Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sparse spiking recurrent networks for seizure-related EEG classification: competitive accuracy and noise tolerance under a leakage-safe protocol

Oliver Ashraf

oliver.gurgis@student.guc.edu.eg

German University in Cairo

Shady Nagy

German University in Cairo

Mohammed Salem

German University in Cairo

Article

Keywords:

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10437482/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Sparse spiking recurrent networks for seizure-related EEG classification: competitive accuracy and noise tolerance under a leakage-safe protocol

Oliver Ashraf[1,*], **Shady Ahmed Nagy**[1], **and Mohammed Abdel-Megeed M. Salem**[1]

[1]Faculty of Media Engineering and Technology, German University in Cairo, New Cairo, Egypt
[*]oliverashraf4@gmail.com

ABSTRACT

Wearable electroencephalography (EEG) systems require seizure-analysis models that remain reliable under noisy acquisition conditions while satisfying strict memory and energy constraints. This study compares three recurrent architectures for five-class seizure-related EEG classification on the Bonn EEG dataset: a conventional convolutional LSTM baseline, a spiking LSTM with temporal attention, and a hybrid Conv1D spiking LSTM with temporal attention. The spiking models convert temporal changes into sparse binary event streams and are trained with surrogate gradients. Evaluation is performed at the full-recording level to avoid leakage from overlapping EEG chunks. In five-fold stratified cross-validation, the baseline, pure spiking model, and hybrid spiking model achieved signal-level accuracies of 87.8%, 86.4%, and 86.6%, respectively. Pairwise McNemar tests found no significant differences, and the 95% confidence intervals of the paired differences bound any accuracy gap at 4.3 percentage points, confirming that the spiking models are competitive with the dense baseline. The spiking models achieve this with binary recurrent communication: the recurrent core of the hybrid model operated with 79.2% hidden-state sparsity, so most recurrent weight accesses can be skipped on event-driven hardware. The pure spiking model was also markedly more tolerant to additive white Gaussian noise, losing 2 percentage points at 20 dB signal-to-noise ratio compared with 26 points for the baseline. A transparent deployment analysis shows that the present spiking models are larger than the baseline, so realizing memory and energy gains on wearable hardware is the next step, through size-matched spiking models, low-precision weights, and event-driven execution. These findings establish that sparse spiking recurrent processing preserves seizure-related EEG classification performance while adding noise tolerance, a necessary foundation for low-power wearable seizure monitoring.

Introduction

Epilepsy is a chronic neurological disorder characterized by recurrent seizures caused by abnormal electrical activity in the brain. Electroencephalography (EEG) remains one of the most important tools for observing neural dynamics and supporting seizure-related analysis. Automated EEG classification has therefore become an active area of machine learning research, especially for long-term monitoring scenarios where manual review is costly and delayed intervention may carry clinical risk.

Deep recurrent models, particularly Long Short-Term Memory (LSTM) networks, are well suited to EEG because they can model temporal dependencies across sequential samples[1]. However, conventional LSTMs rely on dense multiply–accumulate operations and continuous-valued hidden states. These properties are acceptable in offline analysis but become restrictive for wearable or portable medical devices, where battery lifetime, memory footprint, and robustness to motion artifacts are central design constraints.

Spiking Neural Networks (SNNs) offer an alternative computational paradigm in which information is represented through sparse events rather than dense continuous activations. When implemented on event-driven or neuromorphic hardware, binary spikes can replace many dense multiply–accumulate operations with sparse accumulate operations. Modern surrogate-gradient methods make such threshold-based models trainable with backpropagation-like optimization despite the non-differentiability of spike generation[2]. These properties motivate spiking recurrent architectures for EEG analysis, where clinically relevant events may be encoded through temporal changes in the signal.

This work studies whether spiking recurrent models can offer deployment-relevant properties for seizure-related EEG classification without sacrificing useful predictive performance. Rather than claiming clean-data accuracy superiority, the central question is deployment-centered: whether spiking recurrent networks remain competitive with a standard Conv-LSTM baseline while providing sparse binary activity and improved tolerance to noise, and what their memory, latency, and energy profile looks like in software before hardware-specific optimization. Model 3 is a hybrid that combines convolutional feature extraction with spike-based recurrent processing, to test whether a continuous frontend recovers discriminative information that a purely event-based input representation may discard.

Beyond benchmark accuracy, the clinical motivation for this work is the need for seizure-analysis models that can operate directly on wearable EEG hardware. Many wearable or portable EEG systems still depend on transmitting EEG data to external devices or cloud-based servers for computationally intensive analysis. This dependency can introduce communication latency, increase wireless transmission energy, and drain battery life during continuous monitoring. A seizure-alert device is only clinically useful if it remains active when a seizure occurs, so battery depletion can directly increase the risk of a missed warning. On-device inference therefore requires models that fit small memory budgets and run at low energy. Spiking recurrent models address this through sparse event-driven computation, but their accuracy on seizure-related EEG under strict evaluation has not been established. This work establishes it, and reports a transparent deployment profile that identifies the concrete steps from software models to wearable hardware.

The main contributions are:

- a full-signal-level comparison of a conventional Conv-LSTM, a spiking LSTM with temporal attention, and a hybrid Conv1D spiking LSTM with temporal attention on the Bonn EEG dataset[3];
- formal documentation of the implemented spike encoding, spiking LSTM dynamics, temporal attention mechanism, and fast surrogate derivative;
- a leakage-safe five-fold stratified cross-validation protocol in which overlapping chunks are generated only after full EEG recordings are assigned to folds, with McNemar tests and confidence intervals for paired accuracy differences;
- an additive white Gaussian noise (AWGN) stress test at 20 dB, 10 dB, and 0 dB signal-to-noise ratios, which characterizes tolerance to stationary broadband noise; and
- a deployment analysis covering parameter memory at float32 and int8 precision, spike activity and sparsity, dense and spike-aware operation and energy estimates, and CPU inference latency, with an explicit separation between properties demonstrated in this study and properties that remain anticipated.

Related work

EEG seizure analysis and recurrent models

The Bonn EEG dataset introduced by Andrzejak et al. has been widely used as a benchmark for seizure-related EEG classification because it contains controlled five-class recordings from healthy subjects and epilepsy patients across different brain states[3]. The dataset is compact but experimentally flexible: it supports binary, grouped multi-class, and full five-class formulations, making it suitable for controlled comparisons between standard recurrent and spiking recurrent models.

EEG analysis is challenging because EEG signals are noisy, non-stationary, patient-specific, and temporally structured[4]. These characteristics make sequence modeling important in seizure-related classification tasks. LSTM networks address the vanishing-gradient limitations of standard recurrent neural networks by using gated memory states[1]. In EEG analysis, this makes them attractive for temporal pattern recognition, but their dense recurrence increases computational and memory cost. Prior EEG studies have shown that recurrent models can learn seizure-related temporal structure from raw or feature-based EEG sequences[5,6]. For this reason, a conventional LSTM-based model provides a strong baseline for assessing whether spiking recurrent alternatives can preserve useful classification performance.

Spiking recurrent models and surrogate-gradient training

SNNs reduce activity by communicating through discrete events. However, the hard threshold function used to generate spikes is non-differentiable, so modern SNN training commonly relies on surrogate-gradient learning[2]. Recurrent spiking architectures have been proposed to improve temporal memory in SNNs. Bellec et al. introduced long short-term memory spiking networks using adaptive spiking neurons to improve memory over temporal windows[7]. Lotfi Rezaabad and Vishwanath proposed spiking LSTM networks that preserve LSTM-style memory while replacing selected continuous operations with spike-compatible mechanisms[8]. Zhang et al. further showed that LSTM-SNN architectures can be applied to temporal classification tasks outside EEG, supporting the broader feasibility of spiking LSTM-style sequence models[9].

Deployment-oriented spiking recurrent research has also explored hybrid analog-spiking designs. Ponghiran and Roy proposed a hybrid analog-spiking LSTM for energy-efficient edge computing, while Datta et al. investigated accurate, energy-efficient, and low-latency spiking LSTMs[10,11]. At the hardware level, Gruber examined how recurrent LSTM computation can be reinterpreted for Loihi 2 neuromorphic execution, highlighting the importance of fixed-point constraints and hardware mapping[12]. These studies show that spiking recurrent models are promising for temporal learning, but they also highlight practical challenges: simulation overhead, sensitivity to spike rates, training instability, and the need for hardware-aware implementation.

ANN–SNN comparative methodology

Comparative evaluation between artificial neural networks and SNNs is necessary because efficiency claims are meaningful only when measured together with classification performance, latency, robustness, and deployment constraints. Souza de Abreu et al. compared LSTM and spiking models for temporal fault prediction and found that spiking approaches can remain competitive while showing robustness to sparse data and data drift[13]. Sharifian et al. emphasized that neuromorphic efficiency must be treated as a reliability-aware optimization problem rather than a simple replacement of ANN operations with SNN operations[14]. Aydin et al. similarly showed that hybrid ANN-SNN designs can reduce transient inefficiencies and improve low-power temporal inference[15]. These studies motivate the present work's shared evaluation protocol, in which conventional and spiking recurrent models are compared under identical full-signal splits and deployment metrics.

Spike encoding for temporal EEG signals

Continuous EEG signals must be converted into spike-compatible representations before they can be processed by spiking models. Threshold-based temporal encoders are attractive because they emit events when temporal changes exceed a chosen threshold, naturally suppressing near-stationary activity[16,17]. In EEG, this is useful because seizure-related information often appears as transient or rhythmic temporal change. Prior work on temporal spike encoding shows that threshold-triggered event representations can preserve informative signal variation while reducing unnecessary activity[16,18]. At the same time, encoding design remains sensitive to threshold choice: thresholds that are too low can saturate the spike stream with noise, while thresholds that are too high can suppress informative temporal structure.

MAC, AC, and hardware-aware efficiency

The deployment argument for SNNs is closely connected to the distinction between dense multiply–accumulate (MAC) operations and event-driven accumulate (AC) operations. Conventional LSTMs compute dense weighted sums at every timestep, whereas spike-driven layers can reduce effective operation count when activity remains sparse[10,15,19]. Reported hardware estimates show that AC operations can consume less energy than MAC operations under comparable technology assumptions[14,15,19]. However, lower per-operation cost does not automatically guarantee lower end-to-end energy. Practical gains depend on spike activity ratio, conversion overhead, hardware mapping, reliability constraints, and whether any dense analog components remain in the model[11,14]. This is especially important for the present hybrid Model 3, because its Conv1D frontend and analog residual path improve representational capacity but also preserve some dense computation.

