, ECG signal quality, feature selection, and creation of appropriate pre-processing pipelines. Due to the limited number of heart beats captured in short ECG windows, challenges exist in the reliability of rhythm level statistics to completely represent transient types of AF, or subtle differences in atrial activity during AF.

More recently, deep learning defines a new paradigm for ECG-based arrhythmia classification. For example, Hannun et al. [2] showed that deep neural networks achieve cardiologist-level arrhythmia classification from ambulatory single-lead ECG recordings, thus demonstrating the feasibility of automated ECG analysis for scalable monitoring. CNNs perform well at extracting local morphological features from the raw or minimally preprocessed ECG signal, while RNNs (including long short-term memory (LSTM), gated recurrent unit (GRU), and bidirectional LSTM (BLSTM)) model the temporal dependencies in ECG sequences. Thus, CNN-LSTM, CNN-BLSTM, and modified BiLSTM architectures have also been widely used for AF detection and ECG-based arrhythmia [10-12]. More recent work also investigates residual-dense recurrent networks, rhythm-oriented end-to-end models, lightweight feature-fusion networks, time-frequency CNN frameworks, raw ECG generalization models, and localization-aware Transformer-based models, taking advantage of them to improve the representational learning, temporal modeling, wearable-device applicability, or cross-dataset generalization of AF and NSR detection from ECG [13-18]. While effective, a significant portion of the deep learning methods aggregate the temporal information through final hidden states, temporal averaging, global pooling, and other compact summaries of the sequence. In this way, the resulting aggregated information may not explicitly organize the varied types of AF/NSR-related pieces of diagnostic evidence, when AF recognition usually relies on various complementary cues associated with rhythm, atrial activity, regular sinus rhythm evidence, QRS stability, and global temporal context. As a result, diagnostically useful evidence may be compressed into a single global representation, making it difficult for the model to explicitly preserve and utilize complementary decision cues from short ECG.

Attention mechanisms have also been incorporated into ECG classification for enhancing feature selection and interpretability. Temporal attention helps highlight informative time steps, channel attention refines feature channels, and self-attention captures long-range dependencies among temporal tokens [19-22]. ECGNET, a hierarchical attention-based model, has attempted to highlight informative waves, beats, or temporal regions for AF detection and to provide an interpretable view of ECG decision-making [19, 20]. Other attention-based ECG models, on the other hand, are implicitly based on a single (or effectively single) attention distribution over the temporal sequence. This attention can highlight important regions but is not explicitly designed to address complementary roles for diagnosis. It is possible that different types of evidence contribute differently to the final decision. For example, since rhythm irregularity is a diagnostic feature of AF, regular rhythm patterns and stable waveform morphology may be useful for reducing false AF alarms in normal sinus rhythm (NSR) samples. An evidence-oriented attention mechanism that can retrieve multiple complementary diagnostic representations might therefore provide a more natural way of organizing diagnostic evidence for short-window ECG classification.

The application of query-based representation learning provides an insightful perspective on this issue. Attention mechanisms of the transformer variant utilize the interaction between query, key, and value representations in order for queries to retrieve relevant information from contextual tokens [22]. With respect to object detection, learnable

queries have been introduced as task-specific retrievals of specific object-level representations from components of complex contextual features [23]. Moreover, query-oriented modeling can be applied in the field of sequence prediction, whereby queries facilitate the design of temporal context and optimization of information retrieval [24]. Inspired by this paradigm, we introduce query-centric diagnostic evidence retrieval into ECG temporal representation learning and treat short-window AF/NSR discrimination as a query-driven diagnostic evidence retrieval task. In essence, learnable diagnostic queries attend to BLSTM-encoded ECG temporal tokens, allowing different queries to retrieve complementary temporal representations associated with different AF/NSR-related discriminative patterns. In this manner, traditional global temporal aggregation in CNN-BLSTM is replaced with a query-centric diagnostic attention mechanism.

In this regard, we present QDA-BLSTM, which is a query-centric diagnostic attention-enhanced BLSTM network for detecting AF from short-window ECG signals. The architecture of our model includes a four-scale 1D CNN encoder, a channel attention module, a BLSTM temporal encoder, and a query-centric diagnostic attention module. The four-scale CNN encoder makes use of convolutional kernels designed for different receptive fields in order to extract local morphological features from ECG data. The channel attention module recalibrates the fused channels of features. The BLSTM encoder learns the bidirectional dependencies for rhythms within the 5-second ECG window. Diagnostic queries are used as query inputs for the multi-head attention. The keys and values are given by the BLSTM hidden states. The choice of five queries was inspired by the clinical problem of whether AF/NSR classification would involve multiple cues based on rhythm irregularities or regularities, atrial activity, stability of the QRS complex, and temporal context in the general sense. Nonetheless, no limits were assigned to the individual queries, meaning that their representations would be trained during the learning process.

The main contributions of our study are summarised as follows. First, we present QDA-BLSTM, a query-centric diagnostic attention framework for short-window ECG-based atrial fibrillation detection. The proposed model recasts AF/NSR classification as a query-driven temporal evidence retrieval problem, allowing for retrieval of discriminatively informative temporal representations from ECG sequences. Second, we present a unified model—combining multi-scale CNN feature extraction, channel-wise feature recalibration, BLSTM-based temporal modelling, and query-centric attention—such that multiple learnable queries can retrieve complementary temporal representations from a short ECG segment. Third, a comprehensive empirical evaluation is presented using a subject-independent split over an internal dataset from AFDB and NSRDB, plus an independent external validation on CPSC2018. In particular, this evaluation comprises comparisons against robust baseline models and diverse temporal aggregation methods, module and query ablation studies, sensitivity analysis of the number of queries, computational complexity metrics, McNemar test, and sensitivity and robustness analysis of the thresholds. Furthermore, we conduct query attention analysis to visually identify the specific time durations that were focused on by the queries.

2. Materials and Methods

The overall architecture of the QDA-BLSTM framework for short-window ECG-based AF is illustrated in Fig. 1. The framework consists of four components: a multi-scale feature encoder, a SE-like channel attention module, a BLSTM-based temporal modeling module, and a query-centric diagnostic attention (QDA) module followed by a classification layer. In more detail, the multi-scale encoder extracts morphological features from ECG signals via parallel 1D convolutional filters with different receptive fields. The channel attention module enhances informative feature channels and suppresses redundant representations. The refined features are then input into a BLSTM network to capture bidirectional temporal dependencies from short ECG segments. Finally, QDA introduces learnable diagnostic queries to retrieve complementary AF/NSR-related evidence from temporal representations for evidence-oriented feature aggregation and classification. The details of each component are described in the following subsections, including dataset and preprocessing in Section 2.1, the multi-scale 1D CNN encoder in Section 2.2, the SE-like channel attention module in Section 2.3, the BLSTM temporal encoder in Section 2.4, the query-centric diagnostic attention module in Section 2.5, and the overall architecture details in Section 2.6.

