Computer Methods and Programs in Biomedicine 273 (2026) 109101

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Lightweight element-wise product enhanced neural network for efficient arrhythmia detection on embedded devices☆

Haotian Tang[1], Mingke Yan[1], Xidong Wu, Hongkai Lai, Yang Zhang, Hanyu Cui, Liping Xie*

College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, 110169, Shenyang, China

ARTICLE INFO

Keywords:
Artificial intelligence
Embedded device
Health monitoring
ECG

ABSTRACT

Background and Objective: Cardiovascular disease has emerged as a critical global health concern, where early and accurate arrhythmia detection is essential for preventing life-threatening complications and improving clinical outcomes. While deep learning approaches have shown promise in automated ECG-based arrhythmia classification, the development of computationally efficient models for resource-constrained embedded devices presents significant challenges. This study aims to address this gap by proposing a lightweight model optimized for real-time arrhythmia detection on embedded platforms.

Methods: We introduce an elementwise-product enhanced lightweight model (EPLM) featuring spindle-shaped architectures integrated with depthwise separable convolutions. This design reduces model parameters and computational costs while maintaining robust feature extraction capabilities. A novel elementwise product fusion mechanism is employed to combine dual-path features, enhancing high-dimensional feature representation. The model is rigorously evaluated using the standard MIT-BIH, SVDB, INCART, and PTB databases, with performance metrics including classification accuracy, precision, recall, F1-score, computational complexity, inference time and power consumption. Practical validation is conducted via deployment on a Raspberry Pi 5 and a smartphone to assess real-time applicability.

Results: The proposed model achieves classification accuracies of 99.10% and 98.85% on the MIT-BIH and PTB databases, respectively. It requires only 25,981 parameters and 525,400 floating-point operations per second (FLOPs), enabling a single inference time of 2.29 ms. Deployment on Raspberry Pi 5 and Android 10 ×86 virtual machine confirm real-time operation, demonstrating suitability for wearable applications.

Conclusions: The EPLM framework delivers ultra-efficient arrhythmia detection with minimal computational overhead, enabling continuous cardiac monitoring on embedded devices. This advancement holds promise for early clinical intervention, improved patient outcomes, and scalable deployment in resource-limited settings. The model's efficiency and accuracy underscore its potential to transform real-time wearable healthcare technologies.

1. Introduction

Cardiovascular disease (CVD) has become one of the most life-threatening diseases worldwide [1]. In Europe, CVD accounts for over half of all deaths, which is 46 times greater than the combined total of deaths from acquired immune deficiency syndrome (AIDS), malaria, and tuberculosis [2]. In Asia, CVD accounts for a significant 35 percent of all deaths [3]. Arrhythmias, characterized by abnormal heart rhythms, pose a significant risk within the broad category of CVD. Arrhythmias manifest as heartbeats that are too rapid, too slow, or irregular, potentially leading to inadequate blood circulation to critical organs and resulting organ damage. Severe arrhythmia can trigger sudden cardiac arrest, which is potentially fatal [4]. Despite remarkable advancements in cardiovascular care over recent decades, the prevention and management of arrhythmias and sudden cardiac death remain pressing public health challenges. Most sudden cardiac deaths and arrhythmias occur outside of hospitals and frequently present with no

☆ This article is part of a Special issue entitled: 'Secure, smart, and high-performance healthcare systems' published in Computer Methods and Programs in Biomedicine.

* Corresponding author.
E-mail address: xielp@bmie.neu.edu.cn (L. Xie).

[1] Contributed equally to this work.

https://doi.org/10.1016/j.cmpb.2025.109101
Received 22 May 2025; Received in revised form 10 September 2025; Accepted 4 October 2025
Available online 10 October 2025

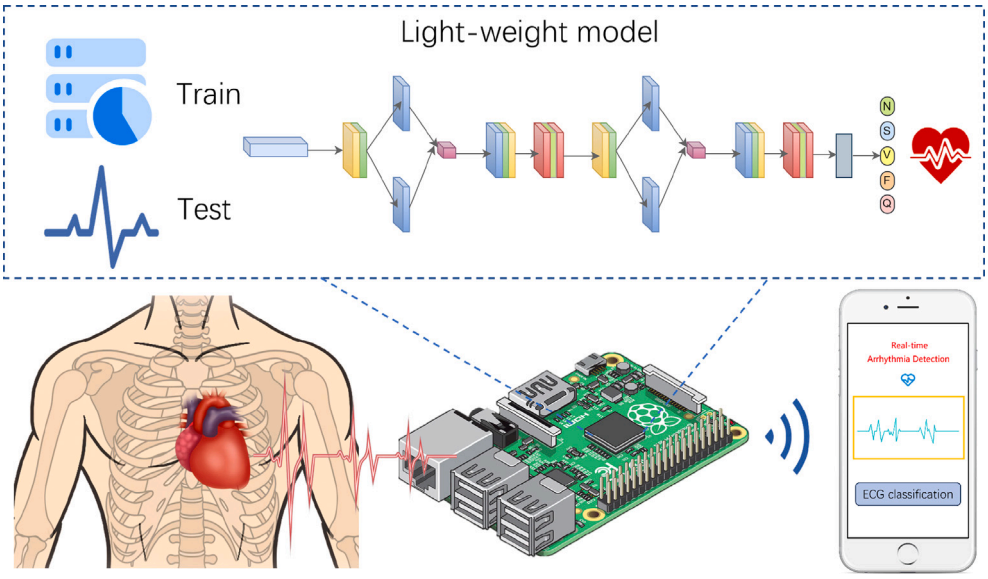

Fig. 1. Illustrator diagram of a light-weight model deployed on an embedded device for detection of arrhythmia.

noticeable symptoms [5]. Real-time detection of arrhythmias is essential for timely diagnosis, potentially preventing severe health problems, such as heart failure, myocardial infarction, and sudden cardiac death. Electrocardiogram (ECG)-based detection is widely recognized as one of the most basic and reliable methods for CVD diagnosis [6], which records the electrical activity of the heart [7]. ECG provides rich information about heart activity and can detect various electrical abnormalities [8]. Traditional long-term ECG detection is usually invasive and hinders patients' daily activities. Wearable devices are increasingly used in the field of medical health [9,10], which can perform long-term, continuous, and real-time heart rhythm detection in daily life [11]. Nevertheless, with the consistent accumulation of data, it has become a challenge to effectively analyze and process these data [12]. Traditional manual analysis methods are time-consuming, labor-intensive, and susceptible to subjective factors, leading to questionable reliability and consistency of analysis results [13].

Machine learning (ML) provides an efficient tool for identifying pathologies and assisting in medical diagnosis [14–17]. It has been successfully applied to ECG-based diagnosis and performed well in arrhythmia detection [18]. However, ML relies on human-defined ECG features, requiring extensive feature engineering and often lacking in automated analysis capabilities. Deep Learning (DL), a subtype of ML can learn patterns from ECG to efficiently and automatically diagnose arrhythmias [19–23]. Although DL has demonstrated its powerful feature learning capabilities [24], it is still a challenge to deploy DL on wearable devices. Due to constraints related to memory, computational efficiency, and energy consumption, most DL models are not directly deployable on target devices [25], and they often require high computational power and large memory [26].

To resolve the conflict between the escalating computational demands of AI-driven applications and the stringent resource constraints of edge devices, researchers have prioritized the development of lightweight models [26]. A dominant strategy in this field focuses on architectural simplification by reducing neural network depth and integrating parameter-efficient operations. Representative models include LiteNet [27], AFSense-ECG [28], and AFibri-Net [29]. He et al. [27] proposed a lightweight network model called LiteNet based on a Lite module, which incorporates squeezed convolutional layers with depth-separable convolutional layers to reduce the model computation. The LiteNet was deployed on a normal personal computer with Intel (R) CPU i3-2370M at 2.40 GHz and 4 GB memory, achieving a single inference time of approximately 20 ms. However, the LiteNet struggled to effectively capture the high-dimensional nonlinear features of ECG signals, leading to limited accuracy. Ukil et al. [28] proposed an AFSense-ECG model based on CNN, which utilizes a CNN structure for automatic AF detection on embedded devices. The model achieved an F1-score of 86.13% on the PhysioNet CinC2017 dataset. However, the model has a large parameter count of approximately 433,675. To deploy it on embedded devices, the model must first be compressed and then converted using TensorFlow Lite Converter. Phukan et al. [29] proposed an AFibri-Net model based on a lightweight convolutional neural network. They compared 54 different combinations of model parameters to achieve the best balance between classification accuracy and model size. For 5s ECG segments, the model achieved the highest accuracy of 93.88% and 95.80% on the Long-Term Atrial Fibrillation database and PhysioNet CinC2017 database, respectively. The model was deployed on the Raspberry Pi 4 computing platform with Broadcom BCM2711, a 1.5 GHz Cortex-A72 quad-core CPU with 8 GB RAM, and achieved a maximum AF detection time of a mere 3 ms. Gu et al. [30] proposed a lightweight CNN model coupled with an efficient hardware architecture. The hardware circuit used a 16-bit floating-point data format to accelerate the CNN computation. The model achieved an accuracy of 97.69% on the MIT-BIH database and a classification time of 0.3 ms per heartbeat on an embedded device for the classification of heart rhythm abnormalities. However, the inherent limitation of reducing network layers and simple model architecture lies in their constrained capacity to learn high-dimensional features, inevitably leading to compromised accuracy and diminished robustness. Meng et al. [31] proposed a Lightweight Fussing Transformer for dynamic ECG heartbeat classification. The model uses CNN and channel attention-based structure for feature extraction. By replacing the self-attention mechanism in the traditional Fussing Transformer with LightConv Attention (LCA), the model performed well on the CPSC2020 dataset with an average accuracy of 99.32%. Although the computational cost of this model is greatly reduced compared to the traditional Transformer model, it still requires 2.69 million parameters and consumes huge computational resources. The central challenge in lightweight model research lies in achieving an optimal trade-off between computational efficiency and task performance.

In this paper, we propose an elementwise-product enhanced lightweight model (EPLM) for real-time arrhythmia detection on resource-constrained devices (Fig. 1). The architecture integrates a spindle-shaped backbone constructed with depthwise separable convolutions to minimize computational overhead while retaining discriminative feature extraction capabilities, and a dual-branch element-wise

product mechanism that synergistically captures high-dimensional feature interactions. By leveraging the two cross-branch fusion, the model efficiently resolves the complexity-performance trade-off, enabling robust feature representation within shallow network layers.

