scientific reports

Check for updates

OPEN

ConvSNN: an energy-efficient hybrid convolutional-spiking neural network for wrist-worn stress and affective state recognition

Muhammad Zaheer Sajid[1]✉, Nour Aburaed[2] & Rytis Maskeliūnas[3]

Continuous stress monitoring on wrist-worn devices matters for real-time affective computing, digital health, and personalized well-being, yet it remains difficult because a wearable model must operate with few sensors, little compute, and a tight energy budget. Chest-mounted systems can draw on high signal-to-noise signals such as ECG, EMG, and respiration, but a wrist device captures only blood-volume pulse, electrodermal activity, skin temperature, and accelerometry, which makes accurate and efficient stress recognition harder. This study presents ConvSNN, a hybrid convolutional–spiking neural network for multimodal affective recognition from wrist-worn physiological signals. ConvSNN pairs a compact four-stage 1D convolutional backbone with a time-to-first-spike (TTFS) classification head that reads aggregated rate-coded and time-coded spike statistics to keep predictions stable. It also adds a spiking confidence gate for uncertainty-aware window acceptance and a temperature-scaling step for probability calibration. We evaluate ConvSNN on the WESAD benchmark using only the six wrist channels of the Empatica E4 wearable (blood-volume pulse, electrodermal activity, skin temperature, and three-axis accelerometry). Under random 85/15 split evaluation it reaches a mean accuracy of 98.07% with a macro-F1 of 0.977 across five random seeds. Under strict leave-one-subject-out (LOSO) cross-validation, ConvSNN reaches a mean accuracy of 94.0% with a macro-F1 of 0.932 over five seeds, which confirms strong subject-independent generalization and amounts to a 4.1-percentage-point drop from the random-split protocol. Temperature scaling lowers the expected calibration error from 0.149 to 0.017 with no change in accuracy, and the spiking confidence gate opens a 5.9-percentage-point accuracy gap between accepted and rejected windows, which supports principled abstention on low-confidence inputs. Based on operation-count emulation, ConvSNN has an estimated per-inference energy of 48.31 μJ at 5.70 ms latency, which points to low-power real-time use on smartwatch-class devices once it is validated on embedded hardware. Taken together, these results show that hybrid convolutional–spiking models can strike an effective balance among recognition performance, subject-independent robustness, calibrated confidence estimation, and energy-efficient deployment for continuous wrist-worn affective monitoring.

Keywords Convolutional neural network, Spiking neural network, Time-to-first-spike, Wearable stress recognition, Neuromorphic computing, Edge inference, Wrist-worn sensors, Temperature scaling

Stress and affective states are among the most consequential psychophysiological states for long-term health and well-being, and they leave measurable signatures in physiological signals such as electrodermal activity, blood-volume pulse, and skin temperature; wearable sensing now enables routine recording of these high-quality signals, together with motion, from wrist-worn devices[1]. Progress in flexible bioelectronics, low-power electronics, and learning algorithms has made ecologically valid stress detection feasible outside the laboratory[2], and a recent systematic review catalogs the open datasets, sensing modalities, and evaluation protocols that have shaped more than a decade of this work[3]. The best-known benchmark in this area is the Wearable Stress and Affect Detection (WESAD) corpus[4]. It provides synchronized recordings from a chest-mounted RespiBAN device (electrocardiography, electromyography, and respiration) together with a wrist-mounted Empatica E4 device that captures blood-volume pulse, electrodermal activity, skin temperature, and three-axis accelerometry.

[1]Department of Electrical and Computer Engineering, George Mason University, Fairfax 22030, VA, USA. [2]College of Engineering and IT, University of Dubai, Dubai, UAE. [3]Software Engineering Department, Faculty of Informatics, Kaunas University of Technology, Kaunas 51423, Lithuania. ✉email: msajid4@gmu.edu

Methods that use the full WESAD modality set have approached near-perfect accuracy under random-split evaluation, whereas wrist-only methods have plateaued in the 80–92% range. This gap is attributable to the absence of ECG-derived heart-rate-variability features at the wrist.

From a deployment standpoint, this accuracy ranking is misleading, because gel-electrode chest straps are tolerated in research studies but not in everyday life. Continuous on-device monitoring must therefore run on a single wrist-worn device within a microjoule-class energy budget, which rules out dense convolutional or transformer inference. This constraint motivates hardware-aware low-precision arithmetic, such as posit-format compute pipelines for wearable inference accelerators[5], and it motivates spike-based neural computation, where inference cost scales with the number of spike events the network emits rather than with the dense activation count at every timestep. The neuromorphic-computing literature offers a comprehensive view of both the algorithmic and the hardware trajectories of spike-based inference[6], and recent topical surveys point to the rapid maturation of dedicated neuromorphic chips for edge deployment[7]. Together they establish spiking neural networks (SNNs) as a strong fit for wrist-worn inference, where every microjoule matters. A third practical concern is uncertainty handling. Real-world wrist signals are corrupted by motion artifacts, sweat saturation, and transient contact loss, so a useful continuous monitor must both classify each window and estimate when its prediction should not be trusted. That, in turn, calls for calibrated probabilities and a discriminative confidence scalar that supports principled abstention.

To address these three constraints together, namely wrist-only sensing, microjoule-class inference, and calibrated abstention, we propose ConvSNN, a hybrid convolutional–spiking neural network architecture for wrist-worn stress and affective state recognition. ConvSNN combines a sample-efficient four-stage 1D convolutional backbone with a time-to-first-spike (TTFS) classification head whose logits come from a dense readout over aggregated rate-coded and time-coded spike statistics, along with a small spiking confidence gate whose sigmoid output is trained against the classifier's continuous probability on the true class. At inference time, a single learned temperature scalar calibrates the classifier output, and an adaptive threshold on the gate confidence supports abstention on low-quality windows. ConvSNN differs from previous wearable stress-recognition models in four concrete ways. First, it works exclusively on the six channels of a consumer-class wrist wearable and never touches chest electrocardiography, electromyography, or respiration. Second, its classification pathway is hybrid rather than fully spiking: it keeps a dense convolutional backbone for representation learning while confining spiking dynamics to the classification head and confidence gate. Third, it provides calibrated per-window abstention through a learned spiking gate. Fourth, it reports an operation-count-based per-inference energy estimate of 48.31 μJ at 5.70 ms latency, which suits the always-on operation of a smartwatch-class compute budget that neither chest-augmented nor fully-spiking-EEG baselines on WESAD report.

Research motivation

The three requirements raised above can be stated compactly. First, the sensor set is restricted to the wrist for all-day wearability, which excludes chest-mounted ECG, EMG, and respiration. Second, inference has to fit within the sub-10 ms latency and microjoule-level energy budget of always-on smartwatch use, which favors spike-based computation whose cost scales with sparse spike activity rather than dense activations. Third, real-world wrist signals are degraded by motion, sweat, and contact loss, so a deployable model has to estimate when its own prediction should not be trusted instead of emitting a label for every window. ConvSNN addresses these three requirements together. As a concrete illustration, consider a smartwatch that monitors a user's affective state throughout the day: while the user is seated at a desk, clean blood-volume-pulse and electrodermal-activity signals let the model confidently report a stress or baseline state at a cost of only tens of microjoules per reading, so continuous all-day monitoring drains a negligible fraction of the battery; when the user then goes for a run, the same channels are corrupted by motion and sweat, and rather than emitting an unreliable label the confidence gate abstains and defers the decision until a clean window is available. This example ties the three requirements together, wrist-only sensing, microjoule-class inference, and calibrated abstention, in a single everyday scenario, and the end-to-end pathway that realizes it is depicted in Fig. 1.