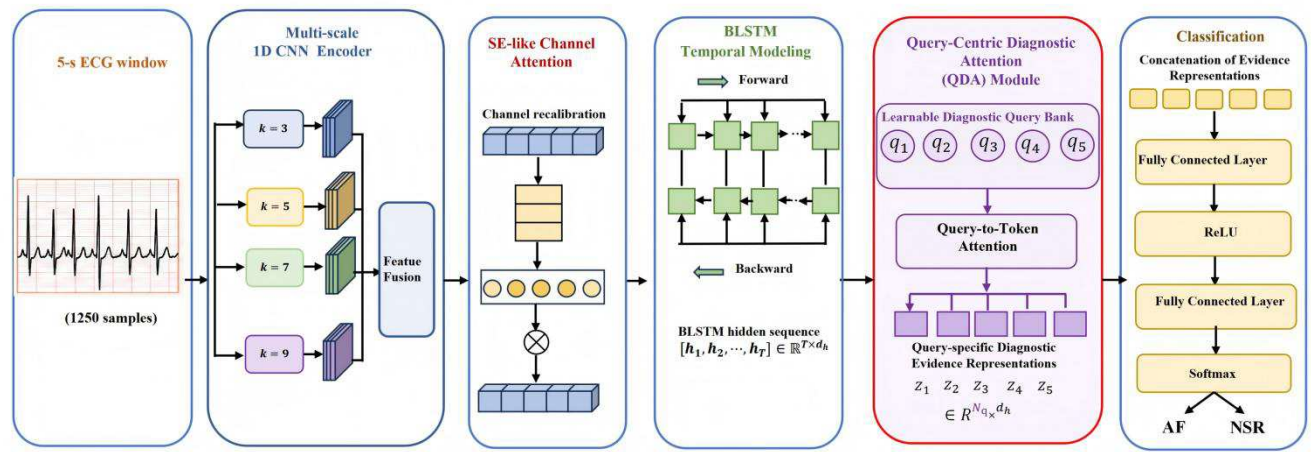


Figure 1. Overall framework of the proposed multi-module QDA-BLSTM architecture for short-window ECG-based AF detection

2.1 Dataset and preprocessing

In this study, signals obtained from short-window ECG recordings were applied to discriminate between atrial fibrillation and normal sinus rhythm, with the samples for atrial fibrillation originating from the MIT-BIH Atrial Fibrillation Database (AFDB). This database aggregates long-term ECG recordings of subjects with atrial fibrillation, mostly paroxysmal AF, at a sampling frequency of 250 Hz. Of the 25 AFDB subjects, only the 23 records containing ECG waveform data were included [25, 26]. The research work utilized samples from the MIT-BIH Normal Sinus Rhythm Database (NSRDB), which includes long-term ECG recordings of subjects without significant arrhythmias at a sampling rate of 128 Hz [25]. The AFDB dataset is not entirely composed of segments classified as AF, as manually prepared rhythm annotations are provided to localize intervals of AFIB. Previous studies have relied on the same principle, classifying each segment according to the rhythm annotations provided in the AFDB database. In the current study, we applied a stricter annotation-overlap criterion in order to diminish noise due to labeling segments with rhythm transitions or mixed rhythms. According to the current approach, 5-second-duration windows were assigned to the AF class if samples belonging to AFIB intervals represented 80% or more of the window content, and all other windows were disregarded.

The method for preparing data included resampling, band-pass filtering, splitting data into windows based on time, and normalizing data. NSRDB ECG signals were resampled from 128 Hz to 250 Hz, while AFDB ECG signals remained at 250 Hz. A Butterworth type band-pass filter (fourth order) was utilized to remove low-frequency baseline noise and high-frequency noise so that the ECG waveforms of interest were retained (the filter was set up using zero-phase filtering; the band-pass filter frequency range was 0.5–40 Hz) [27]. Five seconds of ECG data were used for each window (1250 total samples), and new windows (segments) were created every 2.5 seconds using a stride of 2.5 seconds. Channel 0 of the ECG signal was used for all experiments. To reduce amplitude variability between the ECG records in the analysis, each segment of data underwent z-score normalization. Up to 1500 segments were chosen from each record. The AFDB and NSRDB combined internal dataset included data from 41 subjects (23 subjects from AFDB and 18 subjects from NSRDB). The dataset was divided using a subject-independent, record-wise split, ensuring that all segments from the same original ECG record were assigned to only one subset and that there was no record-level leakage. The distribution of records and samples across the training, validation, and test subsets is summarized in Table 1.

Table 1. Dataset distribution of the internal AFDB/NSRDB experiment. Records denote original ECG recordings before window segmentation, whereas samples denote segmented 5-s ECG windows after preprocessing and class balancing.

Subset	Records	Samples	AF	NSR
Train	28	30352	15176	15176
Validation	4	6000	3000	3000
Test	9	12000	6000	6000

To further evaluate the cross-dataset generalization ability of the proposed model, CPSC2018 was used as an independent external validation dataset [28]. The external validation set was not used for model training, model selection, threshold optimization, or parameter tuning. Only records of AF and NSR were kept for the AF vs NSR binary classification task. Aiming to utilize a dataset with no overlapping records, the preprocessing pipeline for CPSC2018 was kept the same as that of our internal dataset. This consisted of resampling to 250 Hz, 5-s segmentation, 5-s z-score normalization, and a single-channel input setting, enabling the external validation experiment to assess whether our model learns general rhythm-specific representations rather than database-specific representations.

2.2 Multi-scale 1D CNN encoder

ECG signals have different types of waveforms with different temporal widths, such as P waves, QRS complexes, ST segments, and T waves. A single convolution kernel may not be sufficient to capture all multi-scale morphological patterns [2]. Given this, a multi-scale 1D CNN encoder was employed to extract local ECG features at different receptive fields. The kernel set was:

$$K=\{3,5,7,9\} \quad (1)$$

At each convolution scale k , this feature extraction can be expressed as:

$$F_k = \text{Pool}(\phi(\text{BN}(\text{Conv1D}_k(X)))) , k \in K \quad (2)$$

where $\text{Conv1D}_k(\cdot)$ is a one-dimensional convolution of kernel size k , $\text{BN}(\cdot)$ is batch normalization, $\phi(\cdot)$ is the ReLU activation function, and $\text{Pool}(\cdot)$ is max-pooling. The multi-scale features are fused as:

$$F_{\text{cat}} = \text{Concat}(F_3, F_5, F_7, F_9) \quad (3)$$

After concatenation, a fusion convolution is applied to integrate the multi-scale representations:

$$F = \text{Dropout}(\text{Pool}(\phi(\text{BN}(\text{Conv1D}(F_{\text{cat}}))))) \quad (4)$$

where F denotes the fused CNN feature map. In this way, the model is able to capture both the sharp local patterns associated with the QRS complex and also broader local variations in rhythm related to AF.

2.3 SE-like Channel Attention Module

After the multi-scale feature fusion, the SE-like channel attention module is used to recalibrate the channel-wise responses of the fused CNN feature map before temporal modeling [21]. Let the fused feature map be represented as:

$$F \in \mathbb{R}^{B \times C \times T}, \quad (5)$$

where B , C , T are batch size, number of feature channels, and temporal length, respectively. First, global average pooling along the temporal dimension is used to yield a channel descriptor:

$$s = \text{GAP}(F). \quad (6)$$

The channel attention weights are computed through a lightweight gating mechanism:

$$a = \sigma(W_2 \delta(W_1 s)) \quad (7)$$

where W_1 and W_2 are learnable transformations.

The recalibrated feature map is obtained by:

$$\tilde{F} = F \odot a \quad (8)$$

where $\odot$ denotes element-wise multiplication with broadcasting along the temporal dimension. This module thus allows the network to boost informative feature channels and suppress less useful ones before the BLSTM temporal encoder.

2.4 BLSTM temporal encoder

The recalibrated CNN feature map is then decoded into a sequence of temporal tokens to be fed to a BLSTM module. This BLSTM encoder is utilised to model bidirectional temporal dependencies in the 5-s ECG window:

$$H = \text{BLSTM}(\tilde{F}) \quad (9)$$

The BLSTM output is defined as[29]:

$$H = [h_1, h_2, \dots, h_T] \quad (10)$$

where h_t denotes the hidden representation of the t -th ECG token, B is batch size, and T is the number of temporal tokens after CNN encoding. Since a bidirectional LSTM is adopted, the hidden representation consists of both forward and backward temporal context:

$$H \in \mathbb{R}^{B \times T \times d_h} \quad (11)$$

where d_h is the hidden dimension of the BLSTM output. This is a useful modelling step for AF vs. NSR discrimination, as AF is characterised not only by local morphological changes but also by irregular temporal rhythm patterns.