Evaluated on the MIT-BIH arrhythmia database, the EPLM achieves state-of-the-art performance with 99.10% average accuracy while utilizing merely 25,981 parameters. Practical deployments on a Raspberry Pi 5 and an Android ×86 virtual machine platform results in low single-inference latencies of 2.29 ms and 0.62 ms, respectively, confirming the model's suitability for real-time arrhythmia monitoring. This work bridges the gap between high-accuracy arrhythmia diagnosis and edge-device deployment, offering a clinically viable solution for timely cardiovascular intervention. The contributions of this article are as follows:

1. An ultra-lightweight arrhythmia classifier for edge deployment is proposed. The EPLM achieves 99.10% average accuracy on the MIT-BIH database with only 25,981 parameters, which demonstrates unprecedented efficiency-accuracy trade-off optimization for ECG analysis on resource-constrained devices.

2. The model employs a spindle-structured backbone integrating depthwise separable convolutions for feature extraction with minimal FLOPs and a cross-branch element-wise product layer that non-linearly fuses multi-scale features. The hybrid design reduces computational costs while enhancing discriminative power through high-dimensional feature interactions.

3. Deployed on a Raspberry Pi 5 platform and an Android ×86 virtual machine, the EPLM model achieves inference latencies of 2.29 ms and 0.62 ms, respectively. These results demonstrate its suitability for continuous real-time monitoring of arrhythmia.

2. Method

2.1. Database

The data utilized in this study were sourced from four publicly available ECG databases, including Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) Arrhythmia Database [32] and the Physikalisch-Technische Bundesanstalt (PTB) Diagnostic ECG database [33], St. Petersburg INCART 12-lead Arrhythmia database [34] (INCART), and MIT-BIH Supraventricular Arrhythmia database [35] (SVDB). The four databases collectively underpin the development and rigorous validation of the model, thereby guaranteeing a high degree of algorithmic accuracy and extensive applicability. Limb Lead II was chosen for arrhythmia classification in this study, as it is commonly used to simulate single-lead signals in embedded devices for model training [36]. The MIT-BIH Arrhythmia Database contains 48 half-hour two-channel recordings, sampled at 360 Hz and band-pass filtered between 0.1–100 Hz. These recordings were obtained from 47 subjects (25 men, 22 women) aged 23 to 89 and include both normal rhythms and 16 types of arrhythmias. Beat annotations comply with the AAMI EC57 standard, categorizing each beat as normal (N), supraventricular ectopic (S), ventricular ectopic (V), fusion (F), or unclassifiable (Q). The database contains a total of 109,305 annotated beats, with the following distribution: 90,472 N, 2,775 S, 7,229 V, 802 F, and 8,027 Q. The PTB Database contains 549 records from 290 subjects (209 men, 81 women) aged 17 to 87. Each record includes standard 12-lead ECGs and 3 Frank leads. This database contains various cardiovascular diseases, including arrhythmia, myocardial infarction, and others. Among these, the two main categories of Normal(N) and myocardial infarction(M) are frequently used for binary classification tasks [21,37]. We also adopted this selection method, and in the end, there were 10505 and 4045 samples in the N and M classes respectively. The INCART database comprises 75 half-hour 12-lead ECG recordings sampled at 257 Hz, derived from 32 subjects (17 men and 15 women) aged 18 to 80. Although it includes all five AAMI beat categories, such as normal (N), supraventricular ectopic (S), ventricular ectopic (V), fusion (F), and unclassifiable (Q), the F and Q classes account for less than 1% of all heartbeats. To maintain class balance and consistency with other databases, only the three major categories were retained for analysis, resulting in 153,483 N, 1,958 S, and 19,994 V beats. The SVDB comprises 78 half-hour, dual-channel ECG recordings. Following the AAMI standard, beat annotations are categorized into three classes: normal (N), supraventricular ectopic (S), and ventricular ectopic (V). The dataset contains 162,129, 12,191, and 9,931 samples for each class, respectively.

2.2. Data preprocessing

The ECG signals are segmented into uniformly sized cardiac beat segments, which enhances their compatibility with deep neural network architectures while preserving the integrity of their intrinsic characteristics. ECG signals from the MIT-BIH database are segmented into uniform, fixed-size segments based on the R-peaks provided by the database. Each heartbeat segmentation window includes 300 sampling points, which corresponds to a period of 0.8 s at a sampling frequency of 360 Hz [38]. ECG signals from the INCART and SVDB databases were first resampled to 128 Hz, and then segmented into 300-sample segments using the annotated R-peaks provided by each database. The PTB database was downsampled to a sampling frequency of 125 Hz. The ECG signals were then segmented into 10-seconds and normalized. As the PTB database lacks R-peak annotations, we detected them using a first-derivative zero-crossing analysis. This computationally efficient method was selected to support our objective of developing a lightweight model for real-time processing on embedded devices. Using the R-peak as a reference, the ECG signals were further divided into 1.5-second segments, each comprising 188 sampling points.

2.3. Embedded device

To explore the feasibility of deploying ECG classification models on wearable devices, Raspberry Pi 5 and Android 10 ×86 virtual machine are chosen as simulated embedded computing platforms. The Raspberry Pi 5 is equipped with a Broadcom BCM2712 processor that features a 2.4 GHz quad-core 64-bit Arm Cortex-A76 CPU and 8 GB RAM. The operating system for the Raspberry Pi is Raspberry Pi OS. It is a Debian-based Linux operating system specifically optimized for Raspberry Pi hardware, offering excellent performance and compatibility. Python version 3.11.2 is used for programming and Pytorch version 2.2.2 is selected as the deep learning framework for implementing and running the models. The Android 10 ×86 virtual machine is configured with a 4-core ×86 CPU, 2 GB of RAM, and 6 GB of storage for the data partition. The environment is compatible with TensorFlow Lite 2.3.0 and delivers sufficient performance to efficiently run and test lightweight DL models.

2.4. Experimental environment

The EPLM model is trained on a cloud server equipped with a 16-core Intel® Xeon® Gold 6430 processor, a 24 GB NVIDIA TX4090 GPU, and 120 GB RAM. Python 3.10 and the Pytorch 1.2.1-GPU package are used to develop the model. The trained model is saved and deployed on the Raspberry Pi 5 and Android 10 ×86 virtual machine for practical testing and validation.

2.5. Evaluation metrics

Four key metrics were utilized to evaluate the classification performance of the model, including precision, recall, accuracy, and F1-score. Additionally, we employed a confusion matrix to compare the true categories with the predicted categories and computed the AUC-ROC score to evaluate the model's ability to distinguish between the positive

and negative classes across different classification thresholds. Accuracy indicates the overall rate of correct classifications in the entire database. Precision shows the accuracy of the positive predictions made by the model. Recall reflects the ability of the model to correctly identify actual positive instances. F1-score is the reconciled average of precision and recall. The specific calculations of the four metrics are shown below, where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives for each class.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{4}$$

To assess the lightweight nature of the proposed EPLM, we conducted a thorough analysis using three lightweight indicators, including the number of parameters, FLOPS (Floating-Point Operations Per Second), and single inference time and power consumption. The number of parameters corresponds to all trainable weights adjusted by the model through backpropagation during training. It serves as a direct measure of the model's size, with fewer parameters indicating a more lightweight model. FLOPS denotes the number of floating point operations in one second, which is a crucial metric for measuring the computational complexity of the algorithms. A lower FLOPS value indicates a less computationally intensive and potentially more energy-efficient model. Single inference time is the time required for the model to complete the inference, which is a vital metric for assessing the practicality of the model for deployment on hardware, especially where real-time processing is necessary. Power consumption quantifies the total electrical energy used by a model during inference, making it a crucial metric for assessing efficiency in battery-constrained or embedded devices. These metrics provide a comprehensive assessment of the model's performance in terms of the degree of lightweight nature.

2.6. Model

2.6.1. Motivation

Deep learning algorithms automatically learn features through successive layers, with early layers capturing simple patterns and deeper layers recognizing complex representations. Wider networks and deeper architectures enhance a model's ability to represent intricate functions, enabling breakthroughs in tasks. However, the increased width and depth of the models result in higher computational demands. Furthermore, common feature fusion strategies, such as concatenation and attention mechanisms, further exacerbate these constraints. Concatenation merely aggregates features without introducing cross-feature nonlinearities, and attention mechanisms often entail significant parameter and computational costs. As a result, such methods potentially restricting their implementation on devices with limited resources. In recent years, the element-wise product has emerged as a well-established technique for fusing cross-subspace features in diverse architectures. Notable examples include FocalNet [39], Visual Attention Network (VAN) [40], and EfficientMod [41] in computer vision, as well as Mamba [42] and Gated Linear Unit-enhanced Transformers (GLU) [43] in natural language processing. Recent theoretical insights suggest that the element-wise product inherently enables nonlinear interactions between features, effectively approximating a polynomial kernel mechanism in the context of image classification tasks [44]. This property enables implicit mapping of inputs into higher-dimensional representation spaces, thereby enriching feature representation while maintaining computational efficiency. However, these insights have primarily been validated in two-dimensional and spatially structured images data.

In this study, we propose an elementwise-product enhanced lightweight model which integrates a spindle-shaped backbone constructed with depthwise separable convolutions and a dual-branch element-wise product mechanism for analysis of 1D ECG signals. The element-wise product provides an efficient means to implicitly construct high-dimensional features through multiplicative interactions with low computational overhead. By leveraging the cross-branch fusions, the model efficiently resolves the complexity-performance trade-off, enabling robust feature representation within shallow network layers. To investigate the role of the element-wise product in processing 1D time-series ECG signals, we analyze its mathematical expansion. Let a 1D ECG signal be represented as $x \in R^{d \times n}$, where d is the number of input channels and n is the feature dimension per channel. For a dual branch with weights $w_1, w_2 \in R^{d \times d'}$, biases $b_1, b_2 \in R^{d' \times n}$, and the activation function $\sigma(\bullet)$, where d represents the number of input channels, d' represents the number of output channels, n represents the number of features in each channel, the element-wise product operation H is defined as:

$$H = \sigma(w_1^T x + b_1) \odot \sigma(w_2^T x + b_2) \tag{5}$$

Where $\odot$ denotes the Hadamard product. Expanding the (i, j)-th element H^{ij}:

$$H^{ij} = \sigma\left(\sum_{k=1}^{d} w_1^{ik} x^{kj} + b_1^{ij}\right) \cdot \sigma\left(\sum_{p=1}^{d} w_2^{ip} x^{pj} + b_2^{ij}\right) \tag{6}$$

Through polynomial expansion, this yields:

$$\begin{aligned} H^{ij} = & \underbrace{\sum_{k=1}^{d}\sum_{p=1}^{d} w_1^{ik} w_2^{ip} \sigma(x^{kj})\sigma(x^{pj})}_{\text{Quadratic interaction terms}} \\ & + \underbrace{b_2^{ij} \sum_{k=1}^{d} w_1^{ik} \sigma(x^{kj}) + b_1^{ij} \sum_{p=1}^{d} w_2^{ip} \sigma(x^{pj})}_{\text{Bias-Weighted Nonlinear Terms}} \\ & + \underbrace{b_1^{ij} b_2^{ij}}_{\text{Bias interaction}} \end{aligned} \tag{7}$$

The analysis reveals that the pairwise terms $\sigma(x^{kj})\sigma(x^{pj})$, the bias-weighted terms $\sigma(x^{kj})$ and $\sigma(x^{pj})$ exhibits nonlinear dependence on the input x, effectively encoding implicit high-dimensional feature representations. Theoretically, these cross-terms span unique combinatorial configurations, suggesting that the element-wise product operation can implicitly construct features in a near $\mathcal{O}(d^2)$ dimensional spaces while operating within d dimensional computational framework. Crucially, activation-induced sparsity (e.g., ReLU's zero-masking property) dynamically prunes statistically insignificant combinations through nonlinear gating. This adaptive sparsification mechanism achieves an optimal equilibrium between representational capacity and computational tractability. The theoretical analysis indicates that the element-wise product is able to map high-dimensional features in a low-dimensional space with a finite network width. Meanwhile, the element-wise product only involves the multiplication operation of the corresponding elements, and does not require additional parameters or complex mathematical transformations. Through element-wise product, the model can increase the depth of the network without significantly increasing the network width, thus improving model performance with limited parameters.