Research contribution

The main contributions of this work are summarized as follows:

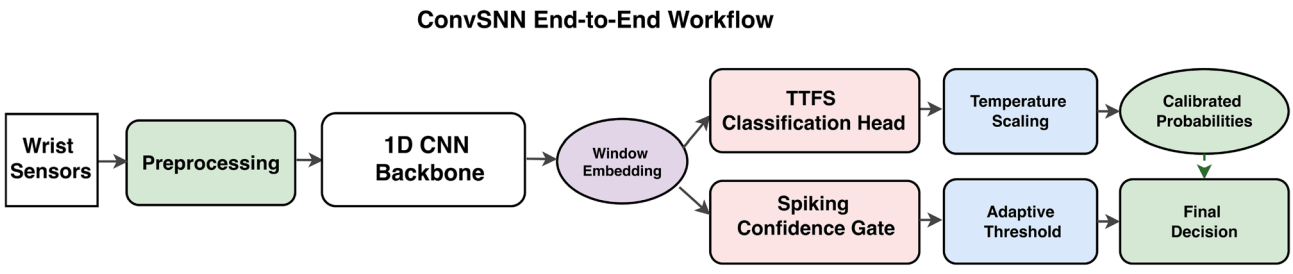


Fig. 1. End-to-end ConvSNN workflow: six wrist channels are preprocessed and passed through a 1D-CNN backbone, then routed in parallel to a TTFS classification head and a spiking confidence gate, producing a calibrated class label and an accept/abstain decision.

- We propose ConvSNN, a compact hybrid convolutional–spiking neural network that joins a four-block 1D convolutional neural network (CNN) backbone with a spiking decision pathway for accurate and energy-efficient wrist-worn stress recognition.
- We design a stable TTFS classification head that works from aggregated rate-coded and time-coded spike statistics rather than direct first-spike decoding, which improves trainability while keeping inference low-latency and spike-based.
- We introduce a spiking confidence gate that predicts the classifier's continuous true-class probability, which lets the system abstain on low-confidence physiological windows.
- We apply adaptive gate thresholding and post-hoc temperature scaling to deliver calibrated confidence estimates without changing the final class predictions.
- We validate ConvSNN on the WESAD wrist-only benchmark under two protocols: a random 85/15 split (mean accuracy 98.07%, macro-F1 0.977 across five seeds) and strict LOSO cross-validation (mean accuracy 94.0%, macro-F1 0.932 across five seeds). This shows both protocol-matched comparability with the WESAD literature and strong subject-independent generalization, with an operation-count-based per-inference energy estimate of 48.31 μ J and 5.70 ms latency.

The rest of the paper is organized as follows. Section "Related work" reviews related literature. Section "Research methodology" formalizes the proposed architecture and the training algorithm. Section "Results" presents the experimental evaluation on WESAD. Section "State-of-the-art comparison" gives a head-to-head state-of-the-art comparison. Section "Discussion" explains why ConvSNN delivers strong wrist-only accuracy at low reported energy. Section "Conclusion" concludes.

Related work

Stress recognition from wearable physiological signals

The WESAD chest channels carry far more information for stress classification than the wrist channels, and methods that use the full multimodal set have approached near-perfect accuracy under random-split evaluation. Cross-modality studies on the full WESAD modality set report accuracies above 99% with both classical and deep-learning models[8], while wrist-only deep-learning frameworks for early-stage stress detection show that high accuracy is achievable from a single physiological stream when chest channels are unavailable[9]. Multimodal fusion networks that combine time-domain and frequency-domain features confirm that merging heterogeneous physiological channels improves accuracy beyond what any single modality can reach[10]. Studies of anxiety detection in real-world conditions report a sizable accuracy gap between laboratory and free-living settings, which suggests that subject-independent generalization, rather than within-laboratory accuracy, is the relevant deployment criterion[11]. Beyond stress and affect, wearable physiological sensing is increasingly paired with on-device intelligence for assistive health applications; for example, a wearable intelligent throat that fuses throat-muscle vibration and carotid-pulse sensing with a lightweight language model restores fluent speech for stroke patients with dysarthria at sub-second latency[29], which underscores the same low-power, real-time, wearable constraints that motivate ConvSNN.

Convolutional and transformer architectures for physiological signals

Convolutional neural networks have long been the workhorse of physiological signal classification, thanks to their inductive bias for local temporal patterns and their relatively modest data requirements. Hybrid CNN–LSTM–attention pipelines reach strong performance on multimodal fatigue detection by combining surface electromyography and inertial measurement units[12]. More recently, transformer-based approaches have been explored for physiological data: stress classification from electrocardiographic spectrograms with vision-transformer models[13], and transformer encoders applied directly to electrodermal activity to capture decomposition components that correlate with affective state[14]. Cross-modal contrastive attention transformers with graph fusion push the accuracy further on multimodal stress data[15], and cross-temporal attention fusion learns soft bidirectional alignments between asynchronous physiological streams in self-supervised settings[16]. A parallel body of work on EEG-based driver-state and fatigue recognition has advanced the preprocessing, feature-extraction, and deep-learning pipeline for physiological signals more broadly: depthwise separable convolutional networks coupled with brain-connectivity features improve recognition for previously unseen drivers[23], entropy-informed deep residual networks capture the nonlinear dynamics of EEG for fine-grained driver-state classification[24], and a recent review synthesizes the nonlinear-dynamics, brain-network, and deep-learning advances that now define this area[25]. These works show that carefully designed convolutional and residual architectures remain strong on the small, subject-limited pools typical of wearable physiological corpora, where high-capacity transformer models need large datasets, extensive pretraining, or heavy regularization before they become competitive. This is what motivates our use of a compact 1D-CNN backbone rather than a transformer in the wrist-only setting targeted here.

Spiking and hybrid spiking–dense architectures

A complementary line of research swaps dense activations for sparse spike trains. Joshi *et al.* propose a Convolutional Spiking Neural Network (CSNN) for stress detection from EEG, in which spiking neurons replace both the convolutional feature extractors and the classification layers; applied to wavelet-decomposed PhysioNet EEG signals, it reaches 98.75% accuracy under 10-fold cross-validation[17]. Adaptive-reset leaky-integrate-and-fire neurons refine the canonical spiking primitive by modulating the reset behavior in response to input statistics[18], and spintronic memtransistor implementations of leaky-integrate-and-fire neurons show that the hardware substrates for SNNs are now mature enough for end-to-end realization[19]. Advances in surrogate-gradient training enable spike-timing learning beyond pure rate coding[20], and adaptive surrogate gradients

improve sequential reinforcement learning in spiking networks[21]. SpiTranNet integrates a transformer with spiking multi-head attention for motor-imagery and sleep-apnea detection, replacing dense activations *inside* the attention mechanism with spiking neurons[22]. Along the same line, spiking attention has also been used for signal restoration rather than classification: STAND-Net, a spiking temporal-attention autoencoder, performs energy-efficient EEG artifact removal, showing that spike-domain attention can preserve fine temporal structure at low compute cost[28]. The recurring trade-off across this line of work is that fully-spiking architectures deliver the strongest energy economy but accumulate surrogate-gradient noise across all layers, which limits their trainability on the small subject pools characteristic of wrist-worn physiological corpora. Hybrid spiking–dense architectures, in which a dense feature extractor is combined with a spiking classification pathway, have not yet been explored systematically for the wrist-only stress-recognition setting targeted here.

Positioning of ConvSNN

These approaches fall into three main directions, each with its own accuracy and deployment trade-offs. The first direction covers high-accuracy multimodal systems that use chest-mounted sensors, especially ECG. These methods can reach very high accuracy on WESAD, but chest-based sensors are not practical for continuous daily use. The second direction covers wrist-only deep-learning models, mostly based on convolutional or recurrent networks. They use a more practical sensor setup, but their accuracy is usually lower than that of chest-augmented systems. The third direction covers spiking neural networks aimed at energy-efficient inference. Many of these models, however, are limited by simplified input settings or are built for EEG signals recorded from scalp electrodes, which are not suitable for smartwatch-based stress monitoring.