2.5 Query-Centric Diagnostic Attention module

The Query-Centric Diagnostic Attention (QDA) module is at the core of the model. Instead of applying global pooling to the output of the BLSTM, our method employs multiple learnable queries to retrieve complementary temporal representations relevant to AF/NSR discrimination from the temporal context of the ECG. This design is inspired by the use of query-key-value attention in Transformer architectures and the use of learnable queries for task-specific representation retrieval [22, 23]. As demonstrated in Fig. 2, the BLSTM hidden sequence is treated as the temporal context, and instead of directly pooling this context, the key idea is that a small number of learnable diagnostic queries retrieve query-specific representations through query-key-value attention, which are then used for classification. Compared to direct global pooling, the proposed QDA module is designed to preserve complementary temporal information from short ECG segments.

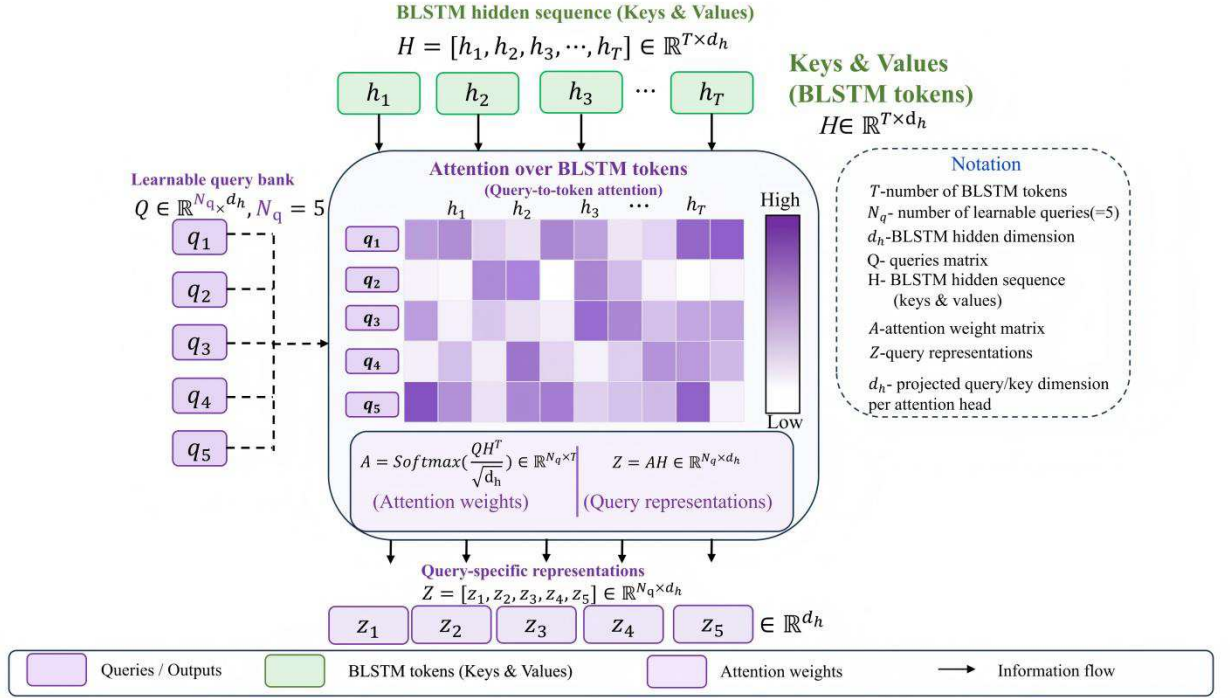


Fig. 2. Detailed structure of the Query-Centric Diagnostic Attention (QDA) module.

The learnable diagnostic queries attend to the BLSTM-encoded ECG temporal tokens through a query-key-value attention operation. The resulting attention matrix shows the relative attention assigned by each query to each ECG token, while the softmax-weighted aggregation produces query-specific representations from which AF/NSR prediction is performed.

For clarity in the following description, the batch dimension B is omitted. A set of $N_q=5$ learnable queries is defined as:

$$Q=[q_1, q_2, q_3, q_4, q_5] \in \mathbb{R}^{N_q \times d_h} \quad (12)$$

where N_q is the number of diagnostic queries and d_h is the size of the hidden dimension of the BLSTM output. The learnable queries are optimized together with the whole network during training.

The attention is computed with the query as the input, and the BLSTM hidden sequence is used as both keys and values. The projected queries, keys, and values are computed as:

$$Q' = QW_Q, K' = HW_K, V' = HW_V \quad (13)$$

where W_Q , W_K , and W_V are learnable projection matrices. For notational simplicity, the attention-head index is omitted. The query-token attention weights are obtained by scaled dot-product attention:

$$A = \text{Softmax}\left(\frac{Q'K'^T}{\sqrt{d_k}}\right) \quad (14)$$

where $A \in \mathbb{R}^{N_q \times T}$, d_k denotes the dimensionality of the projected query and key vectors in each attention head, and $A_{i,t}$ is the attention weight assigned to the t -th ECG token by the i -th learnable query. For M attention heads, $d_k = \frac{d_h}{M}$. The query-specific diagnostic representation is obtained as follows:

$$Z = AV' \quad (15)$$

In this study, $N_q=5$, so the QDA module outputs five query-specific representations for each ECG segment.

The obtained representations

$$Z=[z_1, z_2, z_3, z_4, z_5] \quad (16)$$

are concatenated and passed to the classification head to predict AF/NSR. Compared with direct temporal pooling, the query-centric diagnostic attention module is designed to retrieve complementary temporal representations from short ECG segments, which may encode information related to rhythm irregularity, NSR-related regular rhythm patterns, local waveform morphology, and global temporal context.

2.6 Architecture and experimental details

For a proper comparison, CNN1D, ResNet1D, BLSTM, Transformer Encoder, and CNN-BLSTM were chosen as benchmark models, ensuring that they employed the same data split, preprocessing methods, and training approach. CNN1D and ResNet1D were used to extract features from the dataset without focusing on temporal properties. Meanwhile, BLSTM was applied to prove the advantage of the temporal modeling method. CNN-BLSTM was identified as a key benchmark, as a similar multi-scale CNN encoder and BLSTM model were adopted as in the QDA-BLSTM framework. The difference was that unlike standard CNN-BLSTM, which made use of global average pooling while extracting temporal features, our method implemented the proposed query-centric diagnostic attention after the BLSTM operation. The Transformer Encoder as a benchmark was designed following a standard architecture using patch and positional embedding.

All models were optimised using AdamW with a learning rate of 1×10^{-3} and weight decay of 1×10^{-4} , and were trained with a batch size of 64. To prevent overfitting, dropout (0.4) and label smoothing (0.05) were employed. A cosine annealing learning rate scheduler was used during training. The main experiments were run with a random seed of 2024. To assess robustness and training stability, CNN-BLSTM and QDA-BLSTM were trained using five random seeds (2023 – 2027) under the same subject-independent split protocol. For the multi-seed analysis, two operating-threshold strategies were evaluated: a validation-set threshold selected by maximizing balanced accuracy and a fixed threshold of 0.5. We report accuracy (ACC), sensitivity (SEN), and specificity (SPE) for model performance evaluation. To statistically assess whether CNN-BLSTM and QDA-BLSTM have significantly different performance, we use McNemar’s test. This test works for paired classification results on the same test set, where the two classifiers produce discordant prediction pairs.

