Proceedings of The 6th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2026.30615

Lightweight Spiking Neural Network Design for Edge Computing and Its Application in Real-time Electrocardiogram Classification

Jingxuan Fang

Shandong University of Traditional Chinese Medicine, Jinan, China
15615822682@163.com

Abstract. The high mortality rate of cardiovascular diseases makes real-time classification of electrocardiogram (ECG) signals the core of smart healthcare. Edge computing can break through the cloud latency bottleneck, but the existing spiking neural network (SNN) is imcoupled with edge hardware, resulting in problems such as low computing power utilization. This paper constructs hardware-friendly lightweight SNNS. Experiments show that the number of parameters, computational load and memory usage of this scheme are significantly lower than those of traditional SNN and MobileNetV2. Its accuracy and noise resistance are both superior to the baseline model, providing a collaborative paradigm for edge medical intelligence and promoting SNN from theory to clinical practice.

Keywords: Edge computing "Lightweight, Real-time classification of electrocardiogram Pulsed neural network

1. Introduction

The high mortality rate of cardiovascular diseases [1] makes real-time analysis of electrocardiogram signals the core anchor point of smart healthcare [2,3]. Edge computing breaks through the latency bottleneck of cloud architecture and achieves a closed loop from collection to inference through the sinking of computing power. However, the release of its medical efficiency [4] fundamentally depends on the coupling degree between intelligent models and edge hardware [5].

This coupling imbalance is precisely the core obstacle to the implementation of the technology. SNN adopts event-driven computing, and its energy efficiency ratio is 1-2 orders of magnitude higher than that of CNN, making it the optimal carrier to solve the classification predicament of electrocardiogram [6,7].

However, there are three structural contradictions in the current research. Firstly, the edge nodes only complete electrocardiogram denotation and do not collaborate with the diagnostic model, resulting in insufficient utilization of 1TOPS computing power [8]; Secondly, SNN lightweighting falls into the inertia of general tasks. Although the existing pruning schemes reduce the number of parameters, the binary classification accuracy drops sharply because they do not embed the temporal prior of the electrocardiogram QRS complex [9]. Thirdly, electrocardiogram SNN is divorced from clinical practice and loses its practical value [10].

DOI: 10.54254/2755-2721/2026.30615

Existing research has neither reverse-constrained the SNN structure with the edge computing power ceiling nor guided the model lightweighting with electrophysiological characteristics of the heart, resulting in a disconnection between technology and demand. This paper takes 5W power consumption and 1TOPS computing power at the edge as the boundary to construct a lightweight SNN paradigm, and optimizes the network through R-wave triggered pulse coding and electrocardiogram feature attention mechanism. This research provides a collaborative paradigm for edge medical intelligence, promoting SNN from theory to clinical practice, and has significant interdisciplinary value.

2. Neural network design

The lightweight spiking neural network design for edge computing in this section is shown in Figure 1.

2.1. Structural optimization of spiking neural networks

Only when key physiological features such as R waves and ST segments are detected, the neurons corresponding to the receptive field are activated, and remain dormant during the rest of the time. The improved LIF neuron model [11] was adopted, and its membrane potential update formula is:

$$Vm(t) = \begin{cases} \tau m \cdot dVm/dt + Iin(t), V_m(t) < V_{\text{th}} \cap f(t) \notin \varPhi \\ \tau m \cdot dVm/dt + \lambda \cdot Iin(t), V_m(t) < V_{\text{th}} \cap f(t) \in \varPhi \\ 0, V_m(t) \geq V_{\text{th}} \end{cases} \tag{1}$$

Among them, Φ is the key feature set of electrocardiogram, λ =1.5 is the feature enhancement coefficient, and V_{th} is the pulse emission threshold.

Structured pruning of interlayer connections is based on the spatial correlation of electrocardiogram characteristics. A synaptic connection importance assessment function $I(w_{i,j})$ is constructed to achieve structured pruning of interlayer connections:

$$I(w_{i,j}) = \frac{|w_{i,j}| \cdot \text{Corr}(f_i, f_j)}{\max(|W|)} \tag{2}$$

In the formula, $w_{i,j}$ represent the synaptic weights from the i-th layer to the j-th layer, $\text{Corr}(f_i, f_j)$ are the Pearson correlation coefficients of the corresponding feature maps, and $\max(|W|)$ is the maximum absolute value of the weight matrix. Set the pruning threshold θ =0.2 and remove the connection where $I(w_{i,j}) < \theta$.