ConvSNN occupies a region of this design space that has not been explored before. To the best of our knowledge, ConvSNN is among the first wrist-only WESAD models to bring together a hybrid CNN–SNN classifier, calibrated confidence gating, LOSO evaluation, and per-window energy reporting in a single architecture. It works exclusively on the six channels of a consumer wrist wearable (blood-volume pulse from a photoplethysmography sensor, electrodermal activity from a galvanic skin response sensor, skin temperature from a thermistor, and three-axis accelerometry) and delivers competitive accuracy at a much lower estimated per-inference energy than dense baselines. The wrist-only constraint is deliberate and central: smartwatches such as the Apple Watch, Garmin Vivosmart, Fitbit Sense, Empatica EmbracePlus, and Withings ScanWatch all provide exactly this set of physiological channels at varying sampling rates, and ConvSNN's input front-end resamples them to a common 64 Hz rate without requiring any sensor these devices do not already carry. Within the spiking-architecture family, ConvSNN differs from the fully-spiking convolutional networks above by keeping a dense convolutional backbone, which is essential for stable training on the small wrist datasets where fully-spiking convolutional encoders underfit because of compounded surrogate-gradient noise. It differs from spiking-transformer designs by using a 1D-CNN rather than a transformer, which removes the data-hungriness of self-attention and makes training feasible on the limited subject pools typical of wrist-only physiological corpora. Within the wrist-only family, ConvSNN is the first to add a learned per-window confidence scalar with adaptive thresholding, post-hoc temperature scaling for calibrated probabilities, and reported per-inference energy and latency, all of which are needed for the very deployment scenarios that motivate wrist-only sensing in the first place.

It is worth stating directly how this work is distinguished from prior wearable physiological-recognition systems. The scientific contribution is the specific combination, not previously assembled for the wrist-only setting, of four elements: (i) a hybrid convolutional–spiking factorization that keeps a dense backbone for sample-efficient feature learning while confining spike-domain computation to the decision pathway, which avoids both the underfitting of fully-spiking encoders on small wrist datasets and the data-hungriness of transformer encoders; (ii) a stable time-to-first-spike head decoded through aggregated rate- and time-coded spike statistics rather than a non-differentiable first-spike argmin; (iii) a learned, calibrated spiking confidence gate that supports principled per-window abstention, together with post-hoc temperature scaling; and (iv) a deployment-oriented evaluation that reports subject-independent LOSO performance alongside operation-count energy and latency. To our knowledge, no prior wrist-only WESAD system reports this combination, and this integration, rather than any single component in isolation, is the contribution that distinguishes the present work.

Research methodology

This paper proposes ConvSNN, a hybrid convolutional–spiking neural network architecture for wrist-worn stress and affective state recognition that pairs a sample-efficient dense feature extractor with a sparse spike-domain decision pathway. The end-to-end workflow runs through four sequential phases, shown in Fig. 1. Raw multimodal physiological signals from a wrist-worn device, namely blood-volume pulse, electrodermal activity, skin temperature, and three-axis accelerometry (six channels in total), are first windowed into fixed-length 60-second segments and resampled to a common rate of 64 Hz. Each subject's windows are then standardized with that subject's own mean and standard deviation, which reduces between-subject distributional shift before any model state is involved. The synchronized multichannel waveform is processed in the time domain by a four-stage 1D convolutional backbone that progressively downsamples the temporal axis while expanding the feature axis, producing a fixed-length window embedding $h \in \mathbb{R}^d$ with $d = 128$. As Fig. 1 shows, the window embedding is then decoded by two parallel spiking branches: a time-to-first-spike classification head that produces class logits through a dense readout over aggregated spike statistics, and a spiking confidence gate that emits a per-window certainty score $\rho \in (0, 1)$. A single learned temperature scalar is applied to the classifier logits to obtain calibrated class probabilities. The gate confidence is then compared with a training-derived minimum confidence cutoff, and the final decision rule either returns the predicted class with its probability

vector or abstains when the window is considered too uncertain for reliable reporting. The rest of this section formalizes each component of the workflow.

One-dimensional convolutional backbone

Let $X \in \mathbb{R}^{L \times C_{\text{in}}}$ denote a windowed multichannel input with L time samples and $C_{\text{in}} = 6$ synchronized wrist physiological channels. The convolutional backbone, shown in Fig. 2, is a sequence of four convolutional blocks, each made up of a 1D convolution, group normalization, GELU activation, max-pooling, and dropout. The block sequence is

$$X \to \text{Conv}_{32}^{15} \to \text{Conv}_{64}^{11} \to \text{Conv}_{128}^{7} \to \text{Conv}_{d}^{5} \to h, \tag{1}$$

where Conv_c^k denotes a 1D convolution with c output channels and kernel size k. As annotated in Fig. 2, the first three blocks apply a stride-2 convolution followed by stride-2 max-pooling, for an aggregate temporal downsampling factor of 64×, taking the temporal axis from L to $L/4$, $L/16$, and $L/64$ across the first three stages; the final block applies global adaptive average pooling to obtain a fixed-length embedding $h \in \mathbb{R}^d$ with $d = 128$. We use group normalization instead of batch normalization throughout because, on the small wrist-only WESAD dataset, the running averages of batch statistics lag behind the actual batch statistics during training and yield unstable validation behavior; group normalization removes this dependence on batch statistics and produces smoother training trajectories.

The backbone is parameter-efficient by design. With six input channels and a $d = 128$ output dimension, the four convolutional blocks in Fig. 2 hold roughly 1.7×10^5 trainable parameters, about thirty times fewer than a typical transformer encoder of comparable depth, which is essential for stable training on the small subject pool of WESAD's wrist-only configuration.

Time-to-first-spike classification head

The classification head, shown in Fig. 3, consists of C output leaky integrate-and-fire (LIF) neurons, one per class. A LIF neuron is a standard spiking primitive that integrates its input over time and emits a spike when its membrane potential crosses a threshold. The inputs to the head come from a single linear projection of the window embedding, $u_c^{(0)} = w_c^\top h + b_c$, where $h \in \mathbb{R}^d$ ($d = 128$) is the window embedding, $w_c \in \mathbb{R}^d$ and $b_c \in \mathbb{R}$ are the weight vector and bias of output neuron c (equivalently, stacking the C rows gives $u^{(0)} = W_{\text{cls}} h + b_{\text{cls}}$ with $W_{\text{cls}} \in \mathbb{R}^{C \times d}$ and $b_{\text{cls}} \in \mathbb{R}^C$), and $u_c^{(0)}$ is the constant input current driving neuron c. Each neuron's membrane potential is integrated over $T = 16$ timesteps with a constant input current,

$$V_c^{(t)} = (1 - 1/\tau_m)\, V_c^{(t-1)} + u_c^{(0)}, \tag{2}$$

where τ_m is the membrane time constant. Here $t \in \{1, \ldots, T\}$ is the discrete simulation timestep index, and a superscript (t) on any quantity denotes its value at timestep t; thus $V_c^{(t)}$ and $V_c^{(t-1)}$ are the membrane potentials of neuron c at the current and previous timesteps, and $s_c^{(t)} \in \{0, 1\}$ (used below and in Eq. 4) is the spike emitted by neuron c at timestep t. As the central LIF block of Fig. 3 shows, a neuron emits a spike at the first timestep where $V_c^{(t)} \geq \theta_0$, after which the membrane potential is reset by subtracting θ_0 (soft reset). The first-spike time $\tau_c \in [1, T]$ is recorded, with $\tau_c = T$ indicating that the neuron did not fire within the simulation window.