Wearable edge-AI motivation and research gap

Energy efficiency is a central requirement in embedded, wearable, and edge-AI systems because such platforms must often process continuous sensory data under strict power and memory constraints[20]. Neuromorphic and spiking approaches are promising in this setting because sparse activity can concentrate computation on informative events rather than every sample. EEG-decoding work on neuromorphic hardware further supports the practical relevance of SNNs for biomedical signal processing[17]. Nevertheless, application-specific evaluation remains necessary before spiking recurrent models can replace conventional temporal baselines in seizure-related EEG analysis.

Recent Edge-AI and neuromorphic studies further emphasize that memory movement, always-on sensing, and event sparsity are central deployment concerns for embedded intelligence[21–24]. This supports evaluating spiking EEG models not only by clean-data accuracy, but also by memory footprint, latency, robustness, and hardware compatibility. The literature therefore leaves a clear research gap. Standard LSTMs remain strong for temporal EEG classification, and spiking recurrent models offer plausible memory and energy advantages, but relatively few studies directly compare standard and spiking LSTM-style architectures for seizure-related EEG under a shared leakage-safe protocol. This study addresses that gap by comparing a conventional Conv-LSTM baseline, a pure spiking LSTM with temporal attention, and a hybrid Conv1D spiking LSTM with temporal attention using signal-level cross-validation, noise stress tests, memory, latency, and energy profiling, and McNemar significance testing.

Materials and methods

Dataset

The experiments use the Bonn EEG dataset[3]. The dataset contains 500 single-channel EEG recordings divided evenly across five physiological classes. Each recording contains 4097 samples acquired at approximately 173.61 Hz, corresponding to about 23.6 s per signal. The five implementation classes are denoted Z, O, N, F, and S and represent healthy, interictal, and ictal brain states. Sets Z and O were recorded from five healthy volunteers using surface electrodes, whereas sets N, F, and S were recorded intracranially from five patients undergoing presurgical evaluation[3]. All segments were cut from continuous multichannel recordings after visual inspection for artefacts, and no subject identifiers are provided. These properties matter for interpreting the results and are discussed under Limitations. All reported metrics are computed using the same class ordering across models.

Table 1 summarizes the dataset characteristics used in this study. The dataset is balanced, with 100 recordings per class, so the five-class classification task is not affected by class imbalance.

Table 1. Summary of the Bonn EEG dataset used in this study.

Class	Condition	Recordings	Samples/recording	Frequency	Duration
Z	Healthy - Eyes Open	100	4097	173.61 Hz	23.6 s
O	Healthy - Eyes Closed	100	4097	173.61 Hz	23.6 s
N	Interictal - Opposite Hemisphere	100	4097	173.61 Hz	23.6 s
F	Interictal - Epileptogenic Zone	100	4097	173.61 Hz	23.6 s
S	Ictal - Seizure Activity	100	4097	173.61 Hz	23.6 s

To make the raw signal morphology visible before segmentation and model training, Figure 1 shows representative 5-second EEG snippets from all five classes using the same time scale. The healthy and interictal classes generally show lower-amplitude oscillatory activity, whereas the ictal seizure class contains larger and more abrupt amplitude changes. These visible temporal differences motivate recurrent sequence modeling and spike-based encoders that respond to rapid signal variation.



Figure 1. Representative 5-second raw EEG snippets from the five Bonn EEG classes: healthy eyes open, healthy eyes closed, interictal opposite hemisphere, interictal epileptogenic zone, and ictal seizure activity.

Figure 2 provides a complementary statistical view of the raw recordings by comparing the peak-to-peak amplitude distribution across classes. The ictal seizure class exhibits a substantially larger amplitude range than the healthy and interictal classes, highlighting an important dataset-level difference that is visible in the raw signals and relevant for seizure-related classification.

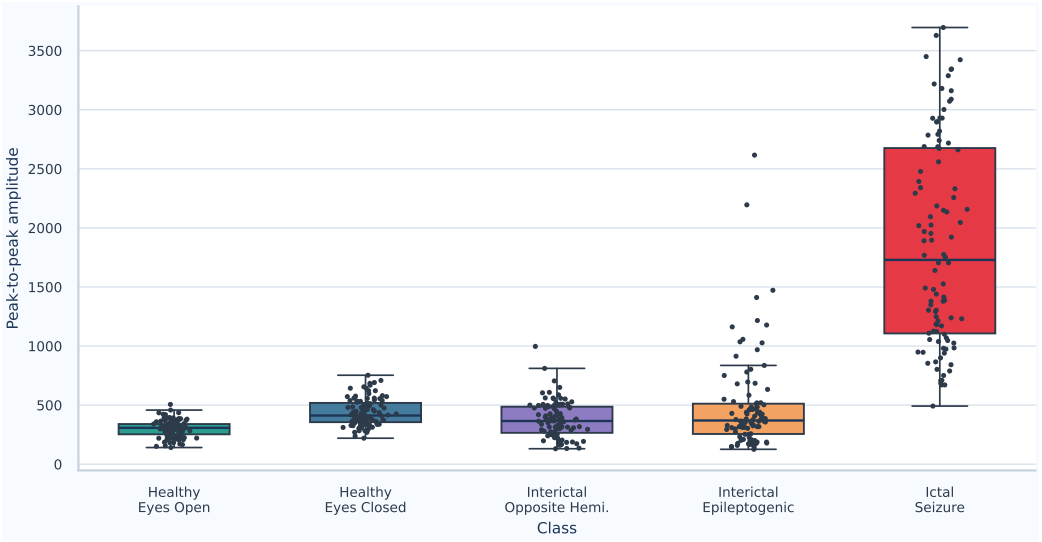

Figure 2. Distribution of peak-to-peak amplitude across the five Bonn EEG classes. The ictal seizure recordings show larger amplitude range than the healthy and interictal recordings.

Full-signal splitting and segmentation

All train-validation-test and cross-validation splits are performed at the full-recording level before chunking. This avoids chunk-level leakage because overlapping windows from the same EEG recording cannot appear simultaneously in training and test partitions. For the primary hold-out experiments, a stratified 70/10/20 split is used. Each recording is then segmented into overlapping windows of 256 samples with stride 128, yielding 31 chunks per recording.

For the final statistical evaluation, five-fold stratified cross-validation is performed over the 500 full EEG recordings. Each fold contains approximately 20 recordings per class, corresponding to approximately 100 held-out full recordings per fold. Models are trained on four folds and evaluated on the held-out fold. The 500 held-out signal-level predictions are then aggregated for metric reporting and McNemar testing. This protocol ensures that evaluation is performed at the signal level rather than the chunk level.

Delta-sigma spike encoding

For the pure spiking model, each normalized EEG window $X = \{x_1, \ldots, x_T\}$ with $T = 256$ is converted into positive and negative spike channels using delta-sigma encoding. The temporal difference and residual update are

$$\Delta x_t = x_t - x_{t-1}, \qquad r_t = r_{t-1} + \Delta x_t. \tag{1}$$

The binary positive and negative spikes are

$$s_t^+ = \Theta(r_t - \theta_{\text{in}}), \qquad s_t^- = \Theta(-r_t - \theta_{\text{in}}), \tag{2}$$

where $\Theta(\cdot)$ is the Heaviside step function and the input spike threshold is explicitly set to

$$\theta_{\text{in}} = 0.18. \tag{3}$$

The residual is updated according to

$$r_t \leftarrow r_t - \theta_{\text{in}} s_t^+ + \theta_{\text{in}} s_t^-, \tag{4}$$

and the spike vector is

$$s_t = [s_t^+, s_t^-] \in \{0,1\}^2. \tag{5}$$

Model 1: Conventional Conv-LSTM baseline

The conventional baseline is designed as a dense, non-spiking reference model. It receives each normalized EEG chunk as a continuous sequence $X \in \mathbb{R}^{256\times 1}$. Two Conv1D blocks extract local temporal features before two standard LSTM layers

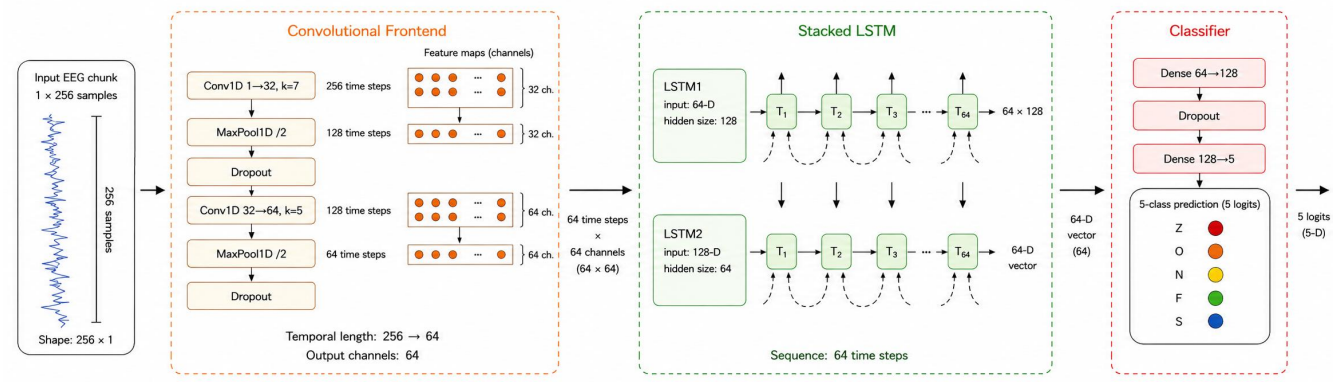

Figure 3. Architecture of Model 1, the conventional Conv-LSTM baseline. The model applies Conv1D feature extraction to the 256-sample EEG chunk, reduces the temporal length to 64 steps, processes the resulting feature sequence using stacked LSTM layers, and maps the final representation to five output logits.

perform recurrent modeling:

$$C^{(1)} = \mathrm{MaxPool}(\mathrm{ReLU}(W^{(1)}_{\mathrm{conv}} * X + b^{(1)}_{\mathrm{conv}})), \tag{6}$$

$$C^{(2)} = \mathrm{MaxPool}(\mathrm{ReLU}(W^{(2)}_{\mathrm{conv}} * C^{(1)} + b^{(2)}_{\mathrm{conv}})). \tag{7}$$

The first convolution maps 1 input channel to 32 filters with kernel size 7. The second convolution maps 32 channels to 64 filters with kernel size 5. Max pooling reduces the temporal length from 256 to 64. The two LSTM hidden sizes are 128 and 64, followed by a dense layer of size 128 and a five-unit output layer for the five classes.