In order to isolate the impact of temporal aggregation on the performance, four aggregation techniques were applied after implementing a common multi-scale CNN, channel-recalibration module, and BLSTM structure: global average pooling, single-query attention, multi-head self-attention with the possibility of temporal mean pooling, and the five-query QDA method proposed by the authors. All configurations were obtained using the same set of data (subject-independent training set, data preprocessing configuration, training settings, criterion for choosing the best checkpoint, and random seed of 2024).

The measure of how QDA-BLSTM is influenced by the number of learnable queries was assessed for values of $N_q \in \{1, 3, 5, 7, 9\}$. All architectural and training factors remained constant except for the input size of the classifier. The use of 2024 as the random seed enabled this experiment to maintain its rigour. The multi-seed efficacy of CNN-BLSTM and the full five-query QDA-BLSTM will be discussed separately.

An input tensor of size $1 \times 1 \times 1250$ corresponding to a single 5-second ECG segment sampled at 250 Hz was used in the evaluation of parameter count and computational complexity. The parameter counts and MACs of the convolutional, normalization, pooling, recurrent, and linear layers were found using THOP. However, since THOP does not consider all aspects regarding the projectors and attention operations of PyTorch’s nn.MultiheadAttention module, the MACs corresponding to the query, key, and value projections and the attention operations were calculated separately and added to the THOP results. FLOPs were calculated as two times the number of MACs.

3. Results

3.1 Main comparison with strong baselines

Table 2 provides a comparison of the performance achieved by our QDA-BLSTM with five other baseline models on the internal AFDB/NSRDB test set. The models compared consist of CNN1D, ResNet1D, BLSTM, Transformer Encoder, CNN-BLSTM, and our QDA-BLSTM. The compared models are trained and evaluated on the same subject-independent split and preprocessing protocol. As shown in Table 2, QDA-BLSTM achieved the highest ACC on the internal AFDB/NSRDB test set. QDA-BLSTM correctly classified 99.42% of all tested ECG segments. The AF segments were classified correctly 99.70% of the time, while the NSR segments were correctly classified 99.15% of the time. ACC values for CNN1D, ResNet1D, BLSTM, Transformer Encoder, and CNN-BLSTM were 90.23%, 97.74%, 97.93%, 69.02%, and 98.17%, respectively. In the same order, SEN values were 99.78%, 97.45%, 96.35%, 99.98%, and 99.90%. Corresponding SPE values were 80.68%, 98.03%, 99.52%, 38.05%, and 96.43%.

Overall, QDA-BLSTM achieved the highest ACC and maintained a favorable balance between SEN and SPE among the compared models. In particular, compared with CNN-BLSTM, QDA-BLSTM increased ACC from 98.17% to 99.42% and SPE from 96.43% to 99.15%, corresponding to improvements of 1.25 and 2.72 percentage points, respectively. Meanwhile, SEN decreased slightly from 99.90% to 99.70%, but remained high. These results suggest that the query-centric diagnostic attention mechanism provides a better balance between AF sensitivity and NSR specificity than conventional global average pooling, particularly by reducing false AF predictions while preserving high AF sensitivity.

Table 2. Main comparison with strong baseline models on the internal test set

Model	ACC (%)	SEN (%)	SPE (%)
CNN1D	90.23	99.78	80.68
ResNet1D	97.74	97.45	98.03
BLSTM	97.93	96.35	99.52
Transformer Encoder	69.02	99.98	38.05
CNN-BLSTM	98.17	99.90	96.43
QDA-BLSTM	99.42	99.70	99.15

3.2 Comparison of temporal aggregation strategies

Table 3 illustrates the comparison of four temporal aggregation strategies with the help of the same CNN-BLSTM backbone and experimental protocol. Using global average pooling, an accuracy of 98.17% was achieved with 99.90% sensitivity and 96.43% specificity. When global average pooling was replaced with single-query attention, it resulted in an accuracy of 99.02%, specificity of 98.30%, along with a minor decline in sensitivity to 99.75%. Multi-headed self-attention helped in achieving 98.89% accuracy, 99.93% sensitivity, and 97.85% specificity.

The suggested five-query QDA was successful in achieving the best ACC and SPE with 99.42% and 99.15%, respectively, alongside still achieving a very high SEN of 99.70%. When compared to global average pooling, five-query QDA achieved an improvement in ACC of 1.25 percentage points and SPE of 2.72 percentage points, whereas slightly reducing SEN by 0.20 percentage points. Compared to multi-head self-attention, QDA managed to increase ACC and SPE by 0.53% and 1.30%, respectively, while slightly decreasing SEN by 0.23 percentage points. Results clearly imply that five-query QDA is capable of better balancing AF sensitivity and NSR specificity at the operating point determined by the validation.

Table 3. Comparison of temporal aggregation strategies on the internal AFDB/NSRDB test set

Aggregation strategy	ACC (%)	SEN (%)	SPE (%)
Global average pooling	98.17	99.90	96.43
Single-query attention	99.02	99.75	98.30
Multi-head self-attention	98.89	99.93	97.85

Aggregation strategy	ACC (%)	SEN (%)	SPE (%)
Five-query QDA	99.42	99.70	99.15

Note: The operating threshold of each model was selected on the validation set by maximizing balanced accuracy.

3.3 Ablation study

Table 4 shows the module-level ablation results of the proposed QDA-BLSTM on our internal AFDB/NSRDB test set. We observe that the full QDA-BLSTM obtains the highest ACC and SPE among all variants, with 99.42% ACC (SEN 99.70%, SPE 99.15%). After removing our multi-scale CNN encoder, the model classifies 98.43% of ECG segments correctly, and the SPE also drops to 96.88%. After removing the SE-like channel attention module, QDA-BLSTM still achieves 98.50% ACC (SEN 99.95%, SPE 97.05%). As for the variant without the QDA module, although its SEN is very high (99.98%), ACC is 96.48% and SPE is 92.98%. We observe that all components contribute to classification performance to different extents: removing the QDA module results in the most obvious decrease in SPE. This suggests that our query-centric diagnostic attention aids in retrieving complementary AF/NSR-related evidence and reducing false AF predictions. The multi-scale CNN encoder and channel attention module improve performance by enhancing local feature extraction and recalibrating the informative CNN channels, respectively. These results show that the improvement of QDA-BLSTM is not simply due to module stacking but rather to the contributions of multi-scale feature extraction, channel recalibration, temporal modelling, and query-centric diagnostic evidence retrieval.

Table 4. Ablation study of the proposed QDA-BLSTM on the internal AFDB/NSRDB test set

Model variant	ACC (%)	SEN (%)	SPE (%)
QDA-BLSTM w/o Multi-scale CNN	98.43	99.97	96.88
QDA-BLSTM w/o Channel Attention	98.50	99.95	97.05
QDA-BLSTM w/o QDA	96.48	99.98	92.98
QDA-BLSTM	99.42	99.70	99.15

Note: w/o denotes “without” the corresponding module

In Table 5, there is an indication of how the elimination of each of the learnable queries impacts the performance of QDA-BLSTM. As mentioned previously in this section, the full QDA-BLSTM achieved 99.42% ACC, 99.70% SEN, and 99.15% SPE. In regards to the comparisons of ACC values derived after the omission of Query-1, Query-3, Query-4, and Query-5, the same model has produced the figures of 98.84%, 98.40%, 99.19%, and 98.77%, respectively. In consideration of the SPE figures for the models discussed, there are the respective figures of 97.68%, 96.80%, 98.38%, and 97.58%. All the values in Table 5 indicate the fact that the removal of Query-2 leads to a drop of ACC to 95.00% and SPE to 90.00%, while SEN remains equal to 100.00%. The figures for the false positives demonstrate how the number has seen a huge leap from 51 in the initial experiment when QDA used a full number of queries to 600, thus demonstrating the crucial role of Query-2 in terms of maintaining NSR specificity and reducing/avoiding false predictions of AF. The high SEN values for all of the various queries that were excluded point out that the AF segments can still be detected whether a certain query is omitted or not; nonetheless, the fall in SPE means that various queries facilitate the differentiation between NSR samples and AF-like patterns. From this standpoint, the multi-query design of QDA-BLSTM plays a role in preserving NSR specificity while achieving a high degree of sensitivity to AF.