To make surrogate-gradient training stable, ConvSNN decodes the head output through a dense readout over aggregated spike features rather than the argmin of the first-spike-time vector, because the argmin operation

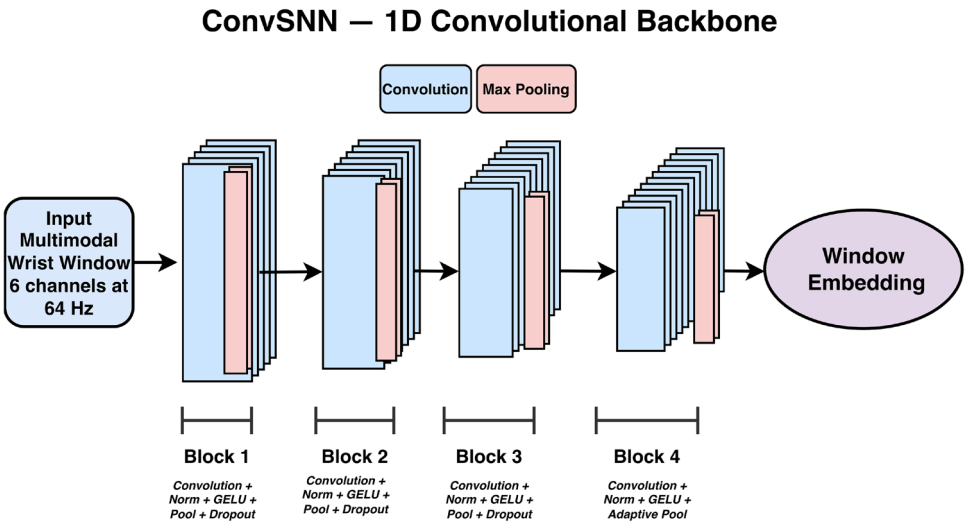


Fig. 2. 1D convolutional backbone of ConvSNN: four sequential blocks of convolution, normalization, GELU, pooling, and dropout that progressively compress the time axis while expanding channel depth, producing the window embedding consumed by both spiking heads.

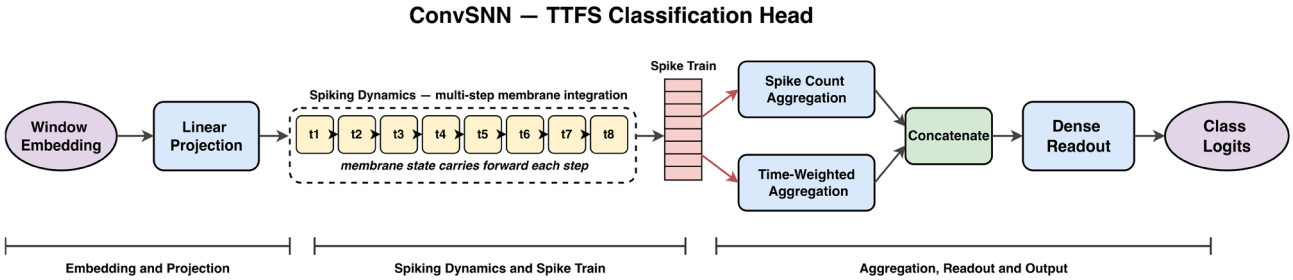


Fig. 3. TTFS classification head: the window embedding is projected and driven through a multi-step LIF neuron bank, whose spike train is summarized by spike-count and time-weighted statistics and combined by a dense readout into class logits.

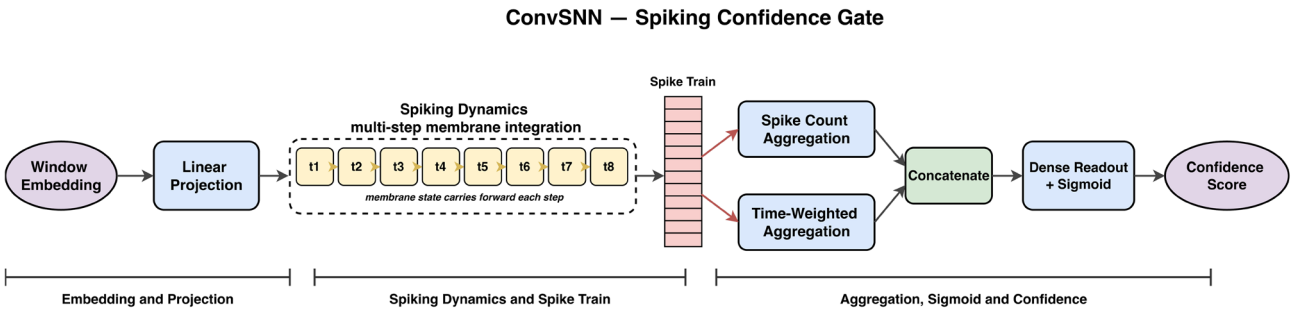


Fig. 4. Spiking confidence gate: structurally similar to the TTFS head but trained against a soft probability target rather than binary correctness, yielding a continuous calibrated confidence score that drives the adaptive accept/abstain threshold.

is non-differentiable and training through it suffers from severe gradient vanishing. From the full spike train $s_c^{(t)} \in \{0,1\}$ over T timesteps, two complementary statistics are computed in parallel, as the lower half of Fig. 3 shows: a rate-coded feature $\sigma_c = \sum_t s_c^{(t)}$ that counts the total number of spikes, and a time-coded feature $\omega_c = \sum_t \frac{T-t}{T} s_c^{(t)}$ that gives more weight to earlier spikes. These per-class features are concatenated into a $2C$-dimensional vector and projected through a small dense readout to produce the classification logits,

$$\hat{p} = \mathrm{softmax}\big(W_{\mathrm{out}}\,[\sigma_1, \ldots, \sigma_C, \omega_1, \ldots, \omega_C]^\top + b_{\mathrm{out}}\big)\,. \tag{3}$$

This formulation has three properties that matter for stable training. First, it preserves the sparse spike-domain dynamics: inference still emits at most a handful of spikes per output neuron over the entire T-step simulation, so the energy budget remains favorable. Second, the dense readout absorbs the discreteness of individual spike events, which removes the gradient-vanishing problem tied to direct first-spike decoding through a non-differentiable argmin. Third, the time-coded feature ω_c keeps the latency advantage of TTFS coding, since classes whose neurons fire earlier contribute more strongly to the final logits.

The surrogate gradient for each spike emission uses a smooth ATan-style relaxation,

$$\frac{\partial s^{(t)}}{\partial V^{(t)}} \approx \frac{\alpha}{2} \frac{1}{1 + (\alpha(V^{(t)} - \theta_0))^2}, \tag{4}$$

with surrogate slope $\alpha = 2.0$. The wider surrogate window given by $\alpha = 2.0$, compared with the more aggressive $\alpha \geq 4$ often used for vision tasks, eases the dying-spike problem, in which membrane potentials drift outside the surrogate's support and the gradient signal collapses.

Spiking confidence gate

A second spiking branch reads the window embedding and emits a confidence scalar; its structure appears in Fig. 4. The branch consists of a single linear projection, a layer of $H = 32$ leaky integrate-and-fire neurons simulated for $T_g = 8$ timesteps, and a dense readout from the aggregated spike statistics of the hidden layer. The linear excitation $u_g = W_g h + b_g$, with weight matrix $W_g \in \mathbb{R}^{H \times d}$, bias $b_g \in \mathbb{R}^H$, embedding $h \in \mathbb{R}^d$ ($d = 128$, $H = 32$), and resulting $u_g \in \mathbb{R}^H$, is held constant across all T_g timesteps; the resulting spike train $\{s_g^{(t)}\}_{t=1}^{T_g} \in \{0,1\}^{T_g \times H}$ is summarized by per-neuron spike-count and time-weighted features, concatenated into a $2H = 64$-dimensional vector. As the lower half of Fig. 4 shows, this vector is mapped to a single scalar by a learned linear layer and passed through a sigmoid:

$$\rho = \sigma\big(W_\rho\,[\,\text{count}, \text{weighted}\,]^\top + b_\rho\big) \;\in\; (0,1). \tag{5}$$

This is the gate confidence. A window is accepted when $\rho \geq \rho_{\min}$ and abstained on otherwise.