For a conventional LSTM layer,

$$\begin{bmatrix} i_t \\ f_t \\ \tilde{c}_t \\ o_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \tanh \\ \sigma \end{bmatrix} (W_x z_t + W_h h_{t-1} + b), \tag{8}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{9}$$

$$h_t = o_t \odot \tanh(c_t). \tag{10}$$

The final representation is passed through the dense layer and classifier to produce five logits.

Model 2: Spiking LSTM with temporal attention

The pure spiking model uses the two-channel spike train as input. A non-convolutional event frontend compresses every four samples into a 64-step representation. For reduced time index τ, the four-sample spike-rate summary is

$$\rho^{(4)}_\tau = \frac{1}{4} \sum_{t \in W_\tau} s_t. \tag{11}$$

The event frontend produces a 10-dimensional event vector. Event branches detect any, repeated, burst, early, and late activity using thresholded spike-rate summaries. The any, repeated, and burst branches are defined as

$$e^{\mathrm{any}}_\tau = \Theta(\rho^{(4)}_\tau - 0.25), \tag{12}$$

$$e^{\mathrm{rep}}_\tau = \Theta(\rho^{(4)}_\tau - 0.50), \tag{13}$$

$$e^{\mathrm{burst}}_\tau = \Theta(\rho^{(4)}_\tau - 0.75). \tag{14}$$

The 10-dimensional event vector is projected into a 128-dimensional binary representation:

$$z_\tau = \Theta(\mathrm{LN}(W_e e_\tau + b_e) - \theta_p), \qquad \theta_p = 0.20. \tag{15}$$

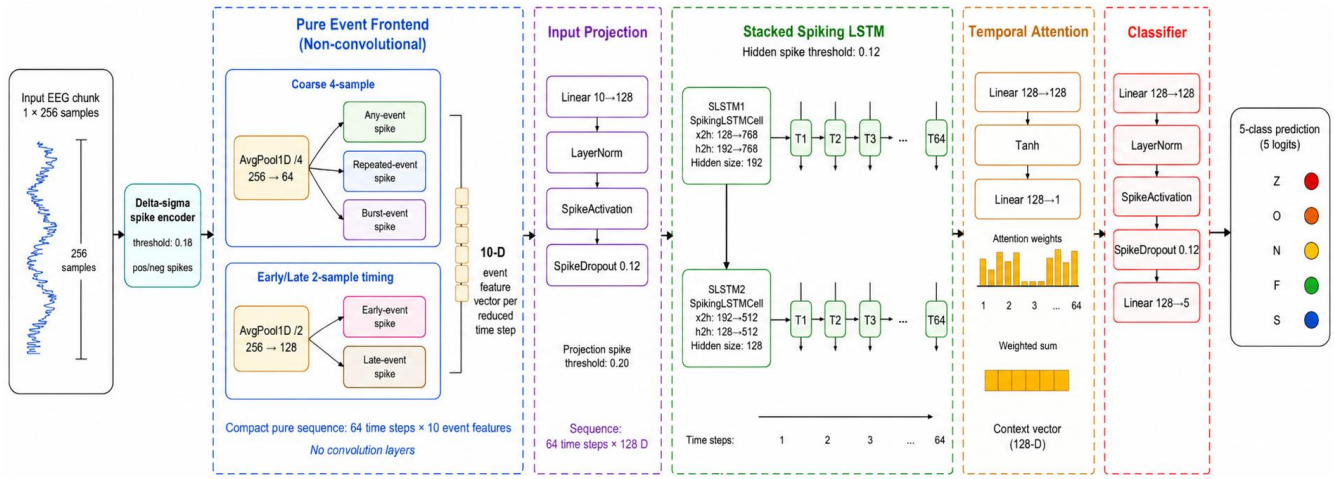

Figure 4. Architecture of Model 2, the pure S-LSTM with temporal attention. The model uses delta-sigma spike encoding, a non-convolutional event frontend, input projection, stacked spiking LSTM layers, temporal attention, and a spiking classifier.

The recurrent core uses two stacked spiking LSTM cells with hidden sizes 192 and 128. The hidden spike threshold is explicitly set to $\theta_h = 0.12$. For a generic spiking LSTM cell,

$$a_t = W_x z_t + W_h h_{t-1} + b, \tag{16}$$

$$\begin{bmatrix} i_t \\ f_t \\ g_t \\ o_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \tanh \\ \sigma \end{bmatrix} (a_t), \tag{17}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t, \tag{18}$$

$$u_t = o_t \odot \tanh(c_t), \tag{19}$$

$$h_t = \Theta(u_t - \theta_h), \qquad \theta_h = 0.12. \tag{20}$$

Temporal attention is applied to the output sequence of the second spiking LSTM layer:

$$q_t = v_a^T \tanh(W_a h_t^{(2)} + b_a), \tag{21}$$

$$\alpha_t = \frac{\exp(q_t)}{\sum_{k=1}^{64} \exp(q_k)}, \tag{22}$$

$$z_{\text{att}} = \sum_{t=1}^{64} \alpha_t h_t^{(2)}. \tag{23}$$

The attention-pooled representation is passed through a spiking classifier to produce five logits.

Model 3: Hybrid Conv1D spiking LSTM with temporal attention

Model 3 is a hybrid architecture that combines a compact continuous Conv1D frontend with spike-based recurrent processing. The convolutional frontend is intended to extract discriminative local temporal features before spike conversion, while the spiking recurrent core restricts recurrent communication to binary events.

The model first applies a continuous convolutional frontend:

$$F^{(1)} = \text{MaxPool}(\text{GELU}(\text{BN}(W_{\text{conv}}^{(1)} * X + b_{\text{conv}}^{(1)}))), \tag{24}$$

$$F^{(2)} = \text{MaxPool}(\text{GELU}(\text{BN}(W_{\text{conv}}^{(2)} * F^{(1)} + b_{\text{conv}}^{(2)}))). \tag{25}$$

The first convolution maps 1 input channel to 48 filters with kernel size 7. The second convolution maps 48 channels to 96 filters with kernel size 5. The output sequence length is reduced from 256 to 64. The 96-channel feature sequence is then projected to 112 features:

$$y_t = \text{GELU}(\text{LN}(W_f F_t^{(2)} + b_f)), \qquad y_t \in \mathbb{R}^{112}. \tag{26}$$

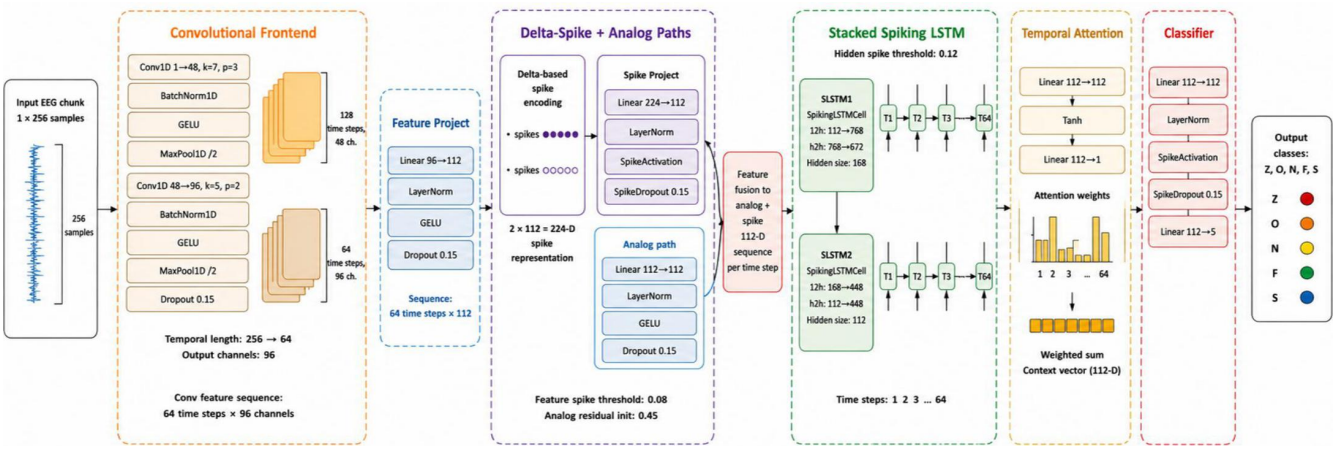

Figure 5. Architecture of Model 3, the hybrid Conv1D S-LSTM with temporal attention. The model combines a compact Conv1D frontend with delta-spike and analog feature paths before stacked spiking LSTM layers, temporal attention, and classification.

Feature-level delta spikes are generated by

$$\Delta y_t = y_t - y_{t-1}, \qquad \Delta y_1 = 0, \tag{27}$$

$$s_t^+ = \Theta(\Delta y_t - \theta_f), \qquad s_t^- = \Theta(-\Delta y_t - \theta_f), \tag{28}$$

with feature spike threshold

$$\theta_f = 0.08 \tag{29}$$

and spike vector

$$s_t = [s_t^+, s_t^-] \in \{0,1\}^{224}. \tag{30}$$

The spike and analog residual paths are fused as

$$z_t^s = \Theta(\mathrm{LN}(W_s s_t + b_s) - \theta_f), \tag{31}$$

$$z_t^a = \mathrm{GELU}(\mathrm{LN}(W_a y_t + b_a)), \tag{32}$$

$$z_t = z_t^s + \lambda z_t^a, \tag{33}$$

where the analog residual coefficient λ is learnable and initialized to

$$\lambda = 0.45. \tag{34}$$

The fused representation is processed by two spiking LSTM layers with hidden sizes 168 and 112. The hidden spike threshold is set to $\theta_h = 0.12$, matching Model 2. The output is passed through the same temporal attention mechanism and a classifier to produce five logits.

Surrogate-gradient training

Forward spike generation uses the hard threshold function

$$\Theta(u-\theta) = \begin{cases} 1, & u \geq \theta, \\ 0, & u < \theta. \end{cases} \tag{35}$$

Since this function is non-differentiable, the implementation uses a fast-sigmoid surrogate derivative during backpropagation, following SuperSpike-style surrogate-gradient learning[25]:

$$\frac{\partial \Theta(u-\theta)}{\partial u} \approx \frac{1}{(\alpha|u-\theta|+1)^2}, \qquad \alpha = 5.0. \tag{36}$$

This is the exact fast surrogate form used in the implementation and is distinct from an arctangent surrogate.