2.6.2. Detailed model architecture

We propose a novel lightweight network architecture for real-time arrhythmia monitoring on resource-constrained devices, as illustrated in Fig. 2. The proposed lightweight architecture addresses the

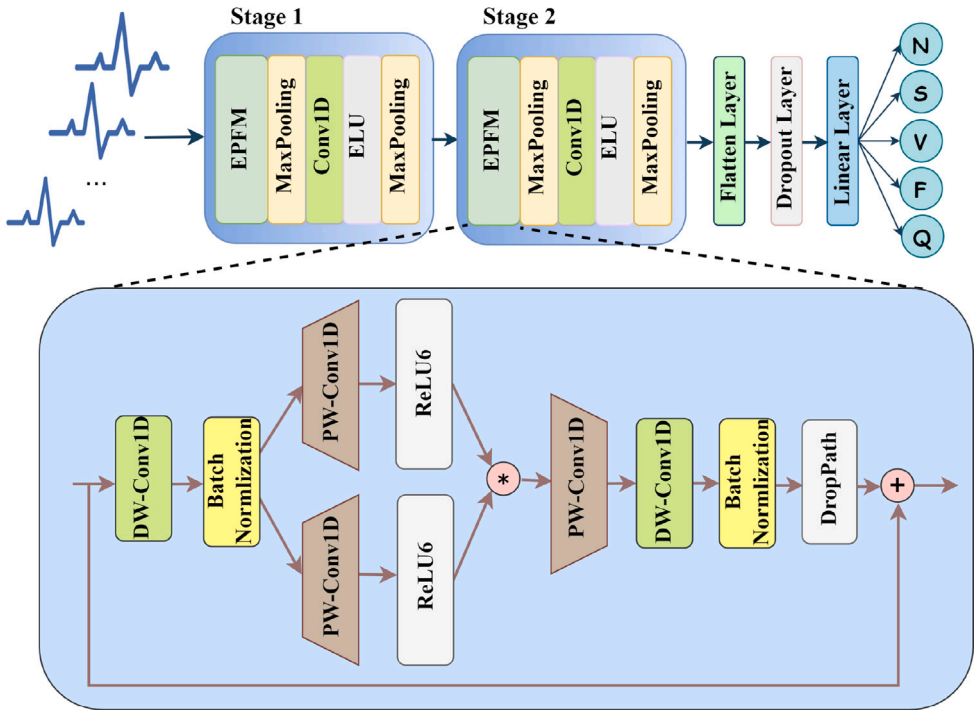

Fig. 2. Structure of EPLM.

complexity-performance trade-off through two synergistic innovations. An elementwise product fusion module (EPFM) employed a special spindle structure composed of depthwise separable convolution and a dual-branch-based element-wise product to extract hidden high-dimensional features. The depthwise separable convolution enhances computational efficiency by decomposing the traditional convolution process into two sequential stages: depthwise convolution (DW) and pointwise convolution (PW). In the depthwise phase, a unique filter is applied to each input channel, generating an output feature map that retains the same channel count as the input. Subsequently, the PW employs a 1×1 convolution kernel to fuse cross-channel information, thereby enriching the feature representation without modifying the spatial dimensions. This approach significantly reduces the computational cost by eliminating the necessity for convolution operations across all channels at each spatial position. The dual-branch element-wise product mechanism synergistically captures high-dimensional feature interactions. By leveraging two cross-branch fusion, the model efficiently resolves the complexity-performance trade-off, enabling robust feature representation within shallow network layers.

In detail, the innovative structure begins with a DW layer to process each input channel individually. This operation captures spatial features within each channel while maintaining a low parameter count. Subsequently, a dual-branch PW is employed, expanding the network width with a channel expansion ratio of 4 to project features into a higher-dimensional space. The outputs of both branches are activated via ReLU and fused through element-wise product, which nonlinearly amplifies features exhibiting significant activations in both branches. The element-wise product potentially amplifies features that are active in both branches, highlighting those with significant responses across both branches. Meanwhile, the element-wise product fusion reduces redundancy in feature representations, resulting in a more compact and efficient feature representation. To restore parameter efficiency, a second PW layer contracts the channel dimension back to its original size. A secondary DW convolution then refines the fused features spatially, enabling the learning of localized anomalies without deepening the network. This dual-branch design mitigates overfitting risks associated with deep architectures while achieving high-dimensional feature extraction under limited computational resources. Following the EPFM, max pooling reduces feature map dimensions by retaining salient activations. To counteract information loss during downsampling, a convolutional layer with Exponential Linear Unit (ELU) activation is introduced, leveraging its negative slope preservation to maintain gradient stability. Finally, the probabilities of different categories are output through the flatten layer and linear layer.

Lightweight models often rely on oversimplified convolutional architectures that sacrifice representational capacity for parameter efficiency, or utilize compressed attention mechanisms that still impose substantial computational demands. In contrast, our model achieves an optimal balance between performance and efficiency. The spindle-shaped depthwise separable convolution backbone enables parameter-efficient spatial feature extraction, while a dual-branch element-wise product mechanism enhances nonlinear representations without introducing significant complexity. This integrated design yields performance competitive with that of attention-based models, while maintaining high computational efficiency, which renders it particularly suitable for resource-constrained applications, such as wearable and embedded ECG monitoring systems.

3. Results

In this section, we conducted validation studies addressing the following three questions. How does element-wise product fusion affect accuracy and computational efficiency in our lightweight ECG model? What is the optimal lightweight model architecture for edge deployment, considering the trade-off between model size and arrhythmia detection performance? Does the proposed lightweight model learn clinically interpretable ECG features?

3.1. Model evaluation

We used the Adam optimizer in the model training phase and set the batch size to 128, the learning rate to 0.001, and the epoch to 50. This configuration can speed up the convergence speed, improve the computational efficiency, and stabilize the training process. At

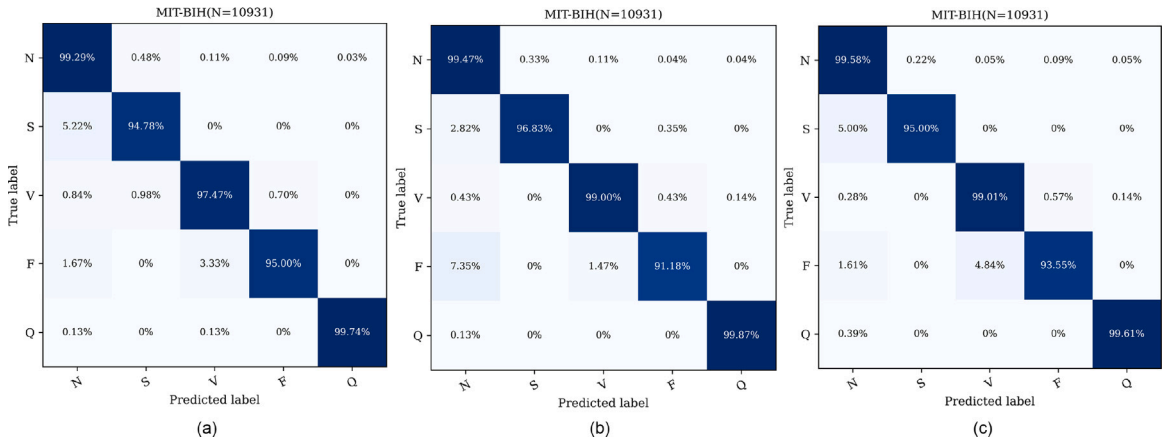

Fig. 3. Confusion matrices of the EPLM with different network widths. (a) is for EPLM (Param 1), (b) is for EPLM (Param 2), and (c) is for EPLM (Param 3).

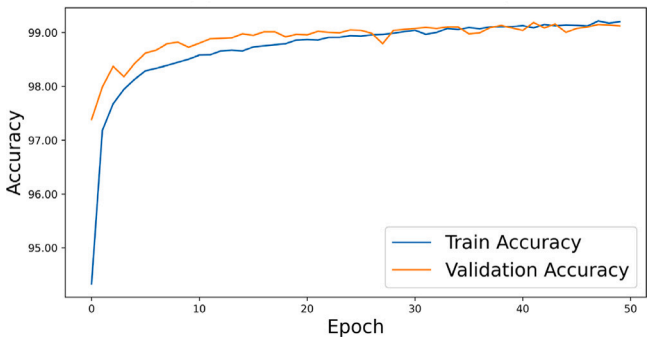

Fig. 4. Training accuracy curve of EPLM.

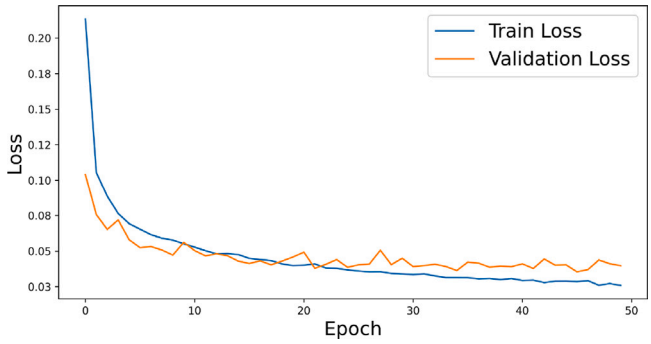

Fig. 5. Training loss curve of EPLM.

Table 1
Network widths of the modules in EPLM.