376 Table 5. Query-level ablation study of QDA-BLSTM on the internal AFDB/NSRDB test set

Model	TN	FP	FN	TP	ACC (%)	SEN (%)	SPE (%)
QDA-BLSTM w/o Query-1	5861	139	0	6000	98.84	100.00	97.68
QDA-BLSTM w/o Query-2	5400	600	0	6000	95.00	100.00	90.00
QDA-BLSTM w/o Query-3	5808	192	0	6000	98.40	100.00	96.80
QDA-BLSTM w/o Query-4	5903	97	0	6000	99.19	100.00	98.38
QDA-BLSTM w/o Query-5	5855	145	3	5997	98.77	99.95	97.58
QDA-BLSTM	5949	51	18	5982	99.42	99.70	99.15

377 Note: w/o denotes "without" the corresponding query. TN = True Negative, FP = False Positive,
 378 FN = False Negative, and TP = True Positive.

3.4 Sensitivity to the number of learnable queries

380 The relationship between the number of learnable queries and classification performance has been assessed in
 381 Table 6. No clear trend was observed in how the number of queries impacts classification performance. ACC and
 382 SPE scores achieved by using one query were 99.02% and 98.30%, respectively, while increasing the number of
 383 queries from one to three led to a deterioration in performance to 98.15% and 96.53%, respectively. The same
 384 tendency was observed with the seven-query configuration, which also demonstrated lower values of specificity.
 385 Hence, the increasing size of the query bank does not guarantee success.

386 The five-query configuration was able to maximize the values of ACC and SPE at the selected operating point,
 387 yielding values of 99.42% ACC, 99.70% SEN, and 99.15% SPE. When increasing the number of queries to nine, no
 388 improvement in ACC or SPE was achieved as compared to the five-query configuration. The results support the
 389 conclusions regarding the use of five queries in the proposed model.

390 Table 6. Sensitivity of QDA-BLSTM to the number of learnable queries

Number of queries	ACC (%)	SEN (%)	SPE (%)
1	99.02	99.75	98.30
3	98.15	99.77	96.53
5	99.42	99.70	99.15
7	98.59	99.63	97.55
9	99.08	99.78	98.37

391 Note: All configurations used the same data split, training protocol,
 392 and random seed of 2024. Thresholds were selected using validation-set
 393 balanced accuracy.

3.5 Computational complexity

395 The number of parameters and computational cost for the four aggregation methods are shown in Table 7. Global
 396 average pooling was the least complex based on the number of parameters (0.155 million) and MACs (63.995 million)
 397 that it used. The proposed five-query QDA method used more parameters and MACs (0.371 million and 75.361

million, respectively), increasing the number of MACs by approximately 17.8% compared with global pooling. Despite the additional computational cost of the five-query QDA, it produced significant improvements in ACC and SPE, while maintaining a similar level of SEN.

The amount of MAC calculations has only marginally increased by 1.1%, from 74.574 million to 75.361 million, when the number of learnable queries was increased from one to five. However, ACC improved from 99.02% to 99.42%. The complete version of multi-head self-attention needed a total of 130.289 million MACs, whereas five-query QDA reduced this number by 42.2% while providing superior results in terms of classification performance. Therefore, it is inaccurate to refer to QDA technology as a method with the lowest level of complexity among aggregation methods.

Table 7. Computational complexity of the temporal aggregation strategies

Aggregation strategy	Parameters (M)	MACs (M)	Approx. FLOPs (M)
Global average pooling	0.155	63.995	127.990
Single-query attention	0.304	74.574	149.148
Multi-head self-attention	0.287	130.289	260.577
Five-query QDA	0.371	75.361	150.721

Note: THOP-derived MACs were supplemented with analytically calculated attention operations. FLOPs were approximated as twice the number of MACs.

3.6 Statistical significance analysis

As shown in Table 8, for the internal test set, there were 249 discordant pairs between CNN-BLSTM and QDA-BLSTM. Of these, 49 samples were classified correctly by CNN-BLSTM but incorrectly by QDA-BLSTM, while 200 samples were classified incorrectly by CNN-BLSTM but correctly by QDA-BLSTM. The p-value obtained was 7.86×10^{-23} , indicating a statistically significant difference between the two models. On the external CPSC2018 validation set, the difference was again statistically significant, with 683 discordant pairs, including 172 samples classified correctly by CNN-BLSTM but incorrectly by QDA-BLSTM and 511 samples classified incorrectly by CNN-BLSTM but correctly by QDA-BLSTM. The p-value was 6.54×10^{-40} , suggesting that the improvement of QDA-BLSTM over CNN-BLSTM is unlikely to be due to random variation.

Table 8. Statistical significance analysis using McNemar’s test

Dataset	Discordant pairs	CNN correct / QDA wrong	CNN wrong / QDA correct	p-value
Internal test set	249	49	200	7.86e-23
CPSC2018 external validation	683	172	511	6.54e-40

3.7 Robustness and threshold sensitivity analysis

For a further evaluation of the robustness of the training and the sensitivity to thresholds, both CNN-BLSTM and QDA-BLSTM were retrained multiple times with the use of five random seeds ranging from 2023 to 2027 under the same subject-independent split. Two threshold strategies were used: the validation-based threshold selected by maximizing balanced accuracy and the fixed threshold of 0.5. The results are reported in the form of mean $\pm$ standard deviation in Table 9.

Under the validation threshold strategy, QDA-BLSTM performed better with a lower standard deviation value as compared to CNN-BLSTM. For example, the QDA-BLSTM method achieved $98.84 \pm 0.66\%$ ACC, $99.75 \pm 0.27\%$ SEN and $97.93 \pm 1.43\%$ SPE as compared to CNN-BLSTM's $97.73 \pm 1.39\%$ ACC, $99.58 \pm 0.29\%$ SEN and $95.88 \pm 2.73\%$ SPE. The fixed threshold analysis also indicated that QDA-BLSTM had less sensitivity to threshold selection. When the fixed threshold of 0.5 was used, CNN-BLSTM achieved an ACC of $95.43 \pm 4.25\%$ along with a SPE of $90.91 \pm 8.50\%$, although the SEN was retained at $99.94 \pm 0.06\%$. In comparison, the QDA-BLSTM method achieved an ACC of $98.33 \pm 0.83\%$ along with a SPE value of $96.73 \pm 1.64\%$.

Table 9. Robustness and threshold sensitivity analysis across five random seeds

Model	Threshold strategy	ACC (%)	SEN (%)	SPE (%)
CNN-BLSTM	Validation BA	97.73 ± 1.39	99.58 ± 0.29	95.88 ± 2.73
QDA-BLSTM	Validation BA	98.84 ± 0.66	99.75 ± 0.27	97.93 ± 1.43
CNN-BLSTM	Fixed 0.5	95.43 ± 4.25	99.94 ± 0.06	90.91 ± 8.50
QDA-BLSTM	Fixed 0.5	98.33 ± 0.83	99.94 ± 0.06	96.73 ± 1.64

Note: Values are reported as mean $\pm$ standard deviation over five random seeds (2023-2027). Validation BA indicates threshold selection using validation-set balanced accuracy, whereas Fixed 0.5 uses a constant decision threshold of 0.5.

3.8 External validation on CPSC2018

To further study generalisation across datasets, we assess the trained CNN-BLSTM and QDA-BLSTM on the external CPSC2018 dataset not used to train, validate or select checkpoints or tuning, giving a more stringent measure of transfer to ECGs from other sources.