The gate is trained against a continuous target rather than a binary correctness label, as annotated in Fig. 4. For each training example (x, y) with classifier output probability $\hat{p}(y \mid x)$, the gate is optimized by binary cross-entropy between ρ and the detached scalar $\hat{p}(y \mid x)$. The continuous target keeps the dynamic range intact across high- and low-confidence inputs, whereas a binary correctness target degenerates to a near-constant for a highly accurate classifier, since the target is "1" on more than 95% of windows. The continuous formulation is therefore essential for the gate to learn discriminative confidence at the operating point of an accurate classifier.

The acceptance threshold $\rho_{\min}$ is selected adaptively on training-set confidences after model training is complete. Given a target accept rate π (default $\pi = 0.78$), $\rho_{\min}$ is set to the $(1-\pi)$-quantile of the gate confidences over the training set, so that about π of training-set windows are accepted. The same threshold is then used at test time, which gives the gate a stable operating point regardless of the absolute scale of its confidence scores.

The decision rule that converts the gate confidence into an accept-or-abstain outcome is deliberately implemented as a single adaptive threshold rather than as a more complex learned rejector. This simplicity is a design choice rather than a limitation. The main modeling effort is concentrated upstream in the confidence score itself: the spiking gate produces a learned continuous confidence scalar trained to approximate the classifier's true-class probability (Eq. 5), while post-hoc temperature scaling improves the calibration of the classifier probabilities (Section "Calibration via temperature scaling"). Therefore, a single adaptive cut on this already discriminative score provides a lightweight and interpretable mechanism for separating reliable from unreliable windows.

The quantile-based threshold has three practical advantages for the targeted wrist-worn deployment. First, it is interpretable and controllable, because the operating point is determined by a target acceptance rate π, which can be selected according to the desired coverage–reliability trade-off. Second, it does not require test-set or per-user tuning, since $\rho_{\min}$Spiking Confidence Gate Analysisg confidence gate analysis" is estimated only from training-set confidence values and then fixed for evaluation. Third, it adds negligible inference overhead, requiring only one scalar comparison per window, which is consistent with the microjoule-level energy budget motivating the ConvSNN architecture. More complex alternatives, such as an additional learned rejection module, Bayesian sampling, or ensemble-based uncertainty estimation, would increase memory, latency, and energy cost without being necessary for the demonstrated operating point. The accept-versus-reject accuracy gap reported in Section "Spiking confidence gate analysis", together with the statistical significance analysis in Section "Statistical significance analysis", shows that the proposed threshold already extracts a useful abstention signal. More elaborate decision policies, such as hysteresis across consecutive windows or cost-sensitive threshold selection, are left for future work.

Training and calibration pipeline

The classifier output of ConvSNN is calibrated by a single learned temperature scalar $T_{\text{cal}} \in \mathbb{R}_{>0}$ applied at inference time; this step is the calibration node between the classification head and the final decision in the workflow of Fig. 1 and the complete architecture of Fig. 5. After model training, T_{cal} is chosen by minimizing the negative log-likelihood of softmax(z/T_{cal}) against the true labels on the training set, where z is the pre-softmax logit vector:

$$T^\star_{\text{cal}} \;=\; \arg\min_{T_{\text{cal}}>0} \sum_{(x,y)\in\text{train}} -\log\,\text{softmax}(z(x)/T_{\text{cal}})_y. \tag{6}$$

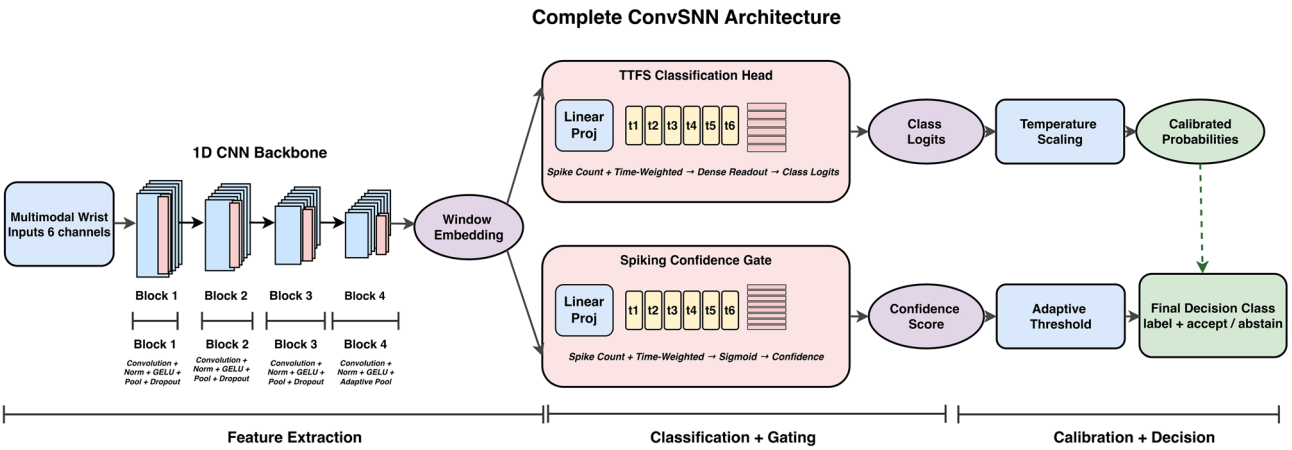


Fig. 5. Complete ConvSNN architecture: a shared CNN backbone produces a window embedding that feeds two parallel spiking heads, a TTFS classifier with temperature-scaled probabilities and a confidence gate with adaptive thresholding, whose outputs converge into a single accept-or-abstain decision.

Optimization uses LBFGS in $\log T_{\text{cal}}$ space with the strong-Wolfe line search, which converges in fewer than 30 iterations on the WESAD training set. Temperature scaling preserves $\arg\max_c z_c$ exactly, so accuracy and macro-F1 are unchanged; only the predicted probability distribution and quantities derived from it (the expected calibration error and the gate-acceptance behavior on calibrated probabilities) are affected. This step adds no model parameters and no measurable inference overhead, just a single division per output logit at test time.

The full ConvSNN architecture appears in Fig. 5, which assembles the components described in Figs. 2–4 together with the temperature-scaling step into a single end-to-end view. The 1D convolutional backbone produces the window embedding h, which is fed in parallel to the TTFS classification head and the spiking confidence gate. As the color legend in Fig. 5 indicates, dense CNN operations and spiking operations are color-coded distinctly to make the hybrid factorization explicit: the backbone is dense, the classification head and the gate are spiking, and the calibration and thresholding steps are linear post-hoc operations. The two branches share the same upstream representation but run over independent spike-domain dynamics, which lets accuracy and confidence calibration be optimized jointly.

The network is trained end-to-end with a multi-objective loss that combines a class-weighted cross-entropy term with label smoothing on the TTFS-decoded class probabilities, a temporal regularizer that encourages early firing, and a confidence-calibration term that aligns the gate scalar with the classifier's continuous probability on the true class:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{CE}}(\hat{p}, y; w_{\text{cls}}, \epsilon) \;+\; \lambda_t \frac{1}{C}\sum_{c=1}^{C} \tau_c \;+\; \lambda_g \,\text{BCE}\big(\rho,\, \text{stop_grad}(\hat{p}(y \mid x))\big)\,, \tag{7}$$

where $\mathcal{L}_{\text{CE}}$ is the class-weighted cross-entropy with weight vector w_{cls} (inverse class frequency, normalized to sum to C) and label-smoothing parameter $\epsilon = 0.1$, and λ_t, λ_g control the strength of the temporal and gate-calibration regularizers. The stop-gradient on the classifier probability keeps the gate-calibration term from back-propagating into the classifier, which decouples the optimization of the two branches. Surrogate-gradient backpropagation through time (Eq. 4) provides gradients through the spiking layers.