Structure	EPFM (Stage 1)	Conv1D (Stage 1)	EPFM (Stage 2)	Conv1D (Stage 2)
Parameters 1	4	8	16	32
Parameters 2	8	16	32	64
Parameters 3	16	32	64	128

the same time, a callback function was used to obtain the model parameters corresponding to the optimal training cycle, and early stopping was applied with a patience value of 10 to prevent overfitting. Most importantly, to rigorously evaluate the robustness of the model's performance, we employed ten-fold cross-validation and computed the mean and standard deviation of each performance metric across all folds. To achieve a balance between network width and computational complexity, we designed three model structures with different widths for training and testing. The specific parameters of the three structures are shown in Table 1, while the detailed network configuration of EPLM (Param2) is provided in Table 2. The classification performance of each model on the MIT-BIH database is presented in Table 3. EPLM (Param3) achieved the best overall results, attaining the highest average accuracy (99.19%), recall (92.91%), F1-score (94.51%), and AUC-ROC (99.74%). EPLM(Param2) achieved the highest average precision of 96.48%. The results indicate that while increasing network width generally enhances classification performance, the marginal improvement diminishes as the model grows wider, suggesting a trend toward performance saturation.

The normalized confusion matrices, which provide detailed class-wise predictive performance, are presented in Fig. 3. These matrices offer comparative insights into the performance of the three model structures: EPLM(Param1), EPLM(Param2), and EPLM(Param3). The elements on the diagonal indicate the percent of samples correctly classified by the model. EPLM(Param3) has a high classification accuracy in the class N. It was able to identify the healthy condition of the patient effectively, but there was a lack of identification of the specific type of arrhythmia during the patient's arrhythmic episode. EPLM(Param1) has high classification accuracy in class F but performs poorly in the remaining classes. The accuracy of the EPLM(Param2), is slightly less than that of the EPLM(Param3) in classifying class N, but it has better classification accuracy in each specific class of arrhythmia, especially in class S, V, and Q. Class S and V heartbeats provide an early warning of serious arrhythmias, and their presence is indicative of underlying cardiovascular diseases such as heart failure or ventricular tachycardia. Some patients experience dizziness, palpitations, and shortness of breath, which can interfere with daily life. The EPLM(Param2), which accurately identifies these types of abnormal heartbeats, makes it a valuable tool for clinical applications. By providing early warnings of potential health risks, EPLM(Param2) can aid in timely interventions and contribute to better patient outcomes.

The accuracy curves and loss curves of EPLM on the training and validation datasets are shown in Figs. 4 and 5, respectively. In the initial pre-training phase, the model exhibits a rapid increase in accuracy, indicating efficient learning of relevant features. Subsequently, after approximately 30 epochs, both the accuracy and loss curves plateau, showing better convergence. This congruity is indicative of a great fitting effect, where the model is not only adept at learning from the training data but also effectively applies the learned knowledge to the validation set, thereby mitigating the risk of overfitting.

Moreover, Gradient-weighted Class Activation Mapping (Grad-CAM) is employed to enhance the model's interpretability and visualize the learned features. As illustrated in Fig. 6, the resulting heatmap shows a strong correspondence with broad and aberrant QRS complexes and

Table 2
Network configuration of EPLM (Param2).

Layers	Type	Input_channel	Output_channel	Kernel (Groups)	Strides
Layer1	DW-Conv1D	1	8	5 (1)	1
Layer2	BN	\	\	\	\
Layer3–1	PW-Conv1D	8	32	1	1
Layer3–2	PW-Conv1D	8	32	1	1
Layer4	PW-Conv1D	32	8	1	1
Layer5	DW-Conv1D	8	8	5 (4)	1
Layer6	BN	\	\	\	\
Layer7	MaxPool	\	\	\	\
Layer8	Conv1D	8	16	5	1
Layer9	MaxPool	\	\	\	\
Layer10	DW-Conv1D	16	32	5 (8)	1
Layer11	BN	\	\	\	\
Layer12–1	PW-Conv1D	32	128	5	1
Layer12–2	PW-Conv1D	32	128	5	1
Layer13	PW-Conv1D	128	32	5	1
Layer14	DW-Conv1D	32	32	5 (16)	1
Layer15	BN	\	\	\	\
Layer16	MaxPool	\	\	\	\
Layer17	Conv1D	32	64	5	1
Layer18	MaxPool	\	\	\	\
Layer19	Flatten	\	\	\	\
Layer20	Linear	\	\	\	\

Table 3
Classification performance of the EPLM with different network widths.

Model	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)	AUC-ROC(%)
EPLM(Param1)	98.86 ± 0.12	96.22 ± 0.98	90.62 ± 1.55	93.13 ± 0.78	99.48 ± 0.11
EPLM(Param2)	99.10 ± 0.11	**96.48 ± 0.84**	92.39 ± 1.45	94.26 ± 0.68	99.67 ± 0.14
EPLM(Param3)	**99.19 ± 0.10**	96.42 ± 0.85	**92.91 ± 1.23**	**94.51 ± 0.65**	**99.74 ± 0.20**

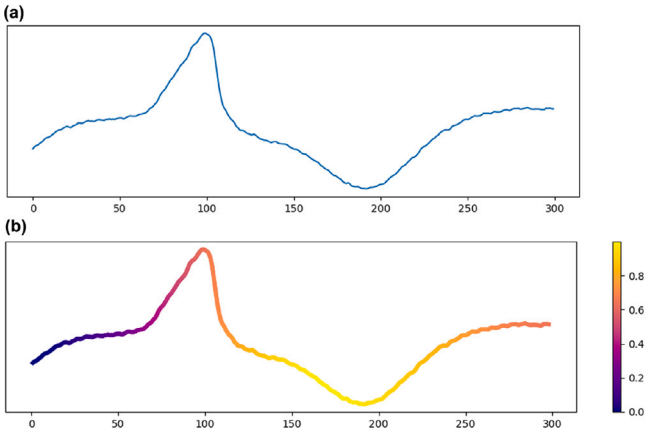

Fig. 6. Visualization of model attention. (a) Ventricular premature beats signal (Class V), (b) Attention visualization of the EPLM on the ventricular premature beats signal using Grad-CAM.

T-wave inversions characteristic of ventricular premature beats (Class V) [45]. Regions of higher intensity in the heatmap indicate that the model assigns greater attention to these specific waveform segments, underscoring their significance in the decision-making process. These findings are consistent with the clinical definition of ventricular premature beats, thereby validating both the interpretability of the model and the clinical relevance of the features it has learned.

3.2. Comparison with current lightweight models

To evaluate the effectiveness of our proposed lightweight model for arrhythmia detection, we conducted a comparative analysis with several state-of-the-art lightweight models specifically designed for ECG classification. Among these, Gu et al. introduced an ultra-lightweight CNN architecture optimized for hardware deployment [30], employing only two convolutional layers followed by a fully-connected layer to achieve efficient arrhythmia detection. Similarly, Phukan et al. developed AFibri-Net [29], a compact convolutional network tailored for atrial fibrillation detection, which systematically explores parameter configurations to balance classification performance and model complexity. Ukil et al. proposed AFSense-ECG [28], a shallow yet effective model that leverages large convolutional kernels to capture long-range temporal patterns in ECG signals, making it suitable for resource-constrained embedded systems. Another hybrid approach by Alamatsaz et al. integrated LSTM modules into a lightweight CNN framework to enhance contextual feature extraction and improve arrhythmia detection capabilities [46]. In addition to domain-specific models, we also compared our approach with widely recognized lightweight neural networks from other domains, namely MobileNet [47] and MicroNet [48], both adapted for ECG classification tasks. These benchmarks provide a broader perspective on the trade-offs between efficiency and accuracy in deep learning models deployed on edge devices.

The performance comparison between the proposed models and current lightweight models is presented in Tables 4 and 5 for the MIT-BIH and PTB databases, respectively. Overall, the three proposed EPLM architectures consistently demonstrate competitive performance across both databases. On the MIT-BIH database, EPLM (Param3) achieves the best classification performance among all models, with an accuracy of 99.19% and an F1-score of 94.51%. Notably, it outperforms the much larger MobileNet architecture while using only approximately 3% of the FLOPS and parameters. EPLM (Param1), with only 6,881 parameters and 149,400 FLOPS, maintains strong performance, achieving an accuracy of 98.86% and an F1-score of 93.13%. EPLM (Param2) strikes a balance between efficiency and performance, attaining 99.10% accuracy and 94.26% F1-score with 25,981 parameters and 525,400 FLOPS, thereby surpassing most competing lightweight models. A similar trend is observed on the PTB database, where EPLM (Param3) again delivers the best classification performance, with an accuracy of 99.13% and an F1-score of 98.92%, significantly exceeding other lightweight models. EPLM (Param2) remains highly competitive, achieving 98.85% accuracy and 98.56% F1-score at a lower computational cost, while

Table 4
Performance comparison with current lightweight models n the MIT-BIH database.

Models	Accuracy(%)	F1-score(%)	Precision(%)	Recall(%)	FLOPS	Params
MobileNet	99.15 ± 0.08	94.50 ± 0.56	95.89 ± 0.56	93.26 ± 0.75	50508928	3181573
MicroNet	99.09 ± 0.09	95.84 ± 1.19	92.64 ± 0.88	94.12 ± 0.49	6472674	1241787
AFibri-net	99.03 ± 0.16	93.95 ± 0.76	96.40 ± 1.02	91.88 ± 1.16	177048	37749
CNN-LSTM	98.73 ± 0.14	92.57 ± 0.74	95.30 ± 0.75	90.26 ± 0.90	3110352	35061
AFSenseECG	99.07 ± 0.08	94.13 ± 0.65	96.19 ± 0.93	92.46 ± 1.48	46346944	435461
1DCNN	97.93 ± 0.21	89.31 ± 0.89	93.21 ± 1.27	86.34 ± 1.78	77100	1933
EPLM(Param1)	98.86 ± 0.12	93.13 ± 0.78	96.22 ± 0.98	90.62 ± 1.55	149440	6881
EPLM(Param2)	99.10 ± 0.11	94.26 ± 0.68	96.48 ± 0.84	92.39 ± 1.45	525400	25981
EPLM(Param3)	99.19 ± 0.10	94.51 ± 0.65	96.42 ± 0.85	92.91 ± 1.23	1957120	100853

Table 5
Performance comparison with current lightweight models on the PTB database.

Models	Accuracy(%)	F1-score(%)	Precision(%)	Recall(%)	FLOPS	Params
MobileNet	98.78 ± 0.39	98.48 ± 0.49	98.59 ± 0.52	98.38 ± 0.55	31210048	3178498
MicroNet	98.29 ± 0.40	97.86 ± 0.50	97.95 ± 0.58	97.80 ± 0.64	4289436	1241394
AFibri-net	98.10 ± 0.41	97.63 ± 0.52	97.58 ± 0.53	97.69 ± 0.64	96744	27482
CNN-LSTM	98.54 ± 0.57	98.19 ± 0.71	98.19 ± 0.74	98.19 ± 0.74	3220032	32034
AFSenseECG	98.82 ± 0.54	98.54 ± 0.68	98.52 ± 0.66	98.56 ± 0.76	24590848	431170
1DCNN	91.64 ± 0.87	89.65 ± 1.06	89.40 ± 1.09	89.93 ± 1.09	47324	798
EPLM(Param1)	97.58 ± 1.24	96.98 ± 1.55	97.09 ± 1.65	96.90 ± 1.52	92432	6782
EPLM(Param2)	98.85 ± 0.16	98.56 ± 0.20	98.63 ± 0.32	98.49 ± 0.24	326176	25786
EPLM(Param3)	99.13 ± 0.18	98.92 ± 0.23	98.98 ± 0.29	98.86 ± 0.28	1217600	100466

Table 6
Performance comparison with current lightweight models on embedded devices.