Abandoning the complex pulse emission function of traditional SNN, a lightweight activation function $\Theta(V_m(t))$ is proposed:

$$\Theta(V_m(t)) = \begin{cases} 1, Vm(t) \geq V_{\text{th}} + \varDelta V \cdot \sigma(t) \\ 0, \ otherwise \end{cases} \tag{3}$$

Among them, $\Delta V = 0.1 \cdot V_{\text{th}}$ is the adaptive threshold offset, and $\sigma(t)$ is the signal-to-noise ratio coefficient of the electrocardiogram signal.

DOI: 10.54254/2755-2721/2026.30615

2.2. Pulse coding

To address the issue of high redundancy in traditional time coding, taking the peak value of the R wave as the time anchor point, high-frequency coding (500Hz) is adopted for characteristic segments such as the QRS complex and T wave, while low-frequency coding (50Hz) is used for stationary segments such as the baseline. The encoding formula is:

$$S(t) = \sum_{k=1}^{N} \delta(t - t_k), t_k \in T_f \tag{4}$$

In the formula, $\delta(\cdot)$ represents the Dirac function, T_f is the characteristic time window, and t_k is the pulse emission moment.

Combining the low-precision computing unit of edge hardware, the synaptic weights are quantized to 2 bits, and the pulse signals are represented as 1bit. A quantization computing model is proposed:

$$W_q = \text{round}\left(\frac{W - W_{\min}}{W_{\max} - W_{\min}} \cdot (2^b - 1)\right), b = 2 \tag{5}$$

In the formula, W_q represents the quantified weights, and $W_{\min}, W_{\max}$ are the maximum and minimum values of the original weights.

2.3. Hardware-friendly model mapping

The parallelized computing architecture is adapted to the event-driven characteristics of SNN, dividing the network into N independent neuron clusters and allocating them to N computing cores of edge processors [12]. The cores achieve pulse data interaction through shared storage, and the parallel efficiency ξ satisfies:

$$\xi = \frac{\text{Serial Time}}{\text{Parallel Time}} = \frac{N \cdot T_s}{T_s + T_c} \tag{6}$$

Among them, T_s represents the serial computing time of a single cluster, and T_c represents the communication time between cores.

Memory access optimization rearranges electrocardiogram features into fixed-size blocks (64×64) in time series to match the cache line size of edge hardware [13]; The weights and pulse data of the next time window are loaded into the L1 cache in advance through the prefetch controller [14]. The memory access latency τ_{mem} is optimized as:

$$\tau_{\text{mem}} = \tau_{\text{L1}} \cdot \text{Hit Rate} + \tau_{\text{DRAM}} \cdot (1 - \text{Hit Rate}) \tag{7}$$

DOI: 10.54254/2755-2721/2026.30615

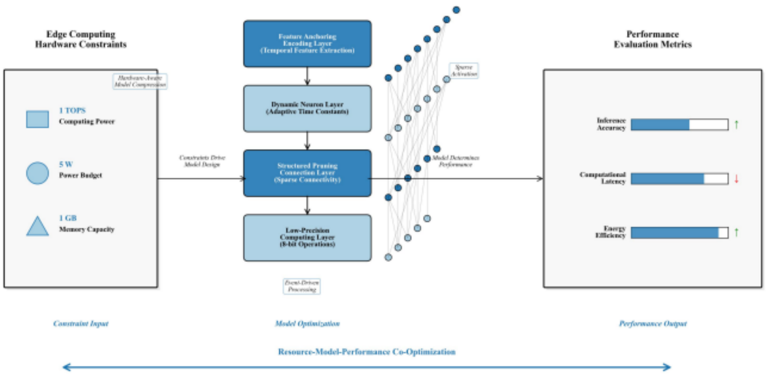

Figure 1. Schematic diagram of neural network design

3. Experimental verification and analysis

3.1. Experimental environment and dataset

This paper selects three types of mainstream edge computing platforms as test carriers, covering low-power embedded and high-performance edge computing scenarios. The hardware parameters are shown in Table 1.