As presented in Table 10, QDA-BLSTM outperformed CNN-BLSTM on the external CPSC2018 validation set. CNN-BLSTM achieved an ACC of 89.71%, a SEN of 84.49%, and an SPE of 96.45%. In contrast, QDA-BLSTM achieved an ACC of 92.97%, a SEN of 89.46%, and an SPE of 97.51%. Compared with CNN-BLSTM, QDA-BLSTM demonstrated an improvement of 3.26 ACC points, 4.97 SEN points, and 1.06 SPE points. The confusion matrix on external data (Table 10) showed that QDA-BLSTM had a reduced number of false negatives and false positives compared with CNN-BLSTM. For CNN-BLSTM, there were 4374 true negatives, 161 false positives, 909 false negatives, and 4953 true positives. When employing the QDA-BLSTM model, a total of 4422 true negatives, 113 false positives, 618 false negatives and 5244 true positives were obtained. Compared with CNN-BLSTM, this represents a reduction in false negatives, which indicates enhanced AF detection on the external dataset, while the reduction in false positives indicates improved NSR recognition.

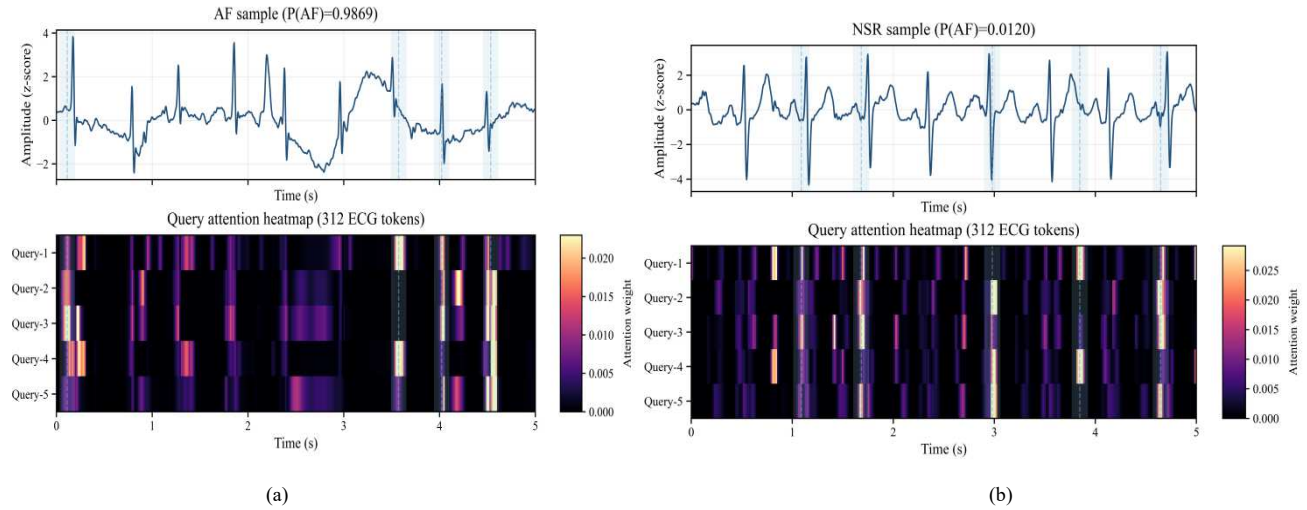
Table 10. External validation performance on CPSC2018

Model	TN	FP	FN	TP	ACC (%)	SEN (%)	SPE (%)
CNN-BLSTM	4374	161	909	4953	89.71	84.49	96.45
QDA-BLSTM	4422	113	618	5244	92.97	89.46	97.51

3.9 Query attention visualization

To qualitatively examine the behaviour of the learned queries, we visualised the query attention responses of QDA-BLSTM in Fig. 3. The upper half of each panel shows the original 5-s ECG waveform, and the lower part

457 displays a heatmap of the attention weights of five learnable diagnostic queries over 312 BLSTM-encoded temporal
 458 tokens. Query-1 to Query-5 refer to the five learnable query embeddings in the QDA module. Darker areas indicate
 459 higher attention responses. The vertical dashed lines provide approximate alignment of high-attention locations with
 460 the corresponding ECG waveform points during this segment.



462 Fig. 3. Visualization of query-centric diagnostic attention responses for AF and NSR examples

464 According to the AF case (Fig. 3(a)), the model yields an AF probability value of 0.9869. While the attention
 465 focus seems to be located at a few time points such as 0.112 s, 3.574 s, 4.022 s, and 4.535 s, they correspond well
 466 with specific points of magnitude where features related to the AF prediction could be found. In the NSR case shown
 467 in Fig. 3(b), the AF probability is only 0.0120. Attention is mainly focused at the time points 1.090 s, 1.683 s, 2.981
 468 s, 3.846 s, and 4.647 s, which also exist within the range of several cardiac cycles. These time points may express
 469 some temporal characteristics, which should result in the proper NSR decisions of the model. Since no event-level
 470 clinical markers were employed, the given spots should be treated as qualitative attention patterns only rather than a
 471 direct reflection of clinical cues.

472 4. Discussion

473 Table 11. Comparative summary of representative AF detection methods involving AFDB

Ref.	Year	Method	Input / Features	Validation / Split	Reported Performance
Yue et al. [5]	2021	FSWT + ML classifiers	5 s ECG segments + frequency-slice wavelet features	10-fold cross-validation	ACC = 98.15%, SEN = 96.43%, SPE = 100%
Duan et al. [8]	2022	RRI histogram + SVM	R-R interval histograms	10-fold cross-validation	ACC = 96.97%, SEN = 95.24%, SPE = 99.94%
Wan et al.[9]	2024	Ensemble learning + multi-feature discrimination	Time + frequency + nonlinear features	cross-validation	ACC > 99%
Petmezas et al. [11]	2021	Hybrid CNN-LSTM	Raw ECG segments	10-fold cross-validation	SEN = 97.87%, SPE = 99.29%
Wang[12]	2021	Modified BiLSTM with attention	2 s ECG segments	subject-independent split	ACC = 99.1%, SEN = 98.6%, SPE = 99.3%

Luo et al. [16]	2024	Directional-statistics-inspired end-to-end model	the distribution patterns of R-R intervals	5-fold cross-validation	ACC = 96.2%, SEN = 98.2%, SPE = 97.3%
Li et al. [14]	2024	LCG-Net (lightweight cross-guidance)	15 s ECG	Not explicitly specified	ACC = 99.66%, SEN = 99.18%, SPE = 99.88%
Morvari & Moghaddam [17]	2024	Hybrid CNN + Ant-Lion Optimization	STFT spectrogram images	10-fold cross-validation	ACC = 99.40% SEN = 99.72% SPE = 98.95%,
Wang et al. [18]	2025	Transformer-Based Object Detection	ECG segments for AF beat localization and classification	cross-validation	ACC = 92.97%, SEN = 77.15%, SPE = 95.96%
Our method	–	QDA-BLSTM	5 s raw ECG segments	subject-independent split	ACC = 99.42%, SEN = 99.70%, SPE = 99.15%

Table 11 summarizes representative AF detection methods evaluated on or utilizing the MIT-BIH Atrial Fibrillation Database (AFDB), which serves as a methodological reference for comparing our proposed QDA-BLSTM with existing approaches. Some are traditional methods based on handcrafted features, while others are deep learning-based methods.