Models	Raspberry Pi 5		Android 10 ×86 virtual machine	
	Power cost (J)	Time (ms)	Power cost (J)	Time (ms)
MobileNet	399.39	18.27	1848.11	84.54
MicroNet	293.14	13.41	577.78	26.43
AFibri-Net	27.43	1.25	10.95	0.50
CNN-LSTM	41.49	1.90	121.95	5.58
AFSenseECG	77.89	3.56	1516.48	69.37
1DCNN	8.20	0.37	3.61	0.17
EPLM(Param1)	46.59	2.13	6.52	0.30
EPLM(Param2)	49.67	2.29	13.48	0.62
EPLM(Param3)	58.88	2.69	85.60	3.92

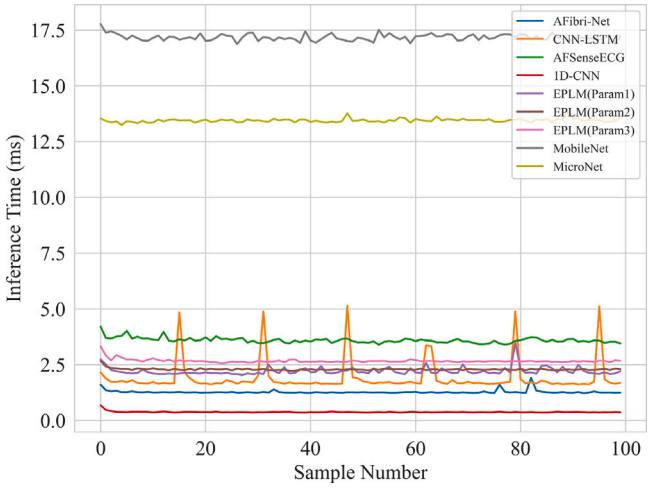

Fig. 7. Performance of different lightweight models for real-time detection.

EPLM (Param1) provides a lightweight yet effective alternative under extreme resource constraints. These results underscore the effectiveness of the element-wise product mechanism in the proposed models, which facilitates high-dimensional nonlinear feature extraction and enhances classification performance within a compact architecture.

3.3. Performance on embedded device

To further evaluate the model's performance under realistic application conditions, we deployed all candidate models on embedded platforms, including Raspberry Pi 5 and an Android 10 ×86 virtual machine. Specifically, the tenth fold of the MIT-BIH ten-fold database was used as the test set. We measured the total power consumption during inference over the entire test set and calculated the average inference time per heartbeat for each model deployed on both platforms. The detailed results are summarized in Table 6. Among the three proposed architectures, EPLM (Param1) achieves the best overall efficiency. On the Android platform, it consumed only 6.52 J of energy with an average inference time of 0.30 ms per heartbeat, performing comparably to the ultra-lightweight 1DCNN model. On Raspberry Pi 5, it also delivered low latency and energy usage. EPLM (Param2) attained short inference times and manageable power consumption on both platforms. Although slightly outperformed in efficiency by models, such as CNN-LSTM and AFibri-Net, it exhibited significantly superior real-time stability and classification accuracy. EPLM (Param3), while having the highest computational cost among the three variants, still maintained lower power cost, faster inference speed, and greater stability compared to models with similar accuracy, such as AFSenseECG, MobileNet, and MicroNet.

Real-time monitoring requires strong stability in addition to efficient performance of the models. We evaluated the real-time monitoring stability of the lightweight models for ECG signal classification on a Raspberry Pi 5. The inference time curves were obtained under realistic deployment conditions. The inference time was measured per input sample, with each sample consisting of 100 heartbeat signals, to ensure statistical reliability of the timing results. All competing models were tested using the same hardware and software environment to guarantee a fair comparison. As shown in Fig. 7, most models exhibited some level of performance fluctuation in terms of classification stability, except for EPLM(Param2) and 1-DCNN. Notably, CNN-LSTM shows the most significant fluctuations. MobileNet and MicroNet exhibit relatively high inference latency compared to other models evaluated in

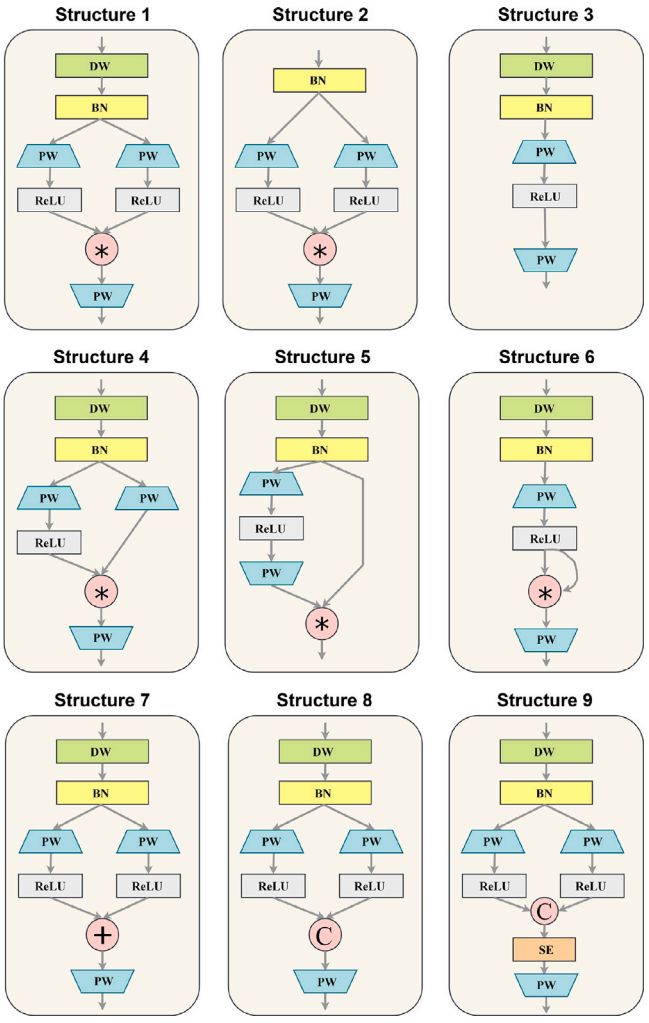

Fig. 8. Schematic diagram of modules in EPLM with different structures.

this study. The EPLM series of models achieves an excellent balance among accuracy, energy consumption, inference speed, and stability, making it particularly suitable for real-time ECG monitoring scenarios in resource-constrained embedded environments. Although EPLM (Param1) has the lightest structure, it struggled with classification performance stability, making it less suitable for real-time applications. EPLM(Param3) exhibits the best classification performance but has a heavier structure and slower classification time, thus unsuitable for deployment on embedded systems. The results suggest that EPLM(Param2) is an optimal choice for real-time ECG signal classification on wearable devices, which is well-suited for the demands of embedded systems, particularly in scenarios requiring efficient and reliable real-time detection.

3.4. Ablation experiment

To validate the effectiveness of the spindle structure in EPLM for extracting high-dimensional nonlinear features and improving classification performance, we compared it against eight alternative structural variants. The architectural details of each variant are illustrated in Fig. 8, while their corresponding classification results are summarized in Table 7. We first performed a series of component-level ablation studies. Structure 1 represents the proposed spindle configuration and achieved the highest classification accuracy of 99.03%. Structure 2 omits the depthwise convolution layer, and Structure 3 removes the

Table 7
Classification performance using different structures.

Modules	Accuracy(%)	F1-score(%)	Recall(%)	Precision(%)
Structure 1	**99.03**	**94.08**	**92.82**	95.46
Structure 2	98.72	92.53	88.86	97.18
Structure 3	98.67	92.63	89.56	96.48
Structure 4	98.76	92.71	91.62	94.19
Structure 5	98.71	92.22	89.35	95.74
Structure 6	98.90	92.71	88.97	**97.77**
Structure 7	97.74	85.49	82.53	94.89
Structure 8	98.84	92.12	88.84	96.41
Structure 9	98.89	92.20	89.16	96.37

Table 8
Classification performance of different activation function combinations.

Combinations	Accuracy(%)	F1-score(%)	Recall(%)	Precision(%)
ReLU6+ReLU6	98.81	91.84	87.50	97.93
ReLU6+ReLU	98.97	91.71	87.76	97.30
ReLU6+ELU	**99.10**	**94.26**	92.39	96.48
ReLU6+LeakyReLU	98.76	91.38	88.03	95.87
ReLU+ReLU6	98.54	90.58	88.03	93.75
ReLU+ReLU	98.69	91.19	87.93	95.52
ReLU+ELU	99.03	94.08	**92.82**	95.46
ReLU+LeakyReLU	98.75	91.88	88.58	96.06
ELU+ReLU6	98.63	90.44	86.25	96.36
ELU+ReLU	98.75	91.57	88.01	96.31
ELU+ELU	98.93	92.81	89.52	97.18
ELU+LeakyReLU	98.54	90.55	88.50	93.02
LeakyReLU+ReLU6	98.78	92.15	87.91	**97.99**
LeakyReLU+ReLU	98.93	93.58	90.56	97.28
LeakyReLU+ELU	98.87	91.72	88.59	96.14
LeakyReLU+LeakyReLU	98.72	90.80	86.41	97.36

element-wise product operation. The superior performance of Structure 1 over both demonstrates that each of these components contributes indispensably to the model's discriminative capability. We further ablated the specific design of the spindle structure. In Structure 4, only one branch is activated before element-wise multiplication. Structure 5 performs multiplication between the activated single-branch features and the original input. Structure 6 uses a self-multiplication of the activated single-branch output. The proposed dual-branch activation followed by element-wise product (Structure 1) outperformed all these variants, indicating that activating both branches before fusion more effectively captures discriminative nonlinear features. Additionally, we compared the element-wise product against other common fusion strategies, including addition (Structure 7), concatenation (Structure 8), and an attention-based fusion incorporating a squeeze-and-excitation (SE) block (Structure 9). The results demonstrate that element-wise product not only captures richer feature interactions than addition or concatenation but also extracts more discriminative nonlinear characteristics than the SE-enhanced attention mechanism. These experiments confirm that the spindle structure of the EPLM successfully enhances feature extraction and classification performance through its dual-branch element-wise product design. To visualize the feature distinguishability of each structure more straightforwardly, we used the t-distributed Stochastic Neighbor Embedding (t-SNE) plot to visualize the distribution of the features extracted by different structures (Fig. 9). The t-SNE plots provide a clear visual representation of how well different structures perform in classifying ECG features. Structure 1 stands out due to its excellent feature separation, showcasing the efficacy of its structure. While other structures exhibit relatively poor performance across multiple classes, including Class F, Class S, and Class Q.