Training and evaluation

The conventional baseline is trained in TensorFlow/Keras using Adam, sparse categorical cross-entropy, batch size 32, and 60 epochs. The spiking models are trained in PyTorch using AdamW, label smoothing of 0.02, batch size 64, learning rate 5×10^{-4}, and weight decay 5×10^{-4}. Model 2 uses random temporal shifts and spike dropout. Model 3 uses random temporal shifts and additive training noise. The baseline was trained without noise augmentation; this difference is relevant to the noise analysis and is discussed there.

All training experiments were executed on a Kaggle GPU environment using two NVIDIA Tesla T4 GPUs, each with 13,757 MB available memory and CUDA compute capability 7.5. The conventional baseline was implemented in TensorFlow/Keras, while the spiking models were implemented in PyTorch.

For signal-level evaluation, class probabilities are averaged over all chunks from the same original EEG recording, and the class with the highest average probability is assigned to the signal. CPU inference latency was measured on an Intel(R) Xeon(R) CPU @ 2.00GHz using the same batched signal-level evaluation protocol for all models, with inference batch size 64. Metrics include accuracy, macro precision, macro recall, macro F1-score, and pairwise exact McNemar tests on binary correctness outputs. Because a non-significant McNemar test does not establish equivalence, each pairwise comparison is also reported as the paired accuracy difference $\hat{d} = (b-c)/n$ with a 95% Wald confidence interval, where b and c are the discordant counts and $n = 500$, using $\widehat{\mathrm{Var}}(\hat{d}) = \left[(b+c) - (b-c)^2/n\right]/n^2$.

Deployment metrics

Parameter memory is computed from trainable parameters under float32 (4 bytes per parameter) and int8 (1 byte per parameter) storage, so that all models are compared at equal weight precision. The int8 values are storage estimates; accuracy after int8 quantization was not evaluated in this study.

Binary spikes and parameter memory must be distinguished. In Models 2 and 3 the hidden outputs h_t and the event inputs are single bits, which reduces activation and recurrent-state traffic, but the synaptic weights that multiply these spikes remain multi-bit values, and the cell state c_t in Eq. (18) remains continuous. Spiking therefore does not by itself reduce weight storage. Storing every parameter of Model 3 at one bit would occupy 50.6 KB, but this would require binarized weights, including those of the continuous Conv1D frontend and analog residual path, which were neither trained nor evaluated here. We report this value only as a theoretical lower bound and not as a property of the evaluated model.

For dense layers, the operation estimate is

$$\mathrm{MAC}_{\mathrm{dense}} = N_{\mathrm{in}} N_{\mathrm{out}}. \tag{37}$$

For spiking layers with activity ratio ρ, the approximate event-driven accumulate count is

$$\mathrm{AC}_{\mathrm{spike}} \approx \rho N_{\mathrm{in}} N_{\mathrm{out}}, \qquad \frac{\mathrm{AC}_{\mathrm{spike}}}{\mathrm{MAC}_{\mathrm{dense}}} \approx \rho, \tag{38}$$

with sparsity defined as $1-\rho$. This spike-aware estimate applies most directly to the spike projection and recurrent transformations. Model 3 remains partly dense because its Conv1D frontend and analog residual branch use continuous operations.

Results

Clean-data signal-level performance

Table 2 reports the five-fold full-signal cross-validation results as mean ± standard deviation across folds. The conventional Conv-LSTM achieved the highest mean accuracy, but the differences were small: the pure spiking model and the hybrid spiking model were 1.4 and 1.2 percentage points below the baseline, respectively.

Table 2. Five-fold full-signal cross-validation results reported as mean ± standard deviation across folds.

Model	Accuracy	Macro precision	Macro recall	Macro F1
Conv-LSTM	87.80% ± 5.22%	88.38% ± 5.95%	87.80% ± 5.22%	87.74% ± 5.19%
S-LSTM + Attn	86.40% ± 5.41%	86.25% ± 5.52%	86.40% ± 5.41%	86.24% ± 5.62%
Conv1D + S-LSTM + Attn	86.60% ± 5.50%	86.65% ± 5.55%	86.60% ± 5.50%	86.61% ± 5.61%

Pairwise McNemar tests were computed from the aggregated held-out correctness vectors. As shown in Table 3, none of the pairwise differences reached the $p < 0.05$ significance threshold. A non-significant test does not by itself establish equivalence, so Table 3 also reports the paired accuracy difference with its 95% confidence interval. Relative to the baseline, the intervals exclude an accuracy deficit larger than 4.3 percentage points for either spiking model, but they do not support equivalence within a narrower margin. In this study, "competitive" accuracy therefore means a deficit that, with 95% confidence, does not exceed about 4 percentage points on this benchmark.

Table 3. Pairwise exact McNemar tests on aggregated five-fold signal-level predictions ($n = 500$). The difference is the accuracy of model A minus model B, with a 95% Wald confidence interval for paired proportions.

Comparison (A vs B)	A correct/B wrong	A wrong/B correct	Difference (pp)	95% CI (pp)	p-value
Conv-LSTM vs S-LSTM	30	23	$+1.4$	$[-1.5, +4.3]$	0.410
Conv-LSTM vs Conv-S-LSTM	32	26	$+1.2$	$[-1.8, +4.2]$	0.512
S-LSTM vs Conv-S-LSTM	30	31	-0.2	$[-3.3, +2.9]$	1.000

Signal-level confusion matrix analysis

Although macro-averaged metrics summarize the overall five-class performance, they do not show which EEG states are most frequently confused. Therefore, signal-level confusion matrices were generated for the three evaluated models, as shown in Figure 6. The matrices are computed on the stratified hold-out test split (100 recordings, 20 per class) rather than on the cross-validation folds, and they are used to describe error structure, not to rank the models. The overall accuracies on this split (91%, 90%, and 95%) are higher than the cross-validated estimates in Table 2, differing from them by 3.2, 3.6, and 8.4 percentage points, or roughly 0.6 to 1.5 fold standard deviations. A single 100-recording split is therefore a noisy estimate, and the cross-validation results remain the primary estimate of generalization.

All three models classify every ictal recording (class S) correctly on this split. This result should be interpreted with care: class S has a much larger amplitude range than the other classes (Figure 2), and it was recorded intracranially, so its separability may partly reflect amplitude scale and recording conditions rather than seizure morphology alone. Most remaining errors occur between physiologically related non-seizure classes, especially between healthy eyes-open and eyes-closed recordings, or between the two interictal groups.

Model 1 mainly confuses healthy eyes-closed recordings O with eyes-open recordings Z, and interictal epileptogenic recordings F with opposite-hemisphere recordings N. Model 2 shows a similar pattern with additional errors for class N, where some N recordings are predicted as O, F, or S. Model 3 made the fewest errors on this split (5, compared with 9 for Model 1 and 10 for Model 2), with fewer Z/O and N/F confusions. With 20 recordings per class, however, these differences correspond to one to four recordings per class, were not tested for significance, and are not reflected in the cross-validated accuracies, where Models 2 and 3 differ by 0.2 percentage points.

Importantly, the errors observed in the confusion matrices do not suggest random failure across the five classes. Instead, the misclassifications are concentrated in classes that are physiologically closer or less visually separable in the raw EEG morphology. This supports the use of signal-level confusion matrices alongside macro-averaged metrics, because two models with similar overall accuracy can still differ in how they distribute their errors across seizure, healthy, and interictal states. The class-wise analysis therefore provides a more detailed view of model behavior than the aggregate scores alone.

The error distribution also indicates where event encoding may affect classification. On this split, the pure spiking model preserves seizure recognition but shows more dispersion in class N, which is consistent with event encoding discarding some fine-grained amplitude information. The hybrid frontend was designed to recover such information before spike conversion; whether it does so reliably cannot be established from a single 100-recording split and would require an ablation evaluated under cross-validation.

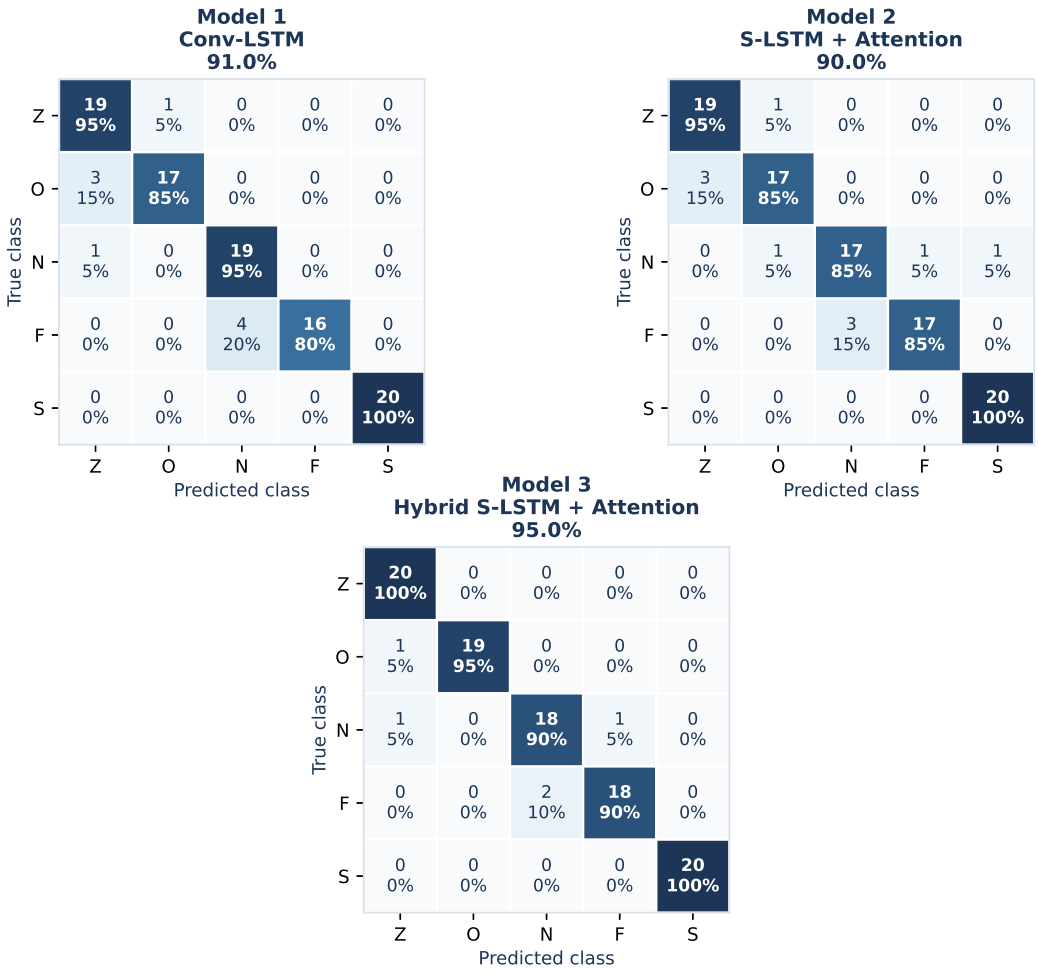

Figure 6. Signal-level confusion matrices for the three evaluated EEG classifiers on the stratified hold-out test split (100 recordings, 20 per class). Model 1 is the conventional Conv-LSTM baseline, Model 2 is the pure S-LSTM with temporal attention, and Model 3 is the hybrid Conv1D S-LSTM with temporal attention. The classes correspond to Z: healthy eyes open, O: healthy eyes closed, N: interictal opposite hemisphere, F: interictal epileptogenic region, and S: ictal seizure. Each cell reports the number of signals and the within-class percentage.