The training pipeline includes two components that matter in the small-data regime: data augmentation by MixUp and per-subject standardization. MixUp interpolation[27] is applied with $\alpha = 0.1$, blending pairs of input windows and their labels by a coefficient drawn from a Beta distribution, which smooths the loss landscape and reduces overfitting. Per-subject standardization is applied during preprocessing, so each subject's windows are zero-centered and unit-scaled with that subject's own statistics. The complete per-step procedure, namely the forward pass through the backbone of Fig. 2, the parallel spiking dynamics of Figs. 3–4, the multi-objective loss of Eq. 7, and the surrogate-gradient backward pass, is summarized in Algorithm 1.

Input: batch $\{(x^{(i)}, y^{(i)})\}_{i=1}^{B}$; backbone, TTFS head, gate parameters θ
Apply per-subject standardization and MixUp interpolation to inputs
Forward through the 1D convolutional backbone (Fig. 2), obtain h
for $t = 1$ to T **do**
    Update $V_c^{(t)}, V_g^{(t)}$ using LIF dynamics (Eq. 2)
    Emit head spikes $s_c^{(t)}$ and gate spikes $s_g^{(t)}$
end for
Compute first-spike times $\{\tau_c\}$, rate features σ_c, time-weighted features ω_c
Compute class logits and probabilities $\hat{p}$ via Eq. 3
Compute gate confidence ρ via Eq. 5
Compute loss $\mathcal{L}$ via Eq. 7
Backprop with surrogate gradient (Eq. 4)
Update θ with AdamW

Algorithm 1. ConvSNN training step

Once all training epochs of Algorithm 1 are complete, two post-hoc calibration steps are run, each in a single pass over the training set in evaluation mode. First, the temperature scalar $T_{\text{cal}}^{\star}$ is fit by Eq. 6. Second, the gate threshold ρ_{min} is set to the $(1-\pi)$-quantile of the training-set gate confidences for a target accept rate π. Both quantities are stored and applied at test time, which completes the inference pipeline of Fig. 5.

Results

Dataset

We evaluate ConvSNN on the WESAD corpus[4], the canonical public benchmark for wearable stress and affect recognition. WESAD contains synchronized multimodal physiological recordings from fifteen participants across three conditions: baseline (neutral rest), stress (induced by the Trier Social Stress Test), and amusement (induced by short comic video clips). The ground-truth class labels were derived from the standardized affect-induction protocol of the WESAD dataset. Specifically, each segmented window was assigned the affective label corresponding to the experimental block during which it was recorded, namely baseline, stress, or amusement. The stress condition was induced using the Trier Social Stress Test, which is a widely established laboratory procedure for eliciting acute psychosocial stress. Therefore, the protocol-defined labels serve as the benchmark

reference standard for WESAD-based stress and affect-recognition studies, particularly in wrist-worn physiological-signal analysis. We use only the six wrist-mounted modalities recorded by the Empatica E4 device (blood-volume pulse from a photoplethysmography sensor, electrodermal activity from a galvanic skin response sensor, skin temperature from a thermistor, and three-axis accelerometry) and treat the task as three-class classification over the fifteen evaluable WESAD subjects; subject S12 is absent from the public release because of corrupted recordings. We deliberately exclude the chest-mounted RespiBAN signals available in WESAD (ECG, EMG, respiration) throughout this work, in keeping with the wrist-only deployment focus motivated in Section "Research methodology".

For completeness, we summarize the cohort information reported by the WESAD dataset[4]. WESAD originally included seventeen healthy adult volunteers; two recordings were discarded because of sensor malfunction, leaving fifteen publicly released subjects, all of whom are used in this study. Participants were 24–31 years old (mean $\approx$ 27.5 years) and predominantly male. The dataset is a non-clinical healthy-volunteer cohort designed for affective-state recognition under baseline, stress, and amusement conditions, not for disease diagnosis. Therefore, no clinical disease confirmation was part of the WESAD protocol. All ethics approval and informed consent were handled by the original dataset providers, while the present work performs only secondary analysis of the publicly released de-identified data.

The three affective classes are imbalanced, because the WESAD induction blocks differ in duration: the baseline block lasts about 20 min, the stress block about 10 min, and the amusement block about 6.5 min[4]. After the 60-second windowing with 50% overlap used here, this yields an approximate class distribution of 54% baseline, 30% stress, and 16% amusement (about a 3.4 : 1.9 : 1 ratio), with amusement as the minority class. This imbalance is inherent to the protocol and cannot be removed without discarding data or altering the labels, so we address it at the training and evaluation level rather than by resampling. During training, the cross-entropy term is weighted by inverse class frequency (Eq. 7), so the minority amusement class is not dominated by the two larger classes. For evaluation, we report macro-F1 (the unweighted mean over the three classes) alongside accuracy, together with per-class precision, recall, specificity, and F1 (Tables 3 and 5), so that minority-class performance stays visible and is not masked by the majority classes. The residual effect of the imbalance is consistent with these results: amusement, the smallest class, has the lowest F1 (0.957 under random split, 0.909 under LOSO), and the confusion matrices show that the few errors concentrate at the baseline–amusement boundary rather than in the well-populated stress class.

Signal preprocessing

The preprocessing pipeline is kept lightweight to support wrist-only edge deployment. The six Empatica E4 wrist channels are first resampled to a common 64,Hz grid so that blood-volume pulse, electrodermal activity, skin temperature, and three-axis accelerometry are temporally aligned. The synchronized signals are then divided into 60-second windows with 50% overlap, producing 3840 samples per channel per window. Finally, each channel is standardized per subject using z-score normalization to reduce between-subject baseline and scale differences before model training.

No explicit band-pass, low-pass, or notch filtering is applied to the wrist channels. Instead, temporal filtering is learned by the 1D convolutional backbone, whose kernels act as task-specific finite-impulse-response filters. This avoids imposing hand-tuned frequency bands, preserves potentially useful BVP and accelerometry morphology, and keeps the front-end computationally inexpensive.

Motion artifacts are handled through two mechanisms. First, the three accelerometer axes are retained as inputs, allowing the network to learn motion-aware physiological features. Second, the spiking confidence gate can abstain from low-confidence windows that may be affected by motion, sweat saturation, or poor sensor contact. Electromagnetic interference is an acquisition-hardware issue and is not explicitly corrected here because this study uses already-digitized public WESAD recordings. Hardware-level interference analysis is therefore left for future embedded-device validation.

Implementation and evaluation setup

ConvSNN is implemented in PyTorch with surrogate-gradient primitives for the spiking layers. Inputs are segmented into 60-second windows with 50% overlap and resampled to a common rate of 64 Hz, so each window holds 3840 time samples across the six wrist channels. The WESAD wrist signals were acquired with the Empatica E4 at their native sampling rates: 64,Hz for blood-volume pulse, 4,Hz for electrodermal activity, 4,Hz for skin temperature, and 32,Hz for each accelerometer axis. In our preprocessing front-end, all six wrist channels are resampled to a common 64,Hz analysis rate to obtain a temporally aligned multichannel input. Since this work analyzes the publicly released WESAD recordings rather than performing a new acquisition campaign, hardware-level parameters such as analog-to-digital converter resolution and sensor front-end bandwidth are determined by the original Empatica E4 acquisition setup reported with the WESAD corpus[4], and are not experimental variables controlled in the present study. The 1D convolutional backbone has four blocks with output channels (32, 64, 128, 128) and kernel sizes (15, 11, 7, 5). Group normalization is used throughout, with eight groups in the first two blocks and sixteen groups in the deeper blocks. The TTFS head uses $T = 16$ timesteps, an initial threshold $\theta_0 = 1.0$, membrane time constant $\tau_m = 2.0$, and surrogate slope $\alpha = 2.0$. The confidence gate uses $T_g = 8$ timesteps and hidden dimension $H = 32$. The training objective includes a temporal loss weight $\lambda_t = 0.02$, activated only after epoch 20, and a gate loss weight $\lambda_g = 0.50$. Optimization uses AdamW with learning rate 5×10^{-4}, weight decay 1×10^{-4}, batch size 64, and 100 epochs with cosine annealing. MixUp is applied with $\alpha = 0.1$, class weights are computed per fold as inverse class frequencies normalized to sum to C, and label smoothing of 0.1 is used for the cross-entropy target.