Optimization experiments are performed to determine the optimal activation functions for the module EPFM and that of the Stage component. The results, as detailed in Table 8, indicate that the highest accuracy and F1-score are achieved when the ReLU6 activation function is used in the EPLM and the ELU activation function is used in the Stage component. The combination of ReLU in EPLM and ELU in Stage

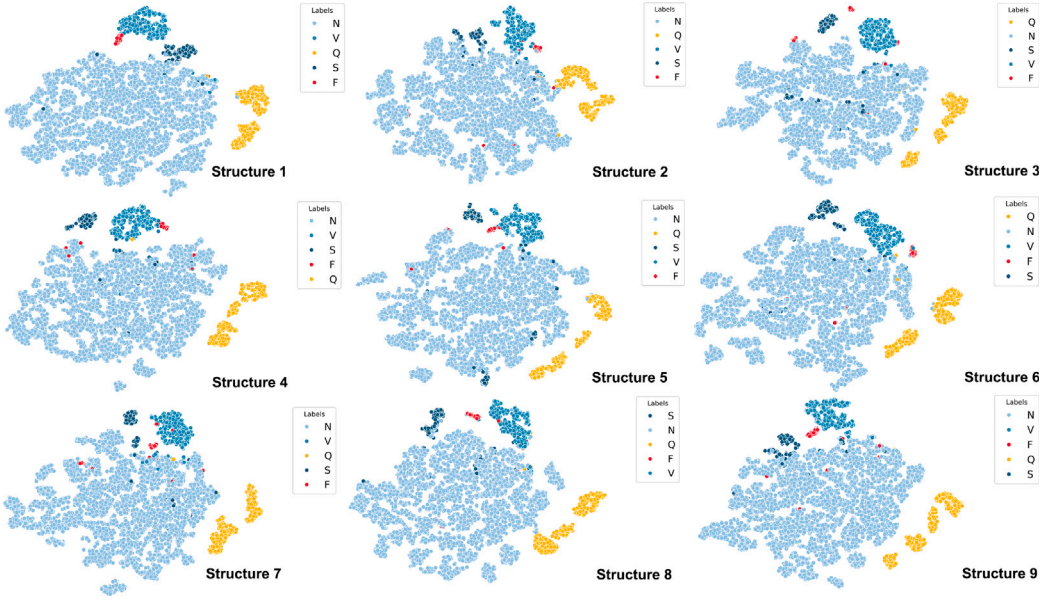

Fig. 9. t-SNE plots of models with different structures.

obtains the highest recall. Meanwhile, the highest precision is attained with ReLU6 in EPLM and ReLU6 in Stage. Based on these findings, the combination of ReLU6 in EPLM and ELU in the Stage component maximizes the accuracy of arrhythmia classification. The choice of ReLU6 is critically important to the mathematical properties derived in Eqs. (5)–(7) in Section 3.1, as it directly ensures the stability and efficiency of the implicit high-dimensional feature mapping. Unlike unbounded functions like ReLU, LeakyReLU, or ELU, the ReLU6 limits the output to the range [0, 6], which is essential to preventing numerical explosion in the expanded polynomial terms of Eq. (7). This constraint guarantees that the pairwise multiplicative terms and bias-weighted terms remain within a stable bounded interval, thereby mitigating gradient instability and training difficulties. Additionally, ReLU6 dynamically prunes insignificant feature combinations by strictly zeroing out negative inputs. This avoids the propagation of noisy or irrelevant combinatorial features, which often occur with activation functions like LeakyReLU or ELU that retain negative values. The use of ELU in the Stage component plays a complementary role. It restores potentially useful negative activations that may arise from element-wise products without compromising stability, further enhancing the model's representational capacity.

We also conducted ablation studies on the channel expansion ratio in the EPLM module. The results shown in Table 9 indicated that the model achieved the best classification performance when the expansion ratio was set to 4. Furthermore, we conducted a performance trade-off analysis between depthwise convolution and standard convolution layers. Specifically, we replaced all the depthwise convolution layers in the original model with standard convolutions and compared the classification performance and parameter size with the original model (Table 10). The results indicate that both convolution types achieve similar high accuracy and F1-score, with depthwise convolution having a slight edge in recall and a smaller parameter size. This suggests that depthwise convolution can maintain comparable performance while reducing model complexity.

3.5. Generalization of models

To verify the cross-dataset generalization performance of the proposed EPLM model, we additionally evaluate the model on the INCART12 and SVDB databases. First, we trained the model on the MIT-BIH database. To ensure label consistency across the MIT-BIH, SVDB, and INCART12 databases, only N, S, and V beats were retained for training. To validate the generalization capability of EPLM, the model is trained on MIT-BIH and then evaluated on the INCART12 and SVDB databases as external test sets. Tables 11 and 12 present the comparison between EPLM and existing lightweight models. The results demonstrate that our proposed model exhibits superior generalization across all evaluation datasets compared with existing lightweight models, underscoring its robustness.

Table 9
Classification performance of different expansion ratios.

Ratio	Accuracy(%)	F1-score(%)	Recall(%)	Precision(%)
2×	99.07	94.08	92.02	**96.55**
4×	**99.10**	**94.26**	**92.39**	96.48
8×	99.08	94.20	92.35	96.38

Table 10
Performance comparison of different convolution layer types.

Convolution types	Accuracy(%)	F1-score(%)	Precision(%)	Recall(%)	Params
Standard	**99.11**	94.18	**92.46**	96.21	26 933
Depthwise	99.10	**94.26**	92.39	**96.48**	**25981**

4. Discussion

4.1. Comparison with state-of-the-art works

To demonstrate the potential and utility of the model architecture we proposed that incorporates element-wise product, we compared it with current state-of-the-art models for ECG classification. Specific data are shown in Table 13. Xu et al. [49] used multimodality data matching-based data augmentation module and a multimodality feature encoding module for arrhythmia classification, achieving an accuracy of 98.83%. However, the model leads to increased computational complexity and high computational costs, making it less suitable for deployment on resource-constrained platforms. El-Ghaish et al. [21] proposed a transformer framework specifically for the classification of ECG rhythm abnormalities. The framework contains a Bidirectional

Table 11
Comparison of generalization performance with other models on the INCART database.

Models	Accuracy(%)	F1-score(%)	Precision(%)	Recall(%)
MobileNet	92.87	70.74	65.32	84.88
MicroNet	94.71	74.03	69.95	87.81
AFibriNet	92.73	70.49	69.50	89.10
CNN-LSTM	94.45	72.73	70.55	88.22
AFSenseECG	93.05	70.31	67.42	86.77
1DCNN	94.40	71.30	69.23	83.44
EPLM(Param2)	94.86	74.56	70.35	85.72

Table 12
Comparison of generalization performance with other models on the SVDB database.

Models	Accuracy(%)	F1-score(%)	Precision(%)	Recall(%)
MobileNet	86.06	55.47	56.20	65.17
MicroNet	86.47	56.88	59.52	65.29
AFibriNet	82.59	48.19	52.06	62.56
CNN-LSTM	86.70	56.51	55.31	64.14
AFSenseECG	85.79	55.77	54.75	64.07
1DCNN	86.10	50.27	58.22	60.85
EPLM(Param2)	88.09	59.80	60.58	65.77

Transformer that captures the time dependence and a Multiscale Convolution and Channel Recalibration Module that implements spatial feature extraction at various granularities. Islam et al. [38] proposed a hybrid model using CNN and Transformer. The model utilizes convolutional layers to capture local information and transformer encoders to extract contextual information from the ECG signals. The two studies use the Transformer to capture long-range dependencies of ECG signals and learn important patterns in the time series to improve classification accuracy. Though achieving excellent classification performance, the attention mechanism in Transformer requires huge computational resources and has some limitations in practical application and deployment. Compared with the latest studies mentioned above, our proposed model demonstrates a significant advantage in terms of classification accuracy and computational efficiency. The proposed EPLM can extract high-dimensional nonlinear features effectively, strengthen the important features, and suppress the irrelevant noise. Besides, the model is successfully deployed on an embedded device. It achieves an average accuracy of up to 99.19% with and a single inference time of only 2.69 ms.

To further test the generalization performance of the model, we then compare the performance of the proposed model with that of the existing work using the PTB-DB database. The overview of performance is shown in Table 14. Acharya et al. [55] proposed a CNN model and achieved a classification accuracy of 93.50%. Diker et al. [56] proposed a traditional Extreme Learning Machine (ELM)-based machine learning algorithm with hidden neuron optimization using differential evolutionary algorithms for ECG classification and obtained a classification accuracy of 97.50%. Alkhawaldeh et al. [58] proposed the DeepResidualBiLSTM model based on residual convolutional neural networks and bi-directional LSTMs, resulting in a classification accuracy of 98.70%. Our proposed EPLM (Param2) model achieved a more accurate classification accuracy of 99.13% than the above works with a lighter and simpler structure. Zhang et al. [57] proposed a 1D Integrated EfficientNet model based on one-dimensional convolutional neural network and inverted residual bottleneck block and accomplished a classification accuracy of 99.60%. The model consists of three sub-models that form an ensemble and use a voting scheme for prediction. The final prediction is decided by majority voting among the models. However, although the individual model is relatively lightweight, the parallel operations of the integrated learning consume a large amount of computational resources and are difficult to deploy practically on wearable devices. Anwar et al. [59] take a traditional machine learning approach and optimize its lightweight degree while earning a classification accuracy of 99.74%. Although it achieved excellent classification accuracy, it still requires tedious feature engineering for manual extraction. In contrast to the outstanding work described above, our proposed model achieves comparable accuracy to state-of-the-art models and outperforms most existing models. Unlike many models that often rely on complex architectures and extensive computational resources, our model maintains high performance while being more lightweight and efficient. These results demonstrate its ability to generalize well across different databases.

4.2. Limitations and future work

Although the EPLM model demonstrates real-time operation on a Raspberry Pi 5 and Android 10 ×86 virtual machine, challenges remain in deploying it on even more resource-constrained devices. To address this, model compression techniques such as pruning and quantization can be employed to further reduce model size and computational complexity, thereby lowering energy consumption and extending battery life for wearable continuous monitoring systems. However, these optimizations must be carefully implemented to maintain diagnostic accuracy when handling the intrinsic complexity of ECG signals, including morphological variations, noise artifacts, and individual-specific patterns, which are crucial for reliable arrhythmia detection. Moreover, while the model focuses on classifying common arrhythmias, it may struggle with rare or complex cases, especially given the diversity and subtlety of ECG manifestations across different patient demographics and clinical conditions. To mitigate this, future work will focus on improving generalization to unknown or rare arrhythmias, potentially through incremental learning techniques that allow the model to adapt to new patterns without full retraining. It should be noted that the use of such continuously evolving models, particularly when fine-tuned with cloud data, raises significant privacy concerns [60,61]. Mechanisms for federated learning and differential privacy must be integrated to ensure that patient data remains protected during model updates, especially when deployed across multiple institutions.