Training and validation dynamics

Training and validation curves were examined to verify that the models converged and to assess overfitting, particularly for the larger spiking models. Figures 7–9 show the learning dynamics for the conventional Conv-LSTM baseline, the pure S-LSTM with temporal attention, and the hybrid Conv1D S-LSTM with temporal attention.

Across the three models, the learning curves show rapid early improvement followed by a plateau, indicating that optimization converged. The degree of generalization separation differs between models and is described below. Note that the validation curves for Model 1 are computed at the chunk level, whereas those for Models 2 and 3 are computed at the signal level, so absolute validation values are not directly comparable across figures.

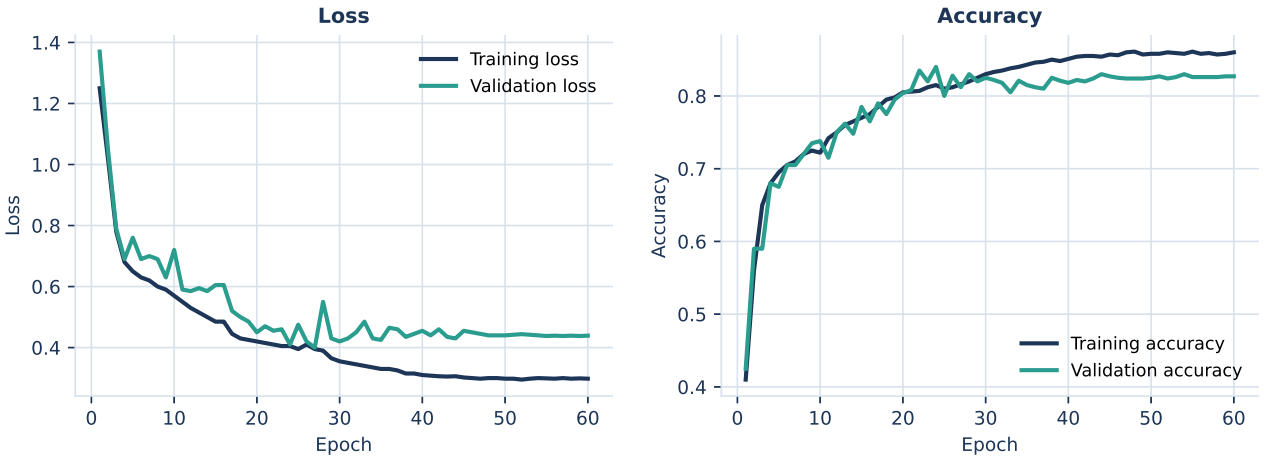

Figure 7. Training and validation dynamics for Model 1, the conventional Conv-LSTM baseline. The left panel shows training and validation loss across epochs, while the right panel shows training and validation accuracy.

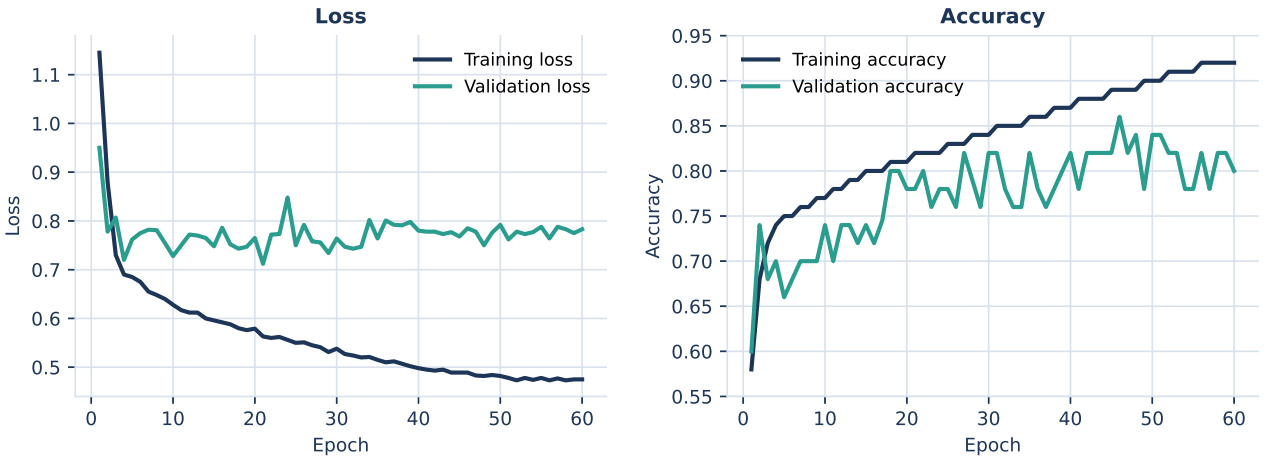

Figure 8. Training and validation dynamics for Model 2, the pure S-LSTM with temporal attention. The left panel shows training and signal-level validation loss across epochs, while the right panel shows training accuracy and signal-level validation accuracy.

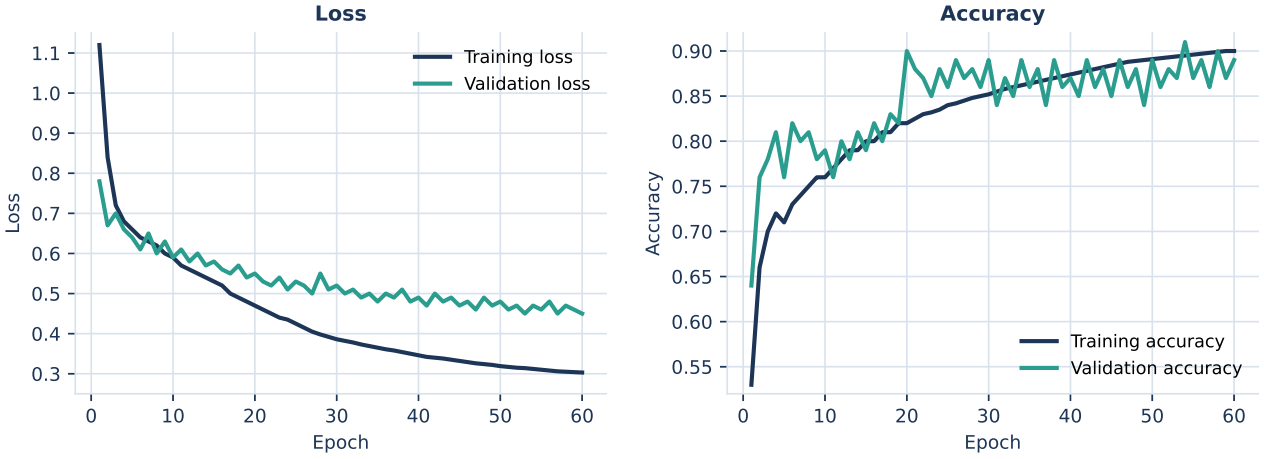

Figure 9. Training and validation dynamics for Model 3, the hybrid Conv1D S-LSTM with temporal attention. The left panel shows training and signal-level validation loss across epochs, while the right panel shows training and signal-level validation accuracy.

Model 1 improves rapidly during the early epochs and then refines slowly; after about epoch 30, training loss continues to decrease (to about 0.30) while validation loss settles near 0.44, and validation accuracy plateaus slightly below training accuracy, indicating moderate generalization separation. Model 2 shows clear overfitting: its signal-level validation loss stops improving after roughly the first five epochs and remains near 0.78, while training loss continues to fall to about 0.47 and training accuracy rises to about 0.92 against a validation accuracy that fluctuates between roughly 0.78 and 0.82. Validation accuracy does not degrade over this period, so the selected checkpoint is not strongly affected, but most of the later training epochs fit the training chunks without improving held-out performance; earlier stopping or stronger regularization would be appropriate for this model. Model 3 shows the closest agreement between curves: validation accuracy tracks training accuracy throughout, and validation loss decreases and stabilizes near 0.45 while training loss reaches about 0.30, indicating mild generalization separation rather than severe overfitting.

Interpretability through temporal attention

The temporal attention mechanism provides an interpretable view of which reduced time steps contribute most strongly to the final decision. Figure 10 shows a seizure-class recording with the raw EEG waveform, delta-encoded spike activity, and attention weights. The attention response concentrates around high-activity epileptiform regions, supporting the interpretation that the model is not relying only on global signal statistics but is emphasizing temporally localized seizure-related structure.

The alignment pattern shows that the attention mechanism assigns non-uniform importance across the EEG recording rather than treating all reduced time steps equally. Higher attention values tend to occur near regions with stronger delta-spike activity, suggesting that the model emphasizes temporally localized changes instead of relying only on global amplitude statistics. This behavior supports the use of temporal attention as an interpretability mechanism for the hybrid spiking model, because it links the final classification decision to specific time regions in the input signal.

At the same time, the attention map should be interpreted as a model-level explanation rather than a clinical localization tool. The figure indicates which reduced time steps were most influential for the trained classifier, but it does not replace expert EEG annotation or patient-specific seizure localization. Its main role in this study is to illustrate that the attention mechanism focuses on temporally meaningful EEG structure during seizure-class prediction. Because Figure 10 shows a single correctly classified recording, it is a qualitative illustration; a quantitative analysis of attention alignment across recordings and classes was not performed.

This interpretation is particularly relevant for Model 3 because the architecture combines a continuous Conv1D frontend with spike-based temporal processing. If the attention weights were uniformly distributed or unrelated to spike activity, the attention layer would provide limited explanatory value. Instead, the observed non-uniform attention profile suggests that the classifier uses the recurrent sequence representation selectively, assigning greater importance to specific time regions after feature extraction and spike-based processing. This supports the claim that temporal attention contributes to interpretability rather than acting only as an additional dense layer.

