Stony Brook University

MICCAI 2026 Paper 4698

MICCAI2026
29TH INTERNATIONAL CONFERENCE ON MEDICAL IMAGE COMPUTING & COMPUTER ASSISTED INTERVENTION
Sept. 27- Oct. 1, 2026
Strasbourg Convention Center, FRANCE

SpikeMamba: Spike-Driven State Space Models for Energy-Efficient Biomedical Sequence Modeling

Si Yong Lee[1], Ryangjin Lee[1], Hawon Park[1], Yoora Kim[1], Byungkon Kang[2], Yoon Seok Yang[2]

[1] Department of Computer Science, Stony Brook University, USA [2] Department of Computer Science, SUNY Korea, Republic of Korea

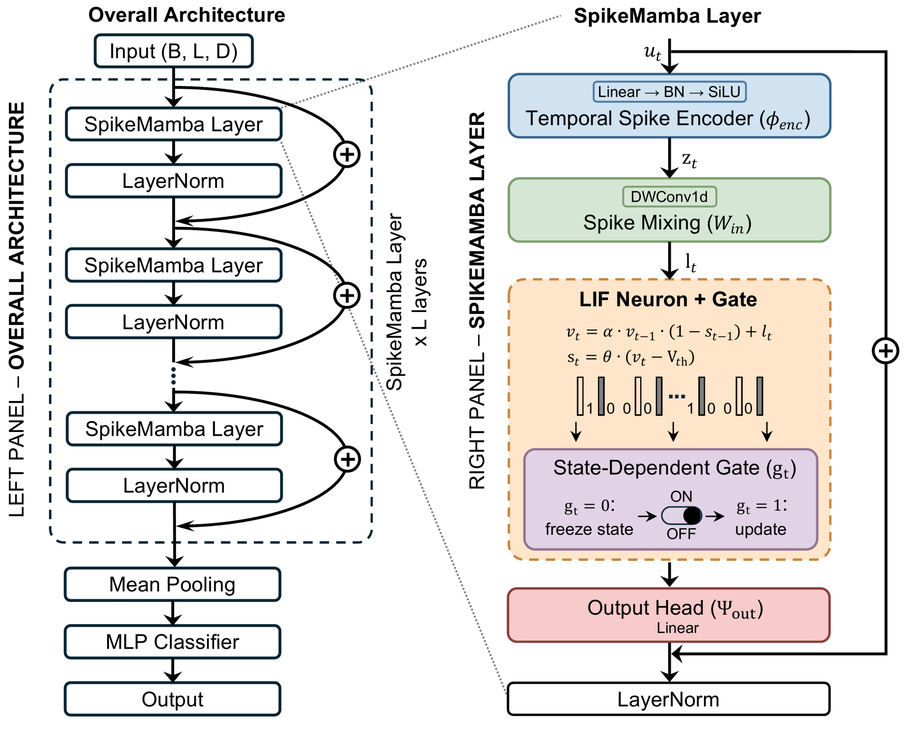
1 Motivation and formulation

Selective SSMs model long biomedical sequences in O(N) time, but still require dense multiply-accumulate (MAC) operations. SpikeMamba uses sparse binary events to reduce the energy cost of recurrent computation.

2 Main contributions

- LIF recurrence gives a spike-domain SSM formulation.
- A learned gate suppresses inactive state updates.
- Five ECG/EEG benchmarks test accuracy and the analytical energy benefit across three model scales.

3 Architecture and temporal encoding



Temporal spike encoder

$$z_t = \mathrm{SiLU}(\mathrm{BN}(W_{\mathrm{enc}} u_t + b_{\mathrm{enc}})), \quad I_t = W_{\mathrm{in}} * z_t$$

BN stabilizes pre-spike activations to avoid all-fire or all-silent behavior. Depthwise 1D convolution generates the input current. Residual layers feed a mean-pooled MLP classifier.

4 State dynamics and event-driven gating

Selective SSM and structural correspondence

$$h_t = \bar{A} \odot h_{t-1} + \bar{B} \odot x_t, \qquad y_t = C h_t$$

$$\bar{A} \leftrightarrow \alpha \quad (\text{decay}), \qquad \bar{B} \odot x_t \leftrightarrow I_t \quad (\text{input})$$

The correspondence concerns state decay and input injection. LIF adds thresholded binary output and a reset.