5. Conclusion

This work proposes a highly lightweight model, EPLM, designed specifically for the detection of arrhythmia for wearable devices. The model incorporates a novel spindle architecture composed of depthwise

Table 13
Comparison with state-of-the-art works on the MIT-BIH database.

Authors	Year	Method (s)	Accuracy(%)	F1-score(%)
Yan et al. [50]	2019	Transformer	98.97	88.73
Xu et al. [51]	2020	CNN-BiLSTM	95.90	95.92
Wang et al. [52]	2021	CNN-Attention	98.64	96.64
Essa et al. [53]	2021	Ensemble of CNN-LSTM & RRHOS-LSTM	95.81	71.06
Zhao et al. [54]	2023	Attention-based TCN	87.81	89.46
EI-Ghaish et al. [21]	2024	MSCNN+BiTransformer	99.11	94.39
Islam et al. [38]	2024	CNN+Attention+Transformer	99.14	94.69
Xu et al. [49]	2024	MM-DANet	98.83	91.96
Ours	**2025**	**EPLM(Param3)**	**99.19**	**94.51**

Table 14
Comparison with state-of-the-art works on the PTB-DB database.

Authors	Year	Method (s)	Accuracy(%)
Acharya et al. [55]	2017	CNN	93.50
Diker et al. [56]	2020	Extreme learning machine	97.50
Zhang et al. [57]	2023	EfficientNet+Ensemble learning	99.60
Alkhawaldeh et al. [58]	2023	BiLSTM	98.70
Anwar et al. [59]	2024	KNN	99.74
Ours	**2025**	**EPLM(Param3)**	**99.13**

convolution, pointwise convolution, and a two-branch-based element-wise product fusion module. The model achieved 99.10% and 98.85% accuracy in the MIT-BIH and the PTB database. In addition, the model is validated real-time arrhythmias monitoring on a Raspberry Pi 5 platform and a Android 10 ×86 virtual machine, achieving a single inference time of 2.29 ms and 0.62 ms, respectively. Our method outperforms current models in terms of efficacy and precision, showing its applicability for real-time monitoring. The model's compact design, minimal parameters, and low computational cost do not affect its accuracy in arrhythmia diagnosis, making it a promising option for real-time monitoring on embedded platforms. As our model has solely been trained on databases that are accessible to the public, we have not yet examined its application in an actual clinical study. In the future, our goal is to design a specialized module for acquiring ECG signals that will be an essential part of a real-time system for monitoring arrhythmia. This module will enable comprehensive clinical trials to determine the effectiveness and feasibility of the EPLM in real clinical environments.

CRediT authorship contribution statement

Haotian Tang: Writing – review & editing, Writing – original draft, Investigation, Conceptualization. **Mingke Yan:** Writing – review & editing, Writing – original draft, Investigation, Conceptualization. **Xidong Wu:** Software, Investigation. **Hongkai Lai:** Writing – review & editing, Methodology. **Yang Zhang:** Writing – review & editing, Methodology. **Hanyu Cui:** Methodology, Funding acquisition. **Liping Xie:** Writing – review & editing, Writing – original draft, Software, Funding acquisition, Conceptualization.

Ethics statement

This research was conducted in full compliance with ethical guidelines and regulations governing biomedical research. The following ethical considerations were adhered to during the course of the study:

Human participants and informed consent

This research didn't involve human participants

Confidentiality and privacy

All data involving human subjects were from the public dataset, which are anonymized or de-identified to ensure confidentiality and protect participants' privacy. Any sensitive or personal information was securely stored and handled in accordance with relevant data protection regulations.

Data integrity and reproducibility

All results presented in this study are based on rigorously analyzed data. The authors commit to ensuring the transparency and reproducibility of the research by making data and methods available upon reasonable request, in line with best practices for scientific research.

Responsible conduct of research

The authors affirm that the research was conducted with the highest ethical standards and scientific integrity. This includes the accurate presentation of results, acknowledgment of prior works, and the proper citation of sources. We also acknowledge the importance of collaboration, peer review, and transparency in the research process.

By following these ethical guidelines, we ensure that the research aligns with the core principles of biomedical science, promoting the well-being and trust of the scientific community and the public.

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.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (No. 62576084), the Natural Science Foundation of Liaoning Province of China (No. 2022-MS-118), and National Training Program of Innovation and Entrepreneurship for Undergraduates, China (No. 251406).

Data availability

The data used in the article was from the public database.

References

[1] A.D. Callow, Cardiovascular disease 2005 — the global picture, Vasc. Pharmacol. 45 (5) (2006) 302–307, http://dx.doi.org/10.1016/j.vph.2006.08.010.
[2] A. Timmis, P. Vardas, N. Townsend, A. Torbica, H. Katus, D. De Smedt, C.P. Gale, A.P. Maggioni, S.E. Petersen, R. Huculeci, et al., European society of cardiology: cardiovascular disease statistics 2021, Eur. Heart J. 43 (8) (2022) 716–799.
[3] D. Zhao, Epidemiological features of cardiovascular disease in Asia, JACC: Asia 1 (1) (2021) 1–13, http://dx.doi.org/10.1016/j.jacasi.2021.04.007.
[4] Z.F.M. Apandi, R. Ikeura, S. Hayakawa, Arrhythmia detection using MIT-bih dataset: A review, in: International Conference on Computational Approach in Smart Systems Design and Applications, ICASSDA, in: International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA), IEEE, 2018, pp. 1–5.
[5] N.T. Srinivasan, R.J. Schilling, Sudden cardiac death and arrhythmias, Arrhythmia Electrophysiol. Rev. 7 (2) (2018) 111, http://dx.doi.org/10.15420/aer.2018:15:2.
[6] O. Yildirim, P. Plawiak, R.S. Tan, U.R. Acharya, Arrhythmia detection using deep convolutional neural network with long duration ECG signals, Comput. Biol. Med. 102 (2018) 411–420, http://dx.doi.org/10.1016/j.compbiomed.2018.09.009.
[7] S. Kaplan Berkaya, A.K. Uysal, E. Sora Gunal, S. Ergin, S. Gunal, M.B. Gulmezoglu, A survey on ECG analysis, Biomed. Signal Process. Control. 43 (2018) 216–235, http://dx.doi.org/10.1016/j.bspc.2018.03.003.
[8] P. Zimetbaum, A. Goldman, Ambulatory arrhythmia monitoring, Circulation 122 (16) (2010) 1629–1636, http://dx.doi.org/10.1161/CIRCULATIONAHA.109.925610.
[9] H. De Melo Ribeiro, A. Arnold, J.P. Howard, M.J. Shun-Shin, Y. Zhang, D.P. Francis, P.B. Lim, Z. Whinnett, M. Zolgharni, ECG-based real-time arrhythmia monitoring using quantized deep neural networks: A feasibility study, Comput. Biol. Med. 143 (2022) 105249, http://dx.doi.org/10.1016/j.compbiomed.2022.105249.
[10] L. Xie, Z. Zhang, Q. Wu, Z. Gao, G. Mi, R. Wang, H.-b. Sun, Y. Zhao, Y. Du, Intelligent wearable devices based on nanomaterials and nanostructures for healthcare, Nanoscale 15 (2) (2023) 405–433, http://dx.doi.org/10.1039/D2NR04551F.
[11] S.H. Kwon, L. Dong, Flexible sensors and machine learning for heart monitoring, Nano Energy 102 (2022) 107632, http://dx.doi.org/10.1016/j.nanoen.2022.107632.
[12] L. Xie, L. Wang, D. Mo, Z. Zhang, M. Liang, Intelligent algorithms powered smart devices for atrial fibrillation discrimination, Biomed. Signal Process. Control. 103 (2025) 107480, http://dx.doi.org/10.1016/j.bspc.2024.107480.
[13] X. Liu, H. Wang, Z. Li, L. Qin, Deep learning in ECG diagnosis: A review, Knowl.-Based Syst. 227 (2021) 107187, http://dx.doi.org/10.1016/j.knosys.2021.107187.

[14] R. Rajkumar, S. Gopalakrishnan, K. Praveena, M. Venkatesan, K. Ramamoorthy, J.J. Hephzipah, DARKNET-53 convolutional neural network-based image processing for breast cancer detection, Mesopotamian J. Artif. Intell. Heal. 2024 (2024) 59–68.

[15] A.A. Khan, M.M.A. Shahid, R.N. Bashir, S. Iqbal, A.S.A. Shahid, J. Maqbool, C. Wechtaisong, Detection of omicron caused pneumonia from radiology images using convolution neural network (CNN), Comput. Mater. Contin. 74 (2) (2023).

[16] A.A. Khan, S. Alsubai, A. Almadhor, N. Kryvinska, A. Al Hejaili, U.G. Mohammad, et al., CD-FL: Cataract images based disease detection using federated learning, Comput. Syst. Sci. Eng. 47 (2) (2023) http://dx.doi.org/10.32604/csse.2023.039296.

[17] P. Thapar, M. Rakhra, D. Prashar, L. Mrsic, A.A. Khan, S. Kadry, Skin cancer segmentation and classification by implementing a hybrid FrCN-(U-NeT) technique with machine learning, PLoS One 20 (6) (2025) e0322659, http://dx.doi.org/10.1371/journal.pone.0322659.

[18] N.A. Trayanova, D.M. Popescu, J.K. Shade, Machine learning in arrhythmia and electrophysiology, Circ. Res. 128 (4) (2021) 544–566, http://dx.doi.org/10.1161/CIRCRESAHA.120.317872.

[19] A.K. Feeny, M.K. Chung, A. Madabhushi, Z.I. Attia, M. Cikes, M. Firouznia, P.A. Friedman, M.M. Kalscheur, S. Kapa, S.M. Narayan, P.A. Noseworthy, R.S. Passman, M.V. Perez, N.S. Peters, J.P. Piccini, K.G. Tarakji, S.A. Thomas, N.A. Trayanova, M.P. Turakhia, P.J. Wang, Artificial intelligence and machine learning in arrhythmias and cardiac electrophysiology, Circ. Arrhythmia Electrophysiol. 13 (8) (2020) http://dx.doi.org/10.1161/CIRCEP.119.007952, e007952–e007952.