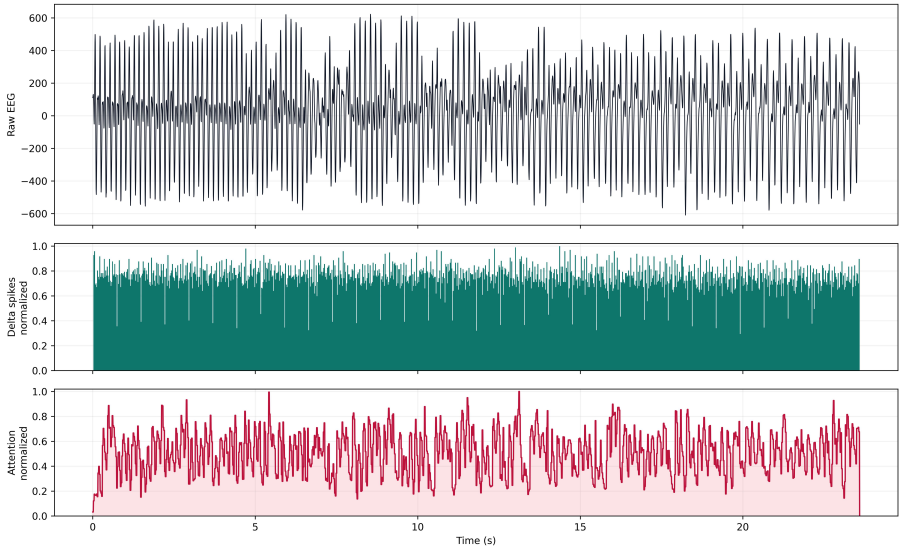

Figure 10. Temporal attention alignment example for a correctly classified ictal seizure recording using Model 3. The top panel shows the raw EEG waveform, the middle panel shows normalized delta-spike activity, and the bottom panel shows normalized temporal attention weights. Higher attention values indicate reduced time steps that contributed more strongly to the final classification decision.

Noise robustness

Additive white Gaussian noise was added to the recordings of the hold-out test split at the stated signal-to-noise ratios, and the trained models were evaluated without retraining. Table 4 includes the clean accuracy on the same split, so that each model's degradation can be read directly.

Table 4. Signal-level accuracy under additive white Gaussian noise on the hold-out test split (100 recordings). The clean column is the accuracy on the same split without added noise.

Model	Clean	20 dB	10 dB	0 dB
Conv-LSTM	91%	65%	30%	21%
S-LSTM + Attn	90%	88%	62%	35%
Conv1D + S-LSTM + Attn	95%	70%	41%	27%

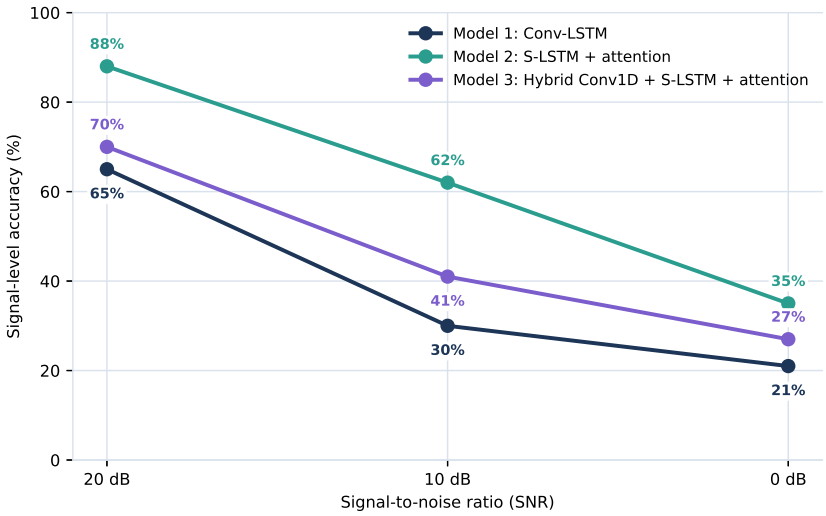

Figure 11. AWGN noise robustness comparison for the three evaluated models. The figure shows signal-level classification accuracy under additive white Gaussian noise at signal-to-noise ratios of 20 dB, 10 dB, and 0 dB. Model 2 shows the smallest degradation at every tested noise level. Model 3 retains slightly higher absolute accuracy than the baseline, but its loss relative to its own clean accuracy is similar.

Table 4 and Figure 11 show that the pure S-LSTM is the most noise-tolerant model at all tested SNR levels. At 20 dB it loses 2 percentage points relative to its clean accuracy, compared with 26 points for the baseline and 25 points for Model 3. This is consistent with the thresholded spike encoder suppressing small fluctuations and responding only to temporal changes that exceed the input threshold. Model 3 achieves higher absolute accuracy than the baseline at every noise level, while its degradation relative to its own clean accuracy is similar to that of the baseline. Because Model 3 was trained with additive noise augmentation and the baseline was not, we attribute the clearest noise-tolerance effect to the pure event-based input pathway of Model 2. This contrast is informative: the Conv1D frontend and analog residual path of Model 3 pass continuous-valued information, including noise, into the recurrent core, whereas the thresholded encoder of Model 2 filters it before recurrent processing. The two spiking designs therefore occupy complementary points in the design space, with Model 2 favoring noise tolerance and Model 3 favoring clean-data class separation on the hold-out split.

These results characterize tolerance to stationary broadband noise only. The limitations of this stress test with respect to realistic wearable EEG artefacts are discussed under Limitations.

Memory and latency deployment profile

Table 5 reports the deployment profile. CPU inference testing was performed on an Intel(R) Xeon(R) CPU @ 2.00GHz using the same batched signal-level evaluation protocol for all models, with inference batch size 64. In the current configuration, the spiking models have wider recurrent layers (hidden sizes 192/128 and 168/112, compared with 128/64 for the baseline) and additional projection layers, and therefore 2.47–2.66 times as many parameters. Their memory profile is thus set by layer width rather than by spiking: the spiking mechanism reduces activation and recurrent-state traffic, while weight storage depends on parameter count and precision, as explained under Deployment metrics. Size-matching the spiking models to the baseline is a direct target for future work and is discussed below.

Table 5. Deployment profile. Parameter memory is given at float32 and int8 precision (int8 values are storage estimates; accuracy after int8 quantization was not evaluated). CPU latency was measured on an Intel(R) Xeon(R) CPU @ 2.00GHz using the same batched signal-level evaluation protocol for all models, with inference batch size 64. The baseline runs in TensorFlow/Keras and the spiking models in PyTorch, so latency differences include framework overhead.

Model	Parameters	Float32 KB	Int8 KB	Relative size	CPU signal latency
Conv-LSTM	167,749	655.3	163.8	1.00×	5.28 ms
S-LSTM + Attn	446,598	1744.5	436.1	2.66×	39.62 ms
Conv1D + S-LSTM + Attn	414,135	1617.7	404.4	2.47×	41.12 ms

The latency values in Table 5 should be interpreted as CPU software latency rather than neuromorphic latency. In the present implementation, inference is executed on a conventional Von Neumann CPU using sequential batch processing. This execution mode penalizes spiking recurrent models because spike states must be simulated timestep by timestep in software, and sparse event updates are still managed through dense tensor operations and Python/PyTorch runtime overhead. Therefore, the measured CPU latency of the hybrid Model 3, 41.12 ms per signal under the batched evaluation protocol, reflects the cost of simulating an SNN on standard hardware rather than the latency expected from dedicated event-driven neuromorphic execution.

On neuromorphic processors such as Intel Loihi 2 or SynSense Speck, latency would be estimated differently. Instead of sequentially simulating all recurrent timesteps on a CPU, spikes are routed asynchronously as events through local memory and parallel neuron cores. In such a setting, the relevant latency depends on the hardware tick duration, event-routing delay, neuron-update schedule, and the number of reduced timesteps processed by the model. For the proposed spiking models, the recurrent sequence contains 64 reduced timesteps; therefore, an approximate neuromorphic latency model can be written as

$$T_{\text{neuro}} \approx 64\,\tau_{\text{tick}} + T_{\text{routing}}. \tag{39}$$

Here, τ_{tick} is the hardware-specific tick duration and T_{routing} accounts for event communication and scheduling overhead. Since the models were not physically deployed on Loihi 2, Speck, or another neuromorphic chip in this study, no measured neuromorphic latency is claimed. The CPU latency is reported for software reproducibility. Spike sparsity and the event-driven formulation indicate where a latency benefit could arise on such hardware, but no such benefit is demonstrated here.

Empirical spike activity was measured during Model 3 training to quantify the sparsity of the spiking recurrent core. At the best validation checkpoint, the average input activity was $\rho_{\text{in}} = 0.3161$, while the average hidden spike activity was $\rho_{\text{hid}} = 0.2077$. Thus, the recurrent spiking core operated with an empirical hidden-state sparsity of approximately $1 - \rho_{\text{hid}} = 79.2\%$. Across the final training plateau, the hidden spike activity remained close to this value, indicating that the spike-aware operation estimates are based on a stable sparse recurrent regime rather than a transient early-training state.

To convert operation counts into concrete theoretical switching energy, standard 45 nm CMOS constants were applied. Following commonly used hardware estimates, a 32-bit floating-point multiply–accumulate (MAC) operation was assigned an energy cost of $E_{\text{MAC}} = 4.6$ pJ, while an accumulate (AC) operation was assigned $E_{\text{AC}} = 0.9$ pJ[14,19,26]. The theoretical inference energy per chunk was therefore estimated as

$$E_{\text{chunk}} = N_{\text{MAC}} E_{\text{MAC}} + N_{\text{AC}} E_{\text{AC}}, \tag{40}$$

where N_{MAC} and N_{AC} denote the estimated number of MAC and AC operations per 256-sample EEG chunk. Signal-level energy was estimated by multiplying the per-chunk energy by the average number of chunks per full Bonn EEG recording, which is 31 in this work.

Table 6. Theoretical CMOS switching-energy estimates using 45 nm operation-energy constants. Dense energy uses the equivalent dense operation count, while spike-aware energy scales recurrent operations by measured spike activity.