Traditional methods use ECG descriptors generated or designed manually followed by classifier construction. One study using the frequency slice wavelet transform and machine learning classifiers was conducted by Yue et al. [5] to detect short-term AF segments, and another one based on histogram features of R-R intervals proposed by Duan et al. [8] relies on the spacing of R peaks and uses a support vector machine. Wan et al. [9] employed ensemble learning on features from time domain, frequency domain, and nonlinear domains to produce their output. They are more interpretable as their algorithm makes decisions based on relatively explicit rhythm, statistical properties, and time/frequency descriptors; however, these algorithms heavily depend on signal preprocessing, R-peak detection, feature generation and feature selection. In short, as discussed in our work, on short windows there are relatively fewer cycles, making rhythm-level statistics less stable, and subtle atrial activity and rhythm patterns may be obscured in pre-defined features.

Methods based on deep learning provide a different approach to automatically learn discriminative ECG representations for identification. Petmezas et al. [11] proposed a hybrid CNN-LSTM model, and Wang [12] proposed a modified BiLSTM with a subsequent feature recalibration and attention mechanism. This demonstrates the benefits of local feature extraction as well as modeling temporal dependencies. More recent works have ventured into rhythm-oriented end-to-end modeling, lightweight feature-fusion networks, optimized CNN-based architectures, and Transformer-based localization frameworks [14, 16-18]. These developments show that AF detection models are evolving from classical CNN-LSTM classifiers to more localized, “smart”, lightweight, and optimized architectures. It is interesting to note that the Transformer Encoder is severely underperforming compared with other models in this study. This performance drop is likely due to insufficient training data and the lack of inductive bias present in vanilla Transformers for modeling local morphological patterns in short-window ECGs.

Despite this positive movement toward reducing these existing limitations and risks, there are still several limitations with existing methods. First, some methods use ECG-derived R-R intervals, wavelet coefficients, and time-frequency images and representations, hence minimizing the benefits of fully end-to-end learning from the raw ECG signal. Second, in many existing works, there is reliance on standard cross-validation and intra-patient sampling procedures, especially with ECG segments taken from the same record or patient being part of both training and test datasets, leading to an inflation of performance estimates and thus they become optimistic estimates. Finally, with many DL models, overall performance may improve as existing components are combined with additional layers (i.e., residual-dense modules, recurrently stacked components, attention fusions, transformed 2D inputs, CNN-based optimization methods, and transformers for object detection). While it is often observed that these methods increase the overall ability to learn features and capture fine-grained details for beat-level localization, these methods also typically increase computational complexity and subsequently decrease the system’s ability to be deployed in scenarios related to wearable or edge devices.

In comparison to these existing methods, we propose a QDA-BLSTM specifically tailored for short-window AF detection from raw single-lead ECG signals. In contrast to classical machine learning methods [5, 8, 9], in which

extracted features such as R-R interval, HRV, wavelet features, entropy, and nonlinear features are typically used, our QDA-BLSTM directly learns discriminative representations from 5-s raw ECG segments. When compared to other works based on CNN-LSTM or BiLSTM, such as those in [11-13], QDA-BLSTM has a specialized mechanism that is query-based, which has the ability to obtain rhythm-based information from BLSTM-encoded ECG tokens within 5-s segments. Based on this query mechanism, QDA-BLSTM can collect information from ECG segments with various contexts, thus resulting in greater robustness in short-window AF detection. With Transformer-based methodologies such as AF-DETR [18], the purpose is more closely related to AF localization or beat-level detection, while our paper looks at segment-level AF/NSR discrimination within short time windows. Hence, we cite AF-DETR as related work even though the objectives differ.

The potential explanation for the observed improvement can be linked to how our QDA aggregation module is set up. The baseline CNN-BLSTM architecture uses global average pooling on the BLSTM output, resulting in the loss of various types of temporal information. However, with the QDA module included in our design, the module can utilize multiple learnable queries to extract different complementary types of temporal information from ECG tokens encoded using the BLSTM. Different queries in this architecture can tune into distinct temporal processes for AF/NSR categorization, though they are not tagged with prior clinical meanings as other methods do. The clearest evidence of this feature can be observed in the ablation test results. Thus, QDA is seen to boost NSR specificity and reduce the false AF prediction rate while also achieving high levels of AF sensitivity. The visualization results shown in Fig. 3 illustrate how the learned queries yield diverse temporal engagement parameters, however, they are qualitative observations and hence do not imply that the queries should be attributed to any particular clinical aspect.

This conclusion is based on further investigations. As seen in Table 3, five-query QDA is markedly better than both global average pooling and multi-head self-attention in terms of ACC and SPE scores while achieving high SEN at the same time. In other words, five-query QDA demonstrates the best combination of AF sensitivity and NSR specificity. In addition, as stated in Table 6, the performance of five-query QDA does not follow a monotonous trend with increasing query numbers, with five queries achieving the optimal value. Meanwhile, Table 7 shows that five-query QDA requires 17.8% more MACs compared with global average pooling and 42.2% fewer MACs than multi-head self-attention, thereby providing a good performance–cost trade-off.

Notably, we selected AF windows based on rhythm annotations using a subject-independent split strategy. Compared with random segment-level splitting, this setting reduces the possibility of information leakage at the record level and provides a stricter evaluation of model generalisation. In this setting, QDA-BLSTM achieved 99.42% ACC, 99.70% SEN, and 99.15% SPE on the internal test set. Compared with representative AFDB-based methods, our framework achieved a favourable trade-off among diagnostic accuracy, AF episode sensitivity, and robustness across independent ECG records. An additional validation on CPSC2018 provides further evidence of the potential applicability of QDA-BLSTM to ECG data collected from different sources.

It should be noted that numerical comparisons should be viewed in a cautious manner. Each study can have different window lengths, different modes of input (e.g., sequences versus transformed time-frequency inputs), different rules for labeling, and so forth. For example, one study uses R-R interval sequences, while another study may use hand-crafted features of rhythm, and yet another study may optimize a CNN or split by patient or perform cross-validation across patients, which makes direct comparison with the current work difficult. The current study utilizes a subject-independent split for internal evaluation and an independent external dataset for validation. The purpose of Table 11 is to indicate that QDA-BLSTM achieves competitive performance under a fairly stringent evaluation process while maintaining an end-to-end short-window ECG learning framework.

While the performance of the proposed QDA-BLSTM is promising, we would also like to acknowledge several current limitations. Firstly, our current model is validated on single-channel ECGs, which lends itself well to wearable monitoring but means that a significant part of the spatial information in multi-lead ECGs is lost. We will look at extending the query-centric mechanism proposed in Section 2.5 to multi-lead data, particularly exploring cross-lead complementary representations. Secondly, we focus on 5-s segments of ECG, which are suitable for fast AF screening, but may be less robust in rhythm-transition regions, very short paroxysmal AF episodes, and noisy recordings. A hierarchical short-to-long scheme of temporal modelling may help boost robustness in practice. In addition, the AF and NSR datasets involved in the internal evaluation were sourced from the AFDB and NSRDB, potentially introducing specific biases during acquisition and in terms of population composition, which may influence classification performance. Although subject-independent division limits the risk of leaking record-level information, this does not eliminate the possibility of such bias having an effect. Although external validation was performed using CPSC2018, further research exploring the classification capabilities of the approach presented here in a broader range of recording environments and patient groups is needed. Additionally, it will be necessary to validate the performance of the proposed approach in real-time applications.

5. Conclusion

In the present study, we proposed a novel query-centric diagnostic attention-based BLSTM network named QDA-BLSTM for short-window ECG-based AF detection. We employed a subject-independent internal evaluation protocol along with an external independent dataset to conduct a feasibility analysis of the proposed network. The experimental results showed that the classification performance of QDA-BLSTM was superior to that of conventional CNN-BLSTM models, with positive results obtained on the external CPSC2018 dataset as well. The main contribution of this work was a novel query-centric diagnostic attention technique that provided the model with the ability to retrieve relevant temporal representations that enabled it to discriminate between AF and NSR with a reasonable performance-cost trade-off. Moreover, we provided an overview of the performance of the proposed framework using comparative experiments, ablation experiments, sensitivity experiments, complexity experiments, statistical analysis, multi-seed evaluation, and attention visualization. In the future, we will enhance our proposed technique with multi-lead ECG data, real-world wearable recordings, longer ECG histories, and verification of the ability to generalize to wearable applications along with its feasibility.