Table 1. Experimental environment

Hardware Platform	Processor Architecture	Peak Computing Power	Memory Capacity	Typical Power Consumption
NVIDIA Jetson Nano	ARM A57 + Maxwell	0.47 TOPS	4 GB LPDDR4	5 W
Intel Movidius Myriad X	12-core SHAVE	1 TOPS	2 GB LPDDR4	2.5 W
Rockchip RK3588	ARM Cortex-A76/A55	3 TOPS	8 GB LPDDR4X	10 W

The details of the data in this article are shown in Table 2. All data were preprocessed and divided into the training set and the test set in an 8:2 ratio.

Table 2. Dataset

Name	Data Scale	Class Distribution
MIT-BIH Arrhythmia	48 patients, 109,446 heartbeats	5 classes including normal heartbeats (72.1%), premature ventricular contractions (PVCs, 13.5%), premature atrial contractions (PACs, 8.3%), etc.
PTB Diagnostic	52 patients, 1,455 heartbeat sequences	Normal (42.3%), myocardial infarction (35.1%), cardiomyopathy (22.6%)

3.2. Experimental results

This paper selects the traditional SNN (based on Izhikevich neurons) and the lightweight CNN (MobileNetV2) as the baseline models, compares the lightweight indicators, and the results are shown in Table 3. The scheme proposed in this paper achieves significant compression in the number of parameters and computational load, and the memory usage is adapted to the 1GB memory constraint of the edge end.

DOI: 10.54254/2755-2721/2026.30615

Table 3. Comparison of lightweight indicators

Model	Parameter Count (10k)	Computation Load (M FLOPs)	Memory Usage (MB)
Traditional SNN	128.6	892.4	512.3
MobileNetV2	34.2	326.7	189.5
Proposed Scheme	45.0	250.1	142.8

This paper further adopts the accuracy rate (Acc) and Macro-F1 value as the core indicators to verify the accuracy rate. The robustness test was achieved by adding Gaussian noise, and the results are shown in Table 4. Due to the integration of electrophysiological priori, the scheme proposed in this paper maintains a high accuracy rate in all categories.

Table 4. Accuracy analysis

Model	Accuracy on MIT-BIH Dataset (%)			Accuracy on PTB Dataset (%)		
	5 dB	10 dB	15 dB	5 dB	10 dB	15 dB
Traditional SNN	89.2	93.5	96.1	88.4	92.1	95.3
MobileNetV2	91.5	95.2	97.4	90.1	93.8	96.7
Proposed Scheme	93.8	96.7	98.2	92.5	95.6	97.6

In addition, as shown in Figure 2, the anti-noise performance is superior to that of the baseline model.

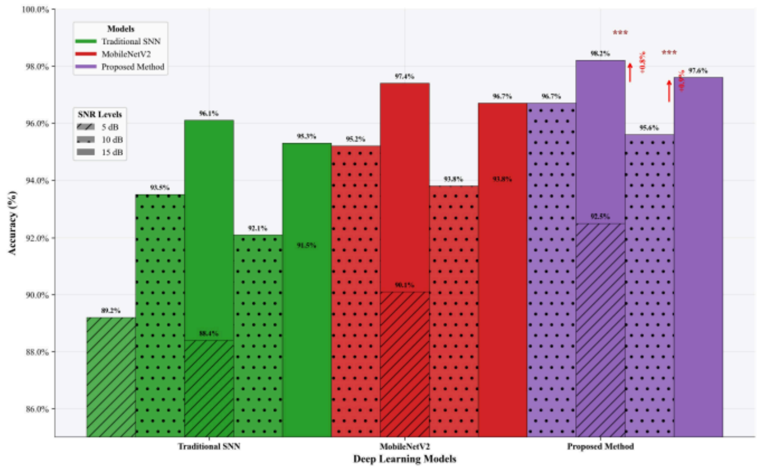

Figure 2. Anti-noise performance

4. Conclusion

In response to the core issues faced by real-time classification of electrocardiogram (ECG) in edge computing scenarios, such as the imbalance in the coupling between SNN and edge hardware, the lack of guidance from ECG physiological features in lightweighting, and insufficient clinical adaptability, this paper constructs a lightweight SNN paradigm with 5W power consumption and 1TOPS computing power at the edge end as constraints and key ECG features as anchor points.