Model	Dense/chunk	Spike-aware/chunk	Energy/signal	Saving
Conv-LSTM	50.21 μJ	–	1.56 mJ	–
S-LSTM + Attn	126.74 μJ	65.60 μJ	2.03 mJ	48.2%
Conv1D + S-LSTM + Attn	125.35 μJ	72.42 μJ	2.24 mJ	42.2%

These estimates make the MAC-to-AC transition quantitatively explicit. For the pure spiking model, measured spike activity reduces the theoretical per-chunk switching energy from 126.74 μJ to 65.60 μJ, corresponding to a 48.2% reduction relative to its dense-equivalent operation count. For the hybrid Conv1D spiking model, spike-aware execution reduces the theoretical per-chunk energy from 125.35 μJ to 72.42 μJ, corresponding to a 42.2% reduction relative to its dense-equivalent operation

count. These savings quantify the effect of sparsity: event-driven execution removes 42–48% of the switching energy that the same networks would require in dense form. Because the savings are measured relative to each model's own dense equivalent, total energy per recording still reflects model size; in the current configuration the larger spiking models are estimated at 2.03 mJ and 2.24 mJ per recording, compared with 1.56 mJ for the smaller baseline. Since sparsity scales operation counts multiplicatively, under the same estimates a size-matched spiking model retaining the measured sparsity would be expected to fall below the baseline; demonstrating this, together with quantized weights for all models, is the natural next step.

Discussion

The central finding of this study is that sparse spiking recurrent processing preserves seizure-related EEG classification performance under a strict, leakage-safe protocol. Both spiking models reached cross-validated signal-level accuracy statistically indistinguishable from the dense Conv-LSTM (86.4% and 86.6% compared with 87.8%), and the confidence intervals bound any accuracy gap at about 4 percentage points. This matters because the spiking models achieve it under a fundamentally more restrictive computational regime: recurrent communication is reduced to binary events, and the recurrent core of Model 3 operated with 79.2% hidden-state sparsity. In other words, roughly four in five recurrent units are silent at any timestep, yet the classifier loses no significant accuracy.

The second finding is noise tolerance. The pure spiking model lost only 2 percentage points at 20 dB SNR, compared with 26 points for the dense baseline, and it retained the highest accuracy at every tested noise level. The thresholded delta-sigma encoder acts as a built-in noise filter, responding only to temporal changes that exceed the input threshold. For wearable EEG, where acquisition conditions are rarely clean, this is a practically relevant property that the dense baseline does not offer.

The two spiking designs also illustrate a useful design trade-off. The pure event-based pathway of Model 2 gives the strongest noise tolerance, while the continuous Conv1D frontend of Model 3 gave the cleanest class separation on the hold-out split at the cost of passing more noise into the recurrent core. Under cross-validation the two models are statistically indistinguishable in clean accuracy ($p = 1.0$), so the choice between them can be guided by the expected acquisition environment.

The deployment analysis identifies what is needed to turn these properties into hardware gains. The present spiking models use wider recurrent layers than the baseline and therefore have 2.5–2.7 times more parameters; their CPU latency reflects software simulation of spiking dynamics; and their estimated total energy, although reduced by 42–48% through sparsity, still reflects their larger size. None of these factors is intrinsic to spiking computation. Each points to a concrete next step: size-matched spiking architectures, quantized weights, and event-driven execution.

The latency results should be interpreted carefully. The current spiking models are implemented as PyTorch software models and are not executed on neuromorphic hardware, so their measured CPU latency reflects the cost of simulating spiking dynamics timestep by timestep on a conventional processor. Modern neuromorphic platforms such as Intel Loihi 2, IBM TrueNorth, BrainScaleS-2, SpiNNaker2, and SynSense Speck are designed around sparse event communication, local memory, and asynchronous or massively parallel spike processing[27–31]. Loihi 2 supports programmable spiking-neuron dynamics and is commonly used through Intel's Lava software framework[27,32]. IBM TrueNorth demonstrated a large-scale digital neurosynaptic architecture with one million neurons and 256 million synapses[28], while mixed-signal platforms such as BrainScaleS-2 and commercial edge sensors such as SynSense Speck illustrate the growing relevance of neuromorphic inference for low-power sensory processing[29,31]. Mapping the spiking models presented here onto such platforms is the natural route to realizing their latency and energy benefits, and is left to future work.

A practical hardware transition would require converting the trained software model into a neuromorphic deployment format. In the present work, the spiking recurrent weights are learned in PyTorch using surrogate-gradient descent, while the forward pass uses threshold-based binary spike generation. For deployment, the learned weights would need to be quantized to the precision supported by the target platform (for example, at most 8-bit synaptic weights on Loihi 2) and mapped onto the chip's synaptic memory. The spike thresholds, recurrent states, and event-routing rules would then be implemented using the neuron and synapse primitives of the target hardware. The continuous cell state of the spiking LSTM and the continuous frontend of Model 3 would require either programmable neuron models or a separate conventional processor. In this setting, the theoretical MAC-to-AC energy reduction estimated in this study is a hardware-mapping hypothesis: energy and battery-life effects can only be established after quantization, placement, routing, and measurement on an actual device.

Scope of deployment claims

To make the deployment evidence easy to assess, Table 7 groups each deployment-related property by the type of evidence behind it: measured on the evaluated models, estimated under stated assumptions, or requiring hardware validation.

Table 7. Evidence status of the deployment-related properties reported in this study.

Evidence	Property
Measured in this study	Competitive cross-validated accuracy of both spiking models, with any gap to the baseline bounded at about 4 percentage points (Table 3). Binary recurrent activity with 79.2% hidden-state sparsity in Model 3. Strong tolerance of Model 2 to additive white Gaussian noise (Table 4). Parameter counts and float32/int8 storage, and CPU software latency (Table 5).
Estimated under stated assumptions	42–48% switching-energy reduction from sparsity under 45 nm operation-energy constants (Table 6). Neuromorphic latency according to Eq. (39). A one-bit-per-parameter lower bound of 50.6 KB for Model 3, which would require binarized weights.
Requires hardware validation	Accuracy after weight quantization. Execution on neuromorphic or embedded hardware. Measured power consumption and battery life. Tolerance to realistic wearable artefacts. Performance on continuous, imbalanced, multichannel clinical EEG.

Comparison with recent five-class Bonn studies

Table 8 places the evaluated models in the context of recent studies that reported five-class classification on the Bonn EEG data[33–38]. All of the studies in Table 8 use dense, non-spiking networks. The comparison should be interpreted with care because published studies differ in preprocessing, segmentation, cross-validation design, and evaluation unit. In particular, Xu et al. and Ahmad et al. used the UCI Epileptic Seizure Recognition version of the data, in which each recording is divided into 1-second segments that are classified individually, so segments from the same recording can appear in both training and test sets. Our protocol assigns whole recordings to folds before segmentation, which excludes this source of optimism and yields a more conservative estimate. Mallick and Baths also classified full recordings and reported 97.75% with a dense Conv1D–BiLSTM, showing the accuracy attainable by dense models on this dataset. Our results complement these studies: they show that replacing dense recurrence with sparse binary spiking dynamics retains accuracy competitive with a dense baseline trained and evaluated under the identical protocol, while adding noise tolerance.

Table 8. Comparison with recent five-class Bonn EEG studies. NR indicates that the information was not reported. Accuracies are as reported by each study and were obtained under different protocols (see text).

Study	Model type	Five-class accuracy	Size / parameters
Xu et al. (2020)	CNN–LSTM	82.00%	NR
Ahmad et al. (2023)	CNN–BiLSTM	84.10%	NR
Shanmugam and Dharmar (2023)	CNN–LSTM	92.50%	NR
Khan et al. (2023)	CWT + CAE + LSTM	93.25%	391,167 CAE encoder parameters; full pipeline NR
Zhao et al. (2024)	ResBiLSTM	91.27%	314,755 parameters; ≈1.23 MB float32
Mallick and Baths (2024)	Conv1D + BiLSTM	97.75%	NR
This work	**Conv1D–S-LSTM–Attn**	**86.60% (5-fold CV)**	**414,135 parameters; 404.4 KB int8**

Architectural justification and component contribution

The architecture of Model 3 was designed to balance discriminative accuracy with event-based recurrent processing. The Conv1D frontend was included because raw EEG contains local temporal patterns, transient amplitude changes, and short rhythmic structures that may be difficult for a recurrent model to extract from sparse event streams alone. The delta-spike path was retained to preserve event-driven sparsity, whereas the analog residual path was added to limit information loss after thresholding. Temporal attention was included because seizure-related activity is not uniformly distributed across the EEG window.

Under cross-validation, the pure S-LSTM, which lacks the Conv1D frontend and analog path, achieved 86.40% ± 5.41%, and Model 3 achieved 86.60% ± 5.50% ($p = 1.0$; 95% CI of the difference −3.3 to +2.9 percentage points), so the two designs are statistically indistinguishable in clean accuracy. They differ in their secondary behavior: Model 3 made fewer errors on the hold-out split, with fewer Z/O and N/F confusions, whereas Model 2 was considerably more tolerant to noise. Isolating the contribution of each component, by removing the frontend, the analog residual path, and the attention layer one at a time under

cross-validation, is a natural extension of this work.

Limitations

Validation on a single dataset. All results are obtained on the Bonn EEG dataset, which limits the generalizability of the conclusions in four ways. First, the recordings are single-channel 23.6-second segments selected from longer multichannel recordings after visual inspection for artefacts, so the input distribution is cleaner than that of continuous clinical or wearable EEG. Second, the healthy classes (Z, O) were recorded with surface electrodes, whereas the interictal and ictal classes (N, F, S) were recorded intracranially. Part of the separability between healthy and epileptic classes, including the perfect recognition of class S on the hold-out split, may therefore reflect differences in recording modality and amplitude scale rather than seizure physiology alone. Third, the dataset is balanced, with 100 recordings per class, whereas seizures in continuous monitoring are rare events; a model can achieve high overall accuracy in such settings while missing clinically important seizures. Fourth, the data originate from ten subjects and provide no subject identifiers, so the recording-level protocol prevents segment leakage but cannot guarantee subject-independent evaluation. The reported accuracies should therefore be read as a controlled comparison between architectures under identical conditions, not as an estimate of clinical performance. Whether the relative ordering of the models, the noise tolerance of the spike encoder, and the sparsity levels transfer to multichannel scalp EEG remains to be tested on continuous datasets such as CHB-MIT[39,40], using patient-wise splits and event-level metrics such as sensitivity, false alarms per hour, and detection latency.