Declarations: Ethics and Consent to Participate declarations: Not applicable.

Author contributions: **Yang Li:** Conceptualization, Methodology, Writing-original draft. **Jianguo chen:** Writing-original draft. **Manhong Shi:** Methodology, Review & editing, Funding acquisition

Declaration of competing interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding : This work was supported by Anhui Province university natural science research key project in 2023, grant number 2023AH051865; The key Discipline Construction Project of Anhui Science and Technology University, grant number XK-XJGY002; Anhui Province Quality Engineering Project, grant number 2025dzxkc067; Anhui Science and Technology University Stabilization and Introduction of Talents, grant number XWWD202101.

Acknowledgments: None.

Data availability statement: The datasets analysed during the current study are publicly available. The MIT-BIH Atrial Fibrillation Database (AFDB) is available at <https://physionet.org/content/afdb/1.0.0/>; the MIT-BIH Normal Sinus Rhythm Database (NSRDB) is available at <https://physionet.org/content/nsrdb/1.0.0/>; and the CPSC2018 dataset is available through the China Physiological Signal Challenge 2018 repository at <https://2018.icbeb.org/Challenge.html>.

References

1. Jogle JA, Chung MK, Armbruster AL, Benjamin EJ, Chyou JY, Cronin EM, Deswal A, Eckhardt LL, Goldberger ZD, Gopinathannair R *et al*: **2023 ACC/AHA/ACCP/HRS Guideline for the Diagnosis and Management of Atrial Fibrillation: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines.** *Circulation* 2024, **149**(1):e1-e156.
2. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY: **Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network.** *Nature medicine* 2019, **25**(1):65-69.
3. Clifford G, Liu C, Moody B, Lehman L, Silva I, Li Q, Johnson A, Mark R: **AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge 2017.** In: *Computing in Cardiology: 2017*; 2017: 1-4.
4. Zabihi M, Bahrami Rad A, Katsaggelos AK, Kiranyaz S, Narkilahti S, Gabbouj M: **Detection of atrial fibrillation in ECG hand-held devices using a random forest classifier.** In: *Computing in Cardiology: 2017*; 2017: 1-4.

5. Yue Y, Chen C, Liu P, Xing Y, Zhou X: **Automatic Detection of Short-Term Atrial Fibrillation Segments Based on Frequency Slice Wavelet Transform and Machine Learning Techniques.** *Sensors (Basel, Switzerland)* 2021, **21**(16):5302.
6. Kumar M, Pachori RB, Rajendra Acharya U: **Automated diagnosis of atrial fibrillation ECG signals using entropy features extracted from flexible analytic wavelet transform.** *Biocybernetics and Biomedical Engineering* 2018, **38**(3): 564 – 573.
7. Rouhi R, Clausel M, Oster J, Lauer F: **An Interpretable Hand-Crafted Feature-Based Model for Atrial Fibrillation Detection.** *Frontiers in Physiology* 2021, **12**:657304.
8. Duan J, Li R, Yang L: **Accurate detection of atrial fibrillation events with R-R intervals from ECG signals.** *PLOS ONE* 2022, **17**(8): e0271596
9. Wan X, Liu Y, Mei X, Ye J, Zeng C, Chen Y: **A novel atrial fibrillation automatic detection algorithm based on ensemble learning and multi-feature discrimination.** *Medical & biological engineering & computing* 2024, **62**(6):1809-1820.
10. Ping Y, Chen C, Wu L, Wang Y, Shu M: **Automatic Detection of Atrial Fibrillation Based on CNN-LSTM and Shortcut Connection.** *Healthcare* 2020, **8**(2):139.
11. Petmezas G, Haris K, Stefanopoulos L, Kilintzis V, Tzavelis A, Rogers JA, Katsaggelos AK, Maglaveras N: **Automated Atrial Fibrillation Detection using a Hybrid CNN-LSTM Network on Imbalanced ECG Datasets.** *Biomedical Signal Processing and Control* 2021, **63**:102194.
12. Wang J: **An intelligent computer-aided approach for atrial fibrillation and atrial flutter signals classification using modified bidirectional LSTM network.** *Information Sciences* 2021, **574**:320-332.
13. Laghari AA, Sun Y, Alhussein M, Aurangzeb K, Anwar MS, Rashid M: **Deep residual-dense network based on bidirectional recurrent neural network for atrial fibrillation detection.** *Scientific reports* 2023, **13**(1):15109.
14. Li Y, Chen M, Jiang Xe, Liu L, Han B, Zhang L, Wei S: **An atrial fibrillation detection algorithm based on lightweight design architecture and feature fusion strategy.** *Biomedical Signal Processing and Control* 2024, **91**:106016.
15. Ben-Moshe N, Tsutsui K, Brimer SB, Zvuloni E, Srnm L, Behar JA: **RawECGNet: Deep Learning Generalization for Atrial Fibrillation Detection From the Raw ECG.** *IEEE Journal of Biomedical and Health Informatics* 2024, **28**(9):5180 – 5188.
16. Luo C, Zhang K, Hu Y, Li X, Cao S, Jin Y, Ren P, Rao N: **Directional statistics-inspired end-to-end atrial fibrillation detection model based on ECG rhythm.** *Expert Systems with Applications* 2024, **247**:123112.
17. Morvari A, Moghaddam RK: **Detecting atrial fibrillation from ECG signal using hybrid convolutional neural network with ant-lion optimization.** *Multimedia Tools & Applications* 2024, **83**(8):24301-24313.
18. Wang P, Song J, Wu P, Li Z, Chen X, Du L, Fang Z: **AF-DETR: Transformer-Based Object Detection for Precise Atrial Fibrillation Beat Localization in ECG.** *Bioengineering* 2025, **12**(10).
19. Mousavi SS, Afghah F, Razi A, Acharya UR: **ECGNET: Learning Where to Attend for Detection of Atrial Fibrillation with Deep Visual Attention.** In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. 2019: 1-4.
20. Mousavi S, Afghah F, Acharya UR: **HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks.** *Computers in biology and medicine* 2020, **127**:104057.

21. Hu J, Shen L, Sun G, Albanie S: **Squeeze-and-Excitation Networks**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2020, **42**(8):2011 – 2023.
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I: **Attention is all you need**. *Advances in Neural Information Processing Systems* 2017, **30**:5998 – 6008.
23. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S: **End-to-End Object Detection with Transformers**. In: *European Conference on Computer Vision (ECCV)*. 2020: 213 – 229.
24. Zhou Z, Wang J, Li YH, Huang YK: **Query-Centric Trajectory Prediction**. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023: 17863 – 17873.
25. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE: **PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals**. *Circulation* 2000, **101**(23):E215.
26. Moody GB, Mark RG: **The impact of the MIT-BIH arrhythmia database**. *IEEE Engineering in Medicine and Biology Magazine* 2002, **20**(3):45-50.
27. Sörnmo L, Laguna P: **Bioelectrical Signal Processing in Cardiac and Neurological Applications**. *Elsevier Academic Press* 2005.
28. Liu F, Liu C, Zhao L, Zhang X, Wu X, Xu X, Liu Y, Ma C, Wei S, He Z *et al*: **An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection**. *Journal of Medical Imaging and Health Informatics* 2018, **8**(7):1368 – 1373.
29. Schuster M, Paliwal KK: **Bidirectional recurrent neural networks**. *IEEE Transactions on Signal Processing* 1997, **45**(11):2673-2681.