Leaky integrate-and-fire dynamics

$$v_t = \alpha v_{t-1}(1 - s_{t-1}) + I_t$$
$$s_t = \Theta(v_t - V_{\mathrm{th}}) \in \{0,1\}^D$$
$$v_t \leftarrow v_t - s_t V_{\mathrm{th}}$$

The leak α is learnable per neuron, with 0 < α < 1.
Θ is the Heaviside function, trained with surrogate gradients.
For binary s, a weighted product w·s is either w or zero.

Gate based on activity and membrane variance

$$g_t = \Theta\big(\sigma\big(W_g[\bar{s}_t \,\|\, \sigma_v^2(t)] + b_g\big) - \tau\big)$$

$$\bar{s}_t = \frac{1}{D}\sum_d s_{t,d}, \qquad \sigma_v^2(t) = \mathrm{Var}(v_t)$$

$$g_t = 0 \implies v_t \leftarrow v_{t-1}, \quad s_t \leftarrow 0$$

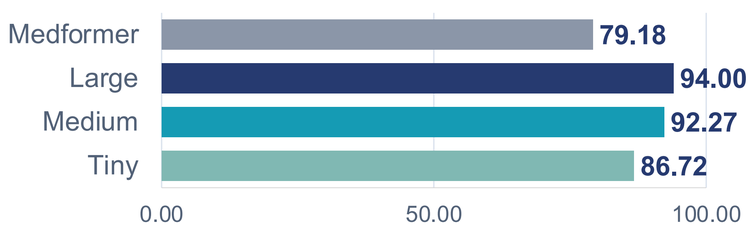
The sigmoid σ combines these statistics with learned weights.
Threshold τ is learnable. A closed gate freezes the state.

5 Classification performance and model scale

F1 (%) with a common 0–100 axis. Selected comparison: Medformer and SpikeMamba-L / M / T. Subject-independent splits.

PTB

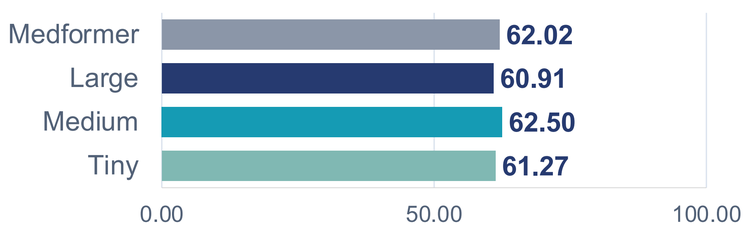
ECG · 2 diagnostic classes



Large: 91.72% accuracy
+8.22 points vs. Medformer

PTB-XL

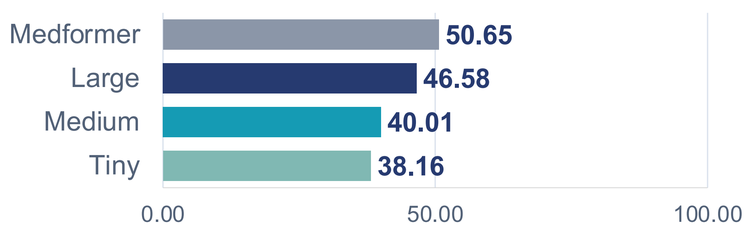
ECG · 5 diagnostic classes



Medium: highest F1, 62.50%
Large: highest accuracy, 74.16%

ADFTD

EEG · Alzheimer's / FTD / control



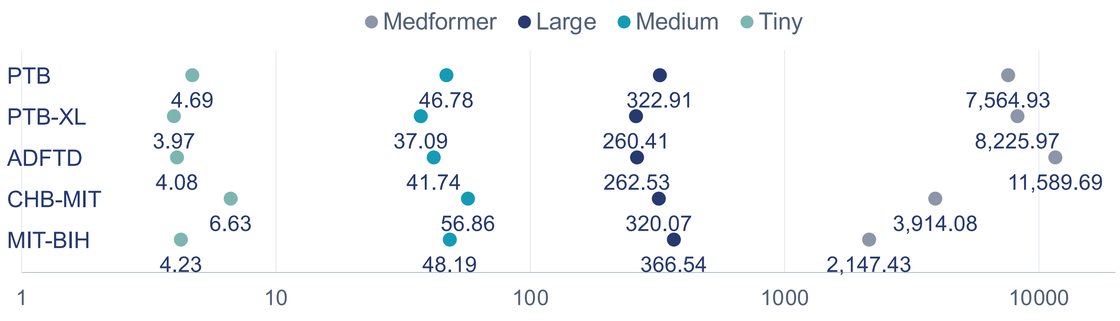
Large: 53.62% accuracy
Medformer retains the highest F1