Synthetic noise model. The noise analysis uses additive white Gaussian noise, which is stationary and spectrally flat. Artefacts in wearable EEG differ substantially: ocular artefacts are low-frequency and high-amplitude, muscle artefacts are bursty and concentrated at higher frequencies, motion artefacts and electrode pops produce abrupt baseline shifts, and power-line interference is narrowband at 50 or 60 Hz. Abrupt, high-amplitude transients are precisely the signal changes that a delta-sigma encoder converts into spikes, so the advantage of the thresholded encoder under AWGN may shrink or reverse under motion and electrode artefacts. In addition, Model 3 was trained with noise augmentation whereas the baseline was not, which confounds the comparison between those two models. The noise results therefore demonstrate tolerance to stationary broadband noise only and should not be read as evidence of tolerance to realistic wearable artefacts. Semi-synthetic evaluation with recorded ocular and muscle artefacts, for example from EEGdenoiseNet[41], together with line noise and electrode-pop models, and ultimately evaluation on wearable recordings, is needed to assess this.

Deployment evidence. Weight quantization, deployment on embedded or neuromorphic hardware, and power measurement were outside the scope of this study, as summarized in Table 7.

Conclusion

This paper compared conventional and spiking recurrent networks for seizure-related EEG classification under a shared, leakage-safe, full-recording evaluation protocol. Five-fold cross-validation showed that sparse spiking recurrent models achieve accuracy competitive with a dense Conv-LSTM, with no significant difference and any gap bounded at about 4 percentage points, while restricting recurrent communication to binary events with about 79% hidden-state sparsity. The pure spiking model was also markedly more tolerant to additive broadband noise, losing 2 percentage points at 20 dB where the dense baseline lost 26.

These results establish the accuracy and noise-tolerance foundation on which low-power wearable seizure monitoring with spiking networks can be built. The transparent deployment analysis shows that sparsity alone removes 42–48% of the estimated switching energy of the spiking networks, and it identifies the steps needed to convert this into system-level gains over dense models: spiking architectures size-matched to dense baselines, quantized weights with measured accuracy, and execution on event-driven hardware. Together with evaluation on realistic artefacts and on continuous, imbalanced clinical EEG, these steps define a clear path from the present software models toward on-device seizure detection.

Declarations

Funding Not applicable.

Conflict of interest/Competing interests The authors declare no competing interests.

Ethics approval and consent to participate Not applicable. The study uses a publicly available anonymized EEG benchmark dataset.

Consent for publication Not applicable.

Data availability The Bonn EEG dataset used in this study is publicly available[3]; the copy used in this study was accessed through Kaggle.

Materials availability Not applicable.

Code availability The implementation notebooks, evaluation code, plotting utilities, and reproducibility instructions are publicly available at: https://github.com/Oliverashraf/spiking-eeg-seizure-classification.

Author contribution Oliver Ashraf implemented the models, conducted the experiments, analyzed the results, and prepared the manuscript draft. Dr. Shady Ahmed Nagy and Prof. Dr. Mohammed Abdel-Megeed M. Salem supervised the study, provided methodological guidance, and reviewed the manuscript.

References

1. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*. 1997;9(8):1735–1780.
2. Neftci EO, Mostafa H, Zenke F. Surrogate gradient learning in spiking neural networks: bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*. 2019;36(6):51–63.
3. Andrzejak RG, Lehnertz K, Mormann F, Rieke C, David P, Elger CE. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. *Physical Review E*. 2001;64:061907.
4. Tian F, Yang J, Zhao S, Sawan M. A new neuromorphic computing approach for epileptic seizure prediction. In: *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE; 2021. p. 1–5.
5. Chambers JD, Cook MJ, Burkitt AN, Grayden DB. Using long short-term memory recurrent neural networks to classify unprocessed EEG for seizure prediction. *Frontiers in Neuroscience*. 2024;18:1472747.
6. Xu Z, Wang T, Cao J, Bao Z, Jiang T, Gao F. BECT spike detection based on novel EEG sequence features and LSTM algorithms. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2021;29:1734–1743.
7. Bellec G, Salaj D, Subramoney A, Legenstein R, Maass W. Long short-term memory and learning-to-learn in networks of spiking neurons. *Advances in Neural Information Processing Systems*. 2018;31.
8. Lotfi Rezaabad A, Vishwanath S. Long short-term memory spiking networks and their applications. In: *International Conference on Neuromorphic Systems*. 2020. p. 1–9.
9. Zhang R, Li R, Liang J, Yue K, Li W, Li Y. Long short-term memory spiking neural networks for classification of snoring and non-snoring sound events. *Chinese Journal of Electronics*. 2024;33(3):793–802.
10. Ponghiran W, Roy K. Hybrid analog-spiking long short-term memory for energy efficient computing on edge devices. In: *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE; 2021. p. 581–586.
11. Datta G, Deng H, Aviles R, Beerel PA. Towards energy-efficient, low-latency and accurate spiking LSTMs. arXiv preprint arXiv:2210.12613. 2022.
12. Gruber A. *Spiking LSTM on Loihi 2: A Neuromorphic Reinterpretation of Recurrent Networks*. Master's thesis, Politecnico di Torino; 2025.
13. Souza de Abreu R, Silva I, Nunes YT, Moioli RC, Guedes LA. Advancing fault prediction: A comparative study between LSTM and spiking neural networks. *Processes*. 2023;11(9):2772.
14. Sharifian S, Taheri M, Rashtchi V, Azarpeyvand A, Herglotz C, Jenihhin M. Reliability-aware hyperparameter optimization for ANN-to-SNN conversion. *WiPiEC Journal*. 2025;11(1):7.
15. Aydin A, Gehrig M, Gehrig D, Scaramuzza D. A hybrid ANN-SNN architecture for low-power and low-latency visual perception. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024. p. 5701–5711.
16. Pokala SD, Bernert M, Nanami T, Kohno T, Lévi T, Yvert B. A frugal spiking neural network for unsupervised multivariate temporal pattern classification and multichannel spike sorting. *Nature Communications*. 2025;16(1):9218.
17. Kumar N, Tang G, Yoo R, Michmizos KP. Decoding EEG with spiking neural networks on neuromorphic hardware. *Transactions on Machine Learning Research*. 2022.
18. Sam A, Boostani R, Hashempour S, Taghavi M, Sanei S. Depression identification using EEG signals via a hybrid of LSTM and spiking neural networks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2023;31:4725–4737.
19. Ma C, Chen X, Li Y, Yang Q, Wu Y, Li G, Pan G, Tang H, Tan KC, Wu J. Spiking neural networks for temporal processing: status quo and future prospects. arXiv preprint arXiv:2502.09449. 2025.
20. Ferreira PM, Wang S, Gao Y, Benlarbi-Delai A. A comparative review of deep and spiking neural networks for edge AI neuromorphic circuits. *Frontiers in Neuroscience*. 2025;19:1676570.
21. Schuman CD, Kulkarni SR, Parsa M, Mitchell JP, Date P, Kay B. Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science*. 2022;2:10–19.

22. Roy K, Jaiswal A, Panda P. Towards spike-based machine intelligence with neuromorphic computing. *Nature*. 2019;575:607–617.

23. Abreu S, Pedretti G, Bohnstingl T, Pehle C, Petrovici MA. Neuromorphic principles for efficient large-scale learning. In: *International Conference on Learning Representations*. 2024.

24. Yang T, Chen Y, Sze V. Designing energy-efficient edge AI systems: A survey. *IEEE Transactions on Circuits and Systems for Artificial Intelligence*. 2024.

25. Zenke F, Ganguli S. SuperSpike: Supervised learning in multilayer spiking neural networks. *Neural Computation*. 2018;30(6):1514–1541.

26. Horowitz M. 1.1 Computing's energy problem and what we can do about it. In: *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*. IEEE; 2014. p. 10–14.

27. Davies M, Wild A, Orchard G, Sandamirskaya Y, Guerra GA, Joshi P, Plank P, Risbud SR. Advancing neuromorphic computing with Loihi: A survey of results and outlook. *Proceedings of the IEEE*. 2021;109(5):911–934.

28. Akopyan F, Sawada J, Cassidy A, Alvarez-Icaza R, Arthur J, Merolla P, Imam N, Nakamura Y, Datta P, Nam GJ, Taba B, Beakes M, Brezzo B, Kuang J, Manohar R, Risk W, Jackson B, Modha DS. TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*. 2015;34(10):1537–1557.

29. Pehle C, Billaudelle S, Cramer B, Kaiser J, Schreiber K, Stradmann Y, Weis J, Leibfried A, Müller E, Schemmel J. The BrainScaleS-2 accelerated neuromorphic system with hybrid plasticity. *Frontiers in Neuroscience*. 2022;16:795876.

30. Mayr C, Hoeppner S, Furber S. SpiNNaker 2: A 10 million core processor system for brain simulation and machine learning. *arXiv preprint arXiv:1911.02385*. 2019.

31. SynSense. *Speck: Dynamic vision sensor with integrated neuromorphic processing for ultra-low-power edge inference*. SynSense product documentation; 2024.

32. Intel Labs. *Intel Labs' Loihi 2 Neuromorphic Research Chip and Lava Software Framework*. Intel Technology Brief; 2021.

33. Xu G, Ren T, Chen Y, Che W. A 1D CNN-LSTM model for epileptic seizure recognition using EEG signal analysis. *Frontiers in Neuroscience*. 2020;14:578126.

34. Ahmad I, Wang X, Javeed D, Kumar P, Samuel OW, Chen S. A hybrid deep learning approach for epileptic seizure detection in EEG signals. *IEEE Journal of Biomedical and Health Informatics*. 2023. doi:10.1109/JBHI.2023.3265983.

35. Shanmugam S, Dharmar S. A CNN-LSTM hybrid network for automatic seizure detection in EEG signals. *Neural Computing and Applications*. 2023;35(28):20605–20617.

36. Khan SU, Jan SU, Koo I. Robust epileptic seizure detection using long short-term memory and feature fusion of compressed time–frequency EEG images. *Sensors*. 2023;23(23):9572.

37. Zhao W, Wang WF, Patnaik LM, Zhang BC, Weng SJ, Xiao SX, Wei DZ, Zhou HF. Residual and bidirectional LSTM for epileptic seizure detection. *Frontiers in Computational Neuroscience*. 2024;18:1415967.

38. Mallick S, Baths V. Novel deep learning framework for detection of epileptic seizures using EEG signals. *Frontiers in Computational Neuroscience*. 2024;18:1340251.

39. Shoeb AH. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology; 2009.

40. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*. 2000;101(23):e215–e220.

41. Zhang H, Zhao M, Wei C, Mantini D, Li Z, Liu Q. EEGdenoiseNet: a benchmark dataset for deep learning solutions of EEG denoising. *Journal of Neural Engineering*. 2021;18(5):056057.