[20] R. Anand, S.V. Lakshmi, D. Pandey, B.K. Pandey, An enhanced ResNet-50 deep learning model for arrhythmia detection using electrocardiogram biomedical indicators, Evol. Syst. 15 (1) (2024) 83–97, http://dx.doi.org/10.1007/s12530-023-09559-0.

[21] H. El-Ghaish, E. Eldele, ECGTransForm: Empowering adaptive ECG arrhythmia classification framework with bidirectional transformer, Biomed. Signal Process. Control. 89 (2024) 105714, http://dx.doi.org/10.1016/j.bspc.2023.105714.

[22] M. Hammad, A.M. Iliyasu, A. Subasi, E.S.L. Ho, A.A.A. El-Latif, A multitier deep learning model for arrhythmia detection, IEEE Trans. Instrum. Meas. 70 (2021) 1–9, http://dx.doi.org/10.1109/TIM.2020.3033072.

[23] X. Wu, M. Yan, R. Wang, L. Xie, Multiscale feature enhanced gating network for atrial fibrillation detection, Comput. Methods Programs Biomed. 261 (2025) 108606, http://dx.doi.org/10.1016/j.cmpb.2025.108606.

[24] X. Wu, M. Yan, H. Tang, D. Wu, L. Xie, MSCGN: Multiscale complementary gating network for time series classification, Biomed. Signal Process. Control. 112 (2026) 108563.

[25] N.D. Lane, S. Bhattacharya, P. Georgiev, C. Forlivesi, F. Kawsar, An early resource characterization of deep learning on wearables, smartphones and internet-of-things devices, 2015, pp. 7–12.

[26] A. Ukil, I. Sahu, A. Majumdar, S.C. Racha, G. Kulkarni, A.D. Choudhury, S. Khandelwal, A. Ghose, A. Pal, Resource constrained CVD classification using single lead ECG on wearable and implantable devices, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in edicine and Biology Society. Annual International Conference, vol. 2021, 2021, pp. 886–889, http://dx.doi.org/10.1109/EMBC46164.2021.9630348.

[27] Z. He, X. Zhang, Y. Cao, Z. Liu, B. Zhang, X. Wang, LiteNet: Lightweight neural network for detecting arrhythmias at resource-constrained mobile devices, Sensors 18 (4) (2018) 1229, http://dx.doi.org/10.3390/s18041229.

[28] A. Ukil, L. Marin, S.C. Mukhopadhyay, A.J. Jara, Afsense-ECG: Atrial fibrillation condition sensing from single lead electrocardiogram (ECG) signals, IEEE Sensors J. 22 (12) (2022) http://dx.doi.org/10.1109/JSEN.2022.3162691, 1–1.

[29] N. Phukan, M.S. Manikandan, R.B. Pachori, AFibri-Net: A lightweight convolution neural network based atrial fibrillation detector, IEEE Trans. Circuits Syst. I. Regul. Pap. 70 (12) (2023) 4962–4974, http://dx.doi.org/10.1109/TCSI.2023.3303936.

[30] M. Gu, Y. Zhang, Y. Wen, G. Ai, H. Zhang, P. Wang, G. Wang, A lightweight convolutional neural network hardware implementation for wearable heart rate anomaly detection, Comput. Biol. Med. 155 (2023) 106623, http://dx.doi.org/10.1016/j.compbiomed.2023.106623.

[31] L. Meng, W. Tan, J. Ma, R. Wang, X. Yin, Y. Zhang, Enhancing dynamic ECG heartbeat classification with lightweight transformer model, Artif. Intell. Med. 124 (2022) 102236, http://dx.doi.org/10.1016/j.artmed.2022.102236.

[32] G.B. Moody, R.G. Mark, The impact of the MIT-BIH arrhythmia database, IEEE Eng. Med. Biology Mag. : Quarterly Mag. Eng. Med. Biology Soc. 20 (3) (2001) 45–50, http://dx.doi.org/10.1109/51.932724.

[33] R. Bousseljot, D. Kreiseler, A. Schnabel, Nutzung der EKG-signaldatenbank CARDIODAT der PTB über das internet, Biomed. Eng. - Biomed. Tech. 40 (s1) (1995) 317–318, http://dx.doi.org/10.1515/bmte.1995.40.s1.317.

[34] E. Yakushenko, St petersburg incart 12-lead arrhythmia database, PhysioBank, PhysioToolkit, PhysioNet (2008).

[35] R. Mark, G. Moody, S. Greenwald, Mit-bih supraventricular arrhythmia database, 1990.

[36] M. Kachuee, S. Fazeli, M. Sarrafzadeh, Ecg heartbeat classification: A deep transferable representation, in: 2018 IEEE International Conference on Healthcare Informatics, ICHI, IEEE, 2018, pp. 443–444.

[37] Y. Jin, Z. Li, M. Wang, J. Liu, Y. Tian, Y. Liu, X. Wei, L. Zhao, C. Liu, Cardiologist-level interpretable knowledge-fused deep neural network for automatic arrhythmia diagnosis, Commun. Med. 4 (1) (2024) 31, http://dx.doi.org/10.1038/s43856-024-00464-4.

[38] M.R. Islam, M. Qaraqe, K. Qaraqe, E. Serpedin, CAT-net: Convolution, attention, and transformer based network for single-lead ECG arrhythmia classification, Biomed. Signal Process. Control. 93 (2024) 106211, http://dx.doi.org/10.1016/j.bspc.2024.106211.

[39] J. Yang, C. Li, X. Dai, J. Gao, Focal modulation networks, Adv. Neural Inf. Process. Syst. 35 (2022) 4203–4217.

[40] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, S.-M. Hu, Visual attention network, Comput. Vis. Media 9 (4) (2023) 733–752, http://dx.doi.org/10.1007/s41095-023-0364-2.

[41] X. Ma, X. Dai, J. Yang, B. Xiao, Y. Chen, Y. Fu, L. Yuan, Efficient modulation for vision networks, in: The Twelfth International Conference on Learning Representations, 2024, URL https://openreview.net/forum?id=ip5LHJs6QX.

[42] T. Dao, A. Gu, Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality, in: International Conference on Machine Learning, ICML, 2024.

[43] N. Shazeer, GLU variants improve transformer, 2020.

[44] X. Ma, X. Dai, Y. Bai, Y. Wang, Y. Fu, Rewrite the stars, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), CVPR, 2024, pp. 5694–5703.

[45] Y.M.H. Elsayed, Premature ventricular contractions from benign to seriousness-a narrative updating review, Arch. Emerg. Med. Intensiv. Care 2 (2) (2019) 1–21.

[46] N. Alamatsaz, L. Tabatabaei, M. Yazdchi, H. Payan, N. Alamatsaz, F. Nasimi, A lightweight hybrid CNN-LSTM explainable model for ECG-based arrhythmia detection, Biomed. Signal Process. Control. 90 (2024) 105884, http://dx.doi.org/10.1016/j.bspc.2023.105884.

[47] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C.B.-G.-A.I. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, 2018, pp. 4510–4520.

[48] Y. Li, Y. Chen, X. Dai, D. Chen, M. Liu, L. Yuan, Z. Liu, L. Zhang, N. Vasconcelos, Micronet: Improving image recognition with extremely low flops, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 468–477.

[49] Z. Xu, M. Zang, T. Liu, Z. Wang, S. Zhou, C. Liu, Q. Wang, P. Surya, S. Prakash, Multimodality data augmentation network for arrhythmia classification, Int. J. Intell. Syst. 2024 (1) (2024) 9954821, http://dx.doi.org/10.1155/2024/9954821.

[50] G. Yan, S. Liang, Y. Zhang, F. Liu, Fusing Transformer Model with Temporal Features for ECG Heartbeat Classification, IEEE, 2019, pp. 898–905.

[51] X. Xu, S. Jeong, J. Li, Interpretation of electrocardiogram (ECG) rhythm by combined CNN and bilstm, IEEE Access 8 (2020) 125380–125388, http://dx.doi.org/10.1109/ACCESS.2020.3006707.

[52] J. Wang, X. Qiao, C. Liu, X. Wang, Y. Liu, L. Yao, H. Zhang, Automated ECG classification using a non-local convolutional block attention module, Comput. Methods Programs Biomed. 203 (2021) 106006, http://dx.doi.org/10.1016/j.cmpb.2021.106006.

[53] E. Essa, X. Xie, An ensemble of deep learning-based multi-model for ECG heartbeats arrhythmia classification, IEEE Access 9 (2021) 103452–103464, http://dx.doi.org/10.1109/ACCESS.2021.3098986.

[54] Y. Zhao, J. Ren, B. Zhang, J. Wu, Y. Lyu, An explainable attention-based TCN heartbeats classification model for arrhythmia detection, Biomed. Signal Process. Control. 80 (2023) 104337, http://dx.doi.org/10.1016/j.bspc.2022.104337.

[55] U.R. Acharya, H. Fujita, S.L. Oh, Y. Hagiwara, J.H. Tan, M. Adam, Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals, Inform. Sci. 415–416 (2017) 190–198, http://dx.doi.org/10.1016/j.ins.2017.06.027.

[56] A. Diker, E. Avci, E. Tanyildizi, M. Gedikpinar, A novel ECG signal classification method using DEA-elm, Med. Hypotheses 136 (2020) 109515, http://dx.doi.org/10.1016/j.mehy.2019.109515.

[57] L. Zhang, S. Chen, F. Lin, W. Ren, K.-K.R. Choo, G. Min, 1DIEN: Cross-session electrocardiogram authentication using 1d integrated EfficientNet, ACM Trans. Multimed. Comput. Commun. Appl. 20 (1) (2023) 1–17, http://dx.doi.org/10.1145/3609800.

[58] R.S. Alkhawaldeh, B. Al-Ahmad, A. Ksibi, N. Ghatasheh, E.M. Abu-Taieh, G. Aldehim, M. Ayadi, S.M. Alkhawaldeh, Convolution neural network bidirectional long short-term memory for heartbeat arrhythmia classification, Int. J. Comput. Intell. Syst. 16 (1) (2023) 197, http://dx.doi.org/10.1007/s44196-023-00374-8.

[59] S.M.S. Anwar, D. Pal, S. Mukhopadhyay, R. Gupta, A lightweight method of myocardial infarction detection and localization from single lead ECG features using machine learning approach, IEEE Sensors Lett. 8 (4) (2024) 1–4, http://dx.doi.org/10.1109/LSENS.2024.3374790.

[60] T. Nay, Enhancing IoT security with AI-driven hybrid machine learning and neural network-based intrusion detection system, Babylon. J. Artif. Intell. 2024 (2024) 158–167.

[61] A.A. Khan, M. Driss, W. Boulila, G.A. Sampedro, S. Abbas, C. Wechtaisong, Privacy preserved and decentralized smartphone recommendation system, IEEE Trans. Consum. Electron. 70 (1) (2023) 4617–4624, http://dx.doi.org/10.1109/TCE.2023.3323406.