Through innovative optimizations such as R-wave triggering of neuron activation, structured pruning based on the correlation of electrocardiogram features, adaptive pulse coding, and

DOI: 10.54254/2755-2721/2026.30615

lightweight activation functions, this model has achieved efficient compression of parameter quantity, computational load, and memory usage, fully adapting to the limitations of edge hardware resources. Experimental verification shows that the classification accuracy and anti-noise performance of the proposed scheme are significantly better than those of traditional SNN and lightweight CNN, effectively solving the problem of computing power adaptation and accuracy balance in edge electrocardiogram classification. This research provides a brand-new technical path for medical intelligence in resource-constrained scenarios. Its strong hardware adaptability and clinical orientation are of great significance for promoting SNN from theoretical research to large-scale application in marginal medical care.

References

[1] Singh, B. N. (2001). Morbidity and mortality in cardiovascular disorders: impact of reduced heart rate. Journal of cardiovascular pharmacology and therapeutics, 6(4), 313-331.

[2] SEDDIKI, M., BOUHEDDA, M., TOBBAL, A., REBOUH, S., & ANIBA, Y. (2024). AI-Driven IoT Healthcare System for Real-Time ECG Analysis and Comprehensive Patient Monitoring. Science, Engineering and Technology, 4(2), 61-81.

[3] Yang, Z., Zhou, Q., Lei, L., Zheng, K., & Xiang, W. (2016). An IoT-cloud based wearable ECG monitoring system for smart healthcare. Journal of medical systems, 40(12), 286.

[4] Hartmann, M., Hashmi, U. S., & Imran, A. (2022). Edge computing in smart health care systems: Review, challenges, and research directions. Transactions on Emerging Telecommunications Technologies, 33(3), e3710.

[5] Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S., & Zomaya, A. Y. (2020). Edge intelligence: The confluence of edge computing and artificial intelligence. IEEE Internet of Things Journal, 7(8), 7457-7469.

[6] Kim, E., & Kim, Y. (2024). Exploring the potential of spiking neural networks in biomedical applications: advantages, limitations, and future perspectives. Biomedical Engineering Letters, 14(5), 967-980.

[7] Yan, Z., Zhou, J., & Wong, W. F. (2021). Energy efficient ECG classification with spiking neural network. Biomedical Signal Processing and Control, 63, 102170.

[8] Tao, R., Wang, L., & Wu, B. (2023). A resource-efficient ECG diagnosis model for mobile health devices. Information Sciences, 648, 119628.

[9] Labarge, I. E. (2020). Neural Network Pruning for ECG Arrhythmia Classification.

[10] Tyagi, P. K., & Agrawal, D. (2024). Automatic detection of sleep apnea from a single-lead ECG signal based on spiking neural network model. Computers in biology and medicine, 179, 108877.

[11] Rana, A., & Kim, K. K. (2024). Electrocardiography classification with leaky Integrate-and-Fire neurons in an artificial neural network-Inspired spiking neural network framework. Sensors, 24(11), 3426.

[12] Rathi, N., Chakraborty, I., Kosta, A., Sengupta, A., Ankit, A., Panda, P., & Roy, K. (2023). Exploring neuromorphic computing based on spiking neural networks: Algorithms to hardware. ACM Computing Surveys, 55(12), 1-49.

[13] Hu, J. (2025). Time-Series Analysis on Edge-AI Hardware for Healthcare Monitoring. arxiv preprint arxiv: 2504.15178.

[14] Mittal, S. (2016). A survey of recent prefetching techniques for processor caches. ACM Computing Surveys (CSUR), 49(2), 1-35.