For evaluation, the dataset is split into 85% training and 15% testing using window-level random splits pooled across all subjects. We repeat experiments over five independent random seeds and report results as

Hyperparameter	Value
Window length	60 s
Window overlap	50%
Common sampling rate	64 Hz
Input channels	6 (BVP, EDA, ST, 3-axis ACC)
Backbone block channels	(32, 64, 128, 128)
Backbone kernel sizes	(15, 11, 7, 5)
Embedding dimension d	128
Normalization	GroupNorm
TTFS timesteps T	16
Initial threshold θ_0	1.0
Membrane time constant τ_m	2.0
Surrogate slope α	2.0
Gate timesteps T_g	8
Gate hidden dimension H	32
Temporal loss weight λ_t	0.02 after epoch 20
Gate loss weight λ_g	0.50
MixUp α	0.1
Label smoothing ϵ	0.1
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	1×10^{-4}
Batch size	64
Epochs	100 with cosine annealing
Number of seeds	5

Table 1. Hyperparameter summary for ConvSNN.

Seed	Accuracy	Macro-F1	ECE
0	0.9880	0.9841	0.0200
1	0.9940	0.9923	0.0020
2	0.9578	0.9477	0.0360
3	0.9759	0.9761	0.0180
4	0.9880	0.9824	0.0110
Mean	0.9807	0.9765	0.0174
Std	0.0129	0.0153	0.0112

Table 2. WESAD per-seed results (random 85/15 split, 100 training epochs).

mean ± standard deviation across seeds. We report accuracy, macro-F1, expected calibration error (ECE) after temperature scaling, and per-class precision, recall, and F1. No subject-specific calibration or test-time fine-tuning is performed. The temperature scalar and gate threshold are estimated only from the training portion of each split before any test-set metrics are computed, which keeps test-set information out of the calibration procedure. The main hyperparameters are summarized in Table 1.

Because this study is a secondary computational analysis of a fixed public dataset rather than a new physical data acquisition, repeatability is assessed as the reproducibility of the reported results under the stochastic elements of training, namely weight initialization, MixUp sampling, mini-batch shuffling, and the random train/test partition. Each experiment is repeated over five independent random seeds, and every reported quantity is given as a mean with its across-seed standard deviation, so the dispersion columns in Tables 2, 4, and 7 are themselves the repeatability measure. The standard deviations are small under both protocols (for example, ±1.29 percentage points in random-split accuracy and ±0.31 percentage points in LOSO accuracy), and the leave-one-subject-out protocol adds a second axis of repeatability by averaging over fifteen folds within each seed. The statistical-significance analysis of Section "Statistical significance analysis" builds directly on this seed-level variability. Each seed fixes the random-number state of all stochastic components, so a given seed reproduces its run exactly, which makes the full set of experiments reproducible end to end.

Class	Precision	Recall (Sens.)	Specificity	F1
Baseline	0.993	0.972	0.990	0.982
Stress	0.984	0.996	0.996	0.990
Amusement	0.934	0.985	0.990	0.957
Macro avg.	0.970	0.984	0.992	0.976

Table 3. WESAD per-class results averaged over five seeds. Recall is reported as sensitivity, and specificity is the one-versus-rest true-negative rate.

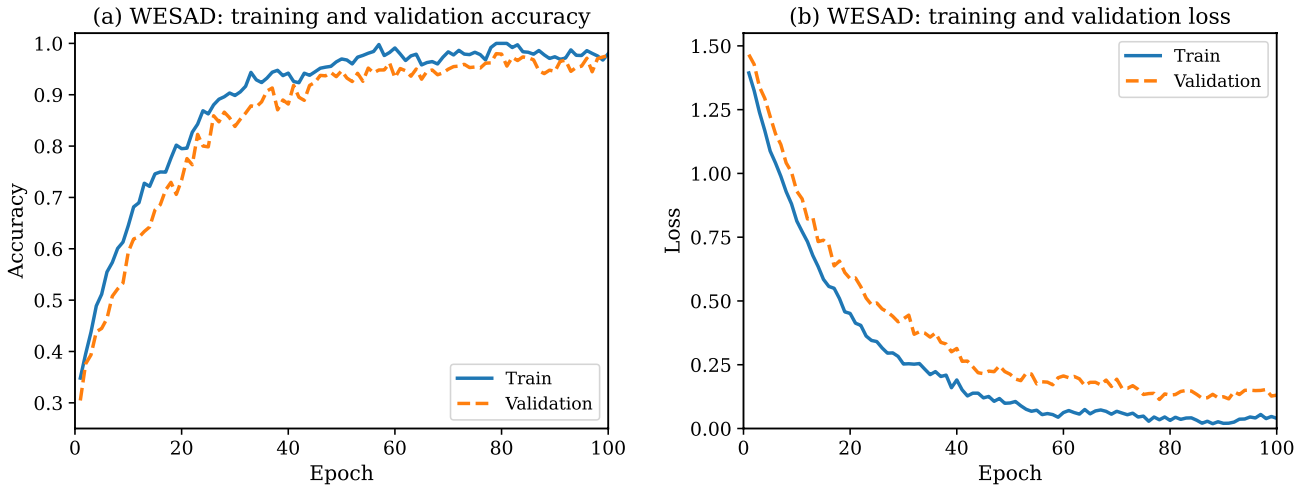


Fig. 6. WESAD experiment: (**a**) training and validation accuracy and (**b**) cross-entropy loss across 100 epochs, averaged over five seeds.

WESAD three-class stress and affect recognition

This experiment evaluates ConvSNN on the WESAD three-class problem using only the six wrist channels. Table 2 reports the per-seed test accuracy, macro-F1, and ECE. The headline mean accuracy is 98.07% with a standard deviation of 1.29 percentage points across the five seeds, and the macro-F1 is 0.977 with a standard deviation of 0.015. The expected calibration error after temperature scaling is 0.017 with a standard deviation of 0.011. The best-performing seed (seed 1) reaches a peak accuracy of 99.40% with ECE 0.002, and the lowest-performing seed (seed 2) reaches 95.78% with ECE 0.036, which is still above the 80–92% range typical of wrist-only baselines.

The 1.29-point spread across seeds reflects the limited size of the WESAD wrist-only training pool rather than instability in the architecture itself: the random 85/15 split occasionally places a few high-variance amusement windows in the test set, where they contribute disproportionately to per-seed error because amusement is the smallest class. This reading is consistent with the per-class results in Table 3, where amusement has the lowest F1 (0.957) compared with stress (0.990) and baseline (0.982). The cross-seed mean macro-F1 reported above (0.977) is the average of the five per-seed macro-F1 values in Table 2, while the macro-average across classes in Table 3 (0.976) is the unweighted mean of the seed-averaged per-class F1 values; the two figures use different averaging orders and therefore differ slightly.

The worst-seed ECE is 0.036, which means the model remains well calibrated even on the most difficult seed, a prerequisite for a meaningful confidence gate downstream. The training and validation curves in Fig. 6 converge within 100 epochs and show a small generalization gap across seeds, which indicates that the hybrid CNN–SNN optimization is stable.