6 Energy model and measured sparsity

$$E_{\mathrm{SpikeMamba}} = E_{\mathrm{MAC}}\,\mathrm{FL}_{\mathrm{Input+MLP}} + E_{\mathrm{AC}}\,R_{\mathrm{eff}}\,\mathrm{FL}_{\mathrm{Spike}}, \qquad R_{\mathrm{eff}} = RG$$

R: firing rate. G: gate activation ratio. FL: operation count.
45 nm CMOS: MAC 4.6 pJ, AC 0.9 pJ. Single pass (T=1).

Estimated energy by model size (mJ, log scale)

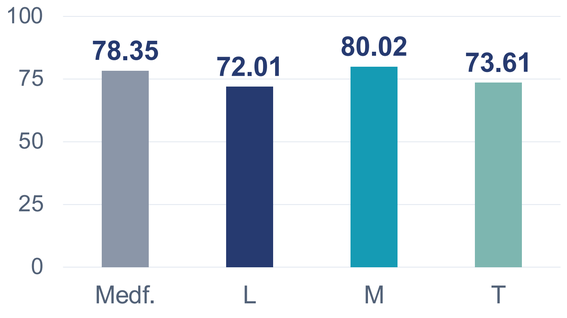


Across profiled variants, firing is 12.37–15.89%. The spike pathway accounts for 96.19–99.71% of operations. Dense input and MLP computation retains MAC cost.

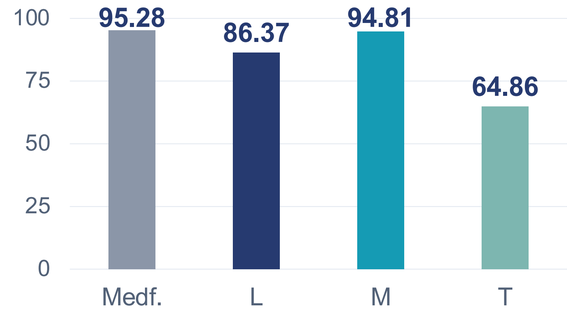
7 Seizure and arrhythmia benchmarks

F1 (%) across model sizes. L/M/T: SpikeMamba. Medf.: Medformer.

CHB-MIT F1 (%)



MIT-BIH F1 (%)



Medium AUROC: 94.48% on CHB-MIT and 99.79% on MIT-BIH.

Medium: +1.67 F1 pts / 68.8× lower energy **Medium: −0.47 F1 pts / 44.6× lower energy**

Both comparisons are relative to Medformer. Energy is an analytical estimate.
CHB-MIT: 5-s EEG, pre-ictal window ends 5 min before onset. MIT-BIH: 5 classes.
Segment-mixed 60/20/20 splits do not establish generalization to unseen subjects.
On ADFTD, higher accuracy coexists with lower F1. The benefit depends on the metric.

8 Model size and estimated energy on PTB

Large: 8 layers, D=512
8,632,106 parameters
322.91 mJ / F1 94.00%
D denotes hidden dimension.

Medium: 4 layers, D=256
1,111,958 parameters
46.78 mJ / F1 92.27%
6.9× lower energy / −1.73 F1 points

Tiny: 4 layers, D=64
75,254 parameters
4.69 mJ / F1 86.72%
68.9× lower energy / −7.28 F1 points

Scan for code

Conclusion Sparse spikes and selective state updates reduce estimated energy, with dataset-dependent F1 trade-offs. MindCore deployment will test on-chip energy and latency.

Contact: siyong.lee@stonybrook.edu / yoonseok.yang@sunykorea.ac.kr