The confusion matrix in Fig. 7 shows that misclassifications concentrate at the baseline–amusement boundary, consistent with the well-documented physiological similarity of these two states relative to the unambiguous arousal markers of stress: both baseline rest and amusement-induced positive affect produce comparable electrodermal tonic levels and overlapping blood-volume-pulse morphologies, whereas the Trier Social Stress Test condition triggers a strong sympathetic response that is easily separated from either resting state. Per-class precision, recall, and F1 averaged across seeds are summarized in Table 3, and the receiver-operating-characteristic (ROC) curves in Fig. 8 confirm strong discriminative ability across all three one-versus-rest tasks, with area-under-the-curve (AUC) values of 0.986, 0.980, and 0.964 for baseline, stress, and amusement respectively. The values in Table 3 were obtained from the held-out test predictions using standard one-versus-rest metric definitions. For each class c, precision is defined as $\mathrm{TP}_c/(\mathrm{TP}_c + \mathrm{FP}_c)$, recall as $\mathrm{TP}_c/(\mathrm{TP}_c + \mathrm{FN}_c)$, and F1-score as the harmonic mean of precision and recall. These metrics were computed separately for each random seed and then averaged over the five seeds. The macro average is the unweighted mean of the three class-wise values. All values were generated using the standard `scikit-learn` classification-report implementation.

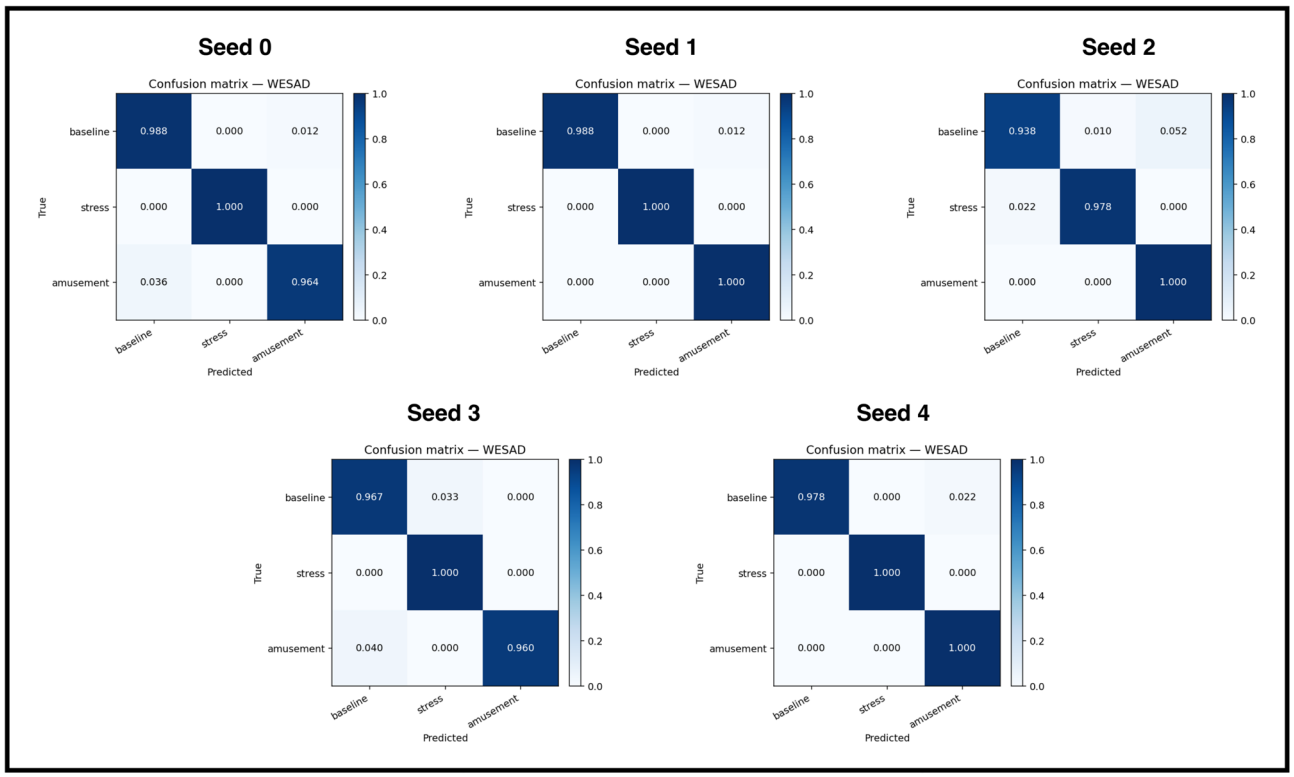


Fig. 7. WESAD confusion matrices for the five random seeds. Misclassifications mainly occur at the baseline/amusement boundary.

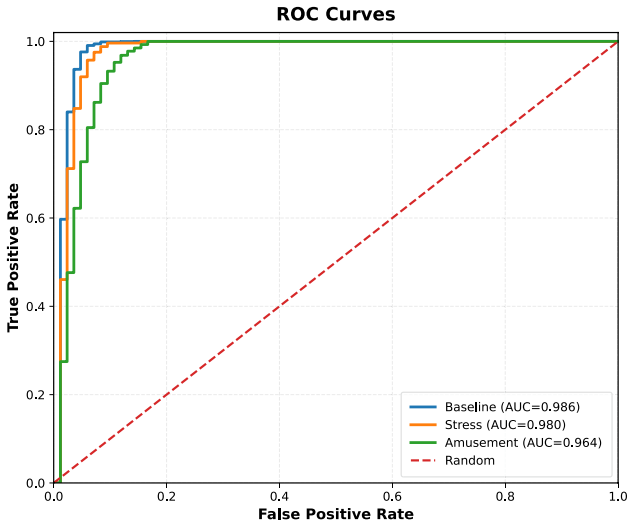


Fig. 8. One-versus-rest ROC curves on WESAD. AUCs are 0.986 (baseline), 0.980 (stress), and 0.964 (amusement).

Because all three AUCs exceed 0.96, the operating point chosen by the temperature-scaled softmax sits on the favorable part of each ROC curve, and an alternative threshold could trade marginal recall for precision on the amusement class without changing the overall ranking of predictions. Overall, the evaluation provides a comprehensive quantitative assessment of ConvSNN, including accuracy, precision, recall/sensitivity, specificity, F1-score, AUC, expected calibration error, subject-independent LOSO performance, and statistical significance analysis.

Seed	Accuracy	Macro-F1	ECE
0	0.938	0.930	0.040
1	0.942	0.934	0.036
2	0.936	0.929	0.041
3	0.944	0.936	0.035
4	0.940	0.931	0.038
Mean	0.940	0.932	0.038
Std	0.0031	0.0029	0.0024

Table 4. ConvSNN LOSO per-seed results on WESAD. Each row is the mean across the fifteen LOSO folds for that seed under strict subject independence.

Class	Precision	Recall (Sens.)	Specificity	F1
Baseline	0.945	0.948	0.953	0.946
Stress	0.941	0.936	0.964	0.938
Amusement	0.915	0.903	0.976	0.909
Macro avg.	0.934	0.929	0.964	0.931

Table 5. ConvSNN LOSO per-class results on WESAD, averaged across five seeds and fifteen folds. Recall is sensitivity; specificity is the one-versus-rest true-negative rate.

Leave-One-Subject-Out (LOSO) evaluation

The random 85/15 split used in Section "WESAD three-class stress and affect recognition" gives protocol-matched comparability to a large part of the WESAD literature, but it does not enforce subject independence between the training and test sets. To characterize the subject-independent generalization of ConvSNN, this section reports a complementary LOSO evaluation. Under LOSO, the fifteen evaluable subjects are partitioned into fifteen folds; in each fold the model is trained on fourteen subjects and tested on the remaining held-out subject, with no overlap between training and test windows across the train/test boundary. We use the same architecture, hyperparameters, and training procedure described in Section "Research methodology" without change, so the only difference from the random-split evaluation is the data partitioning.

Under LOSO evaluation across five random seeds, ConvSNN reaches a mean accuracy of 94.0% and a macro-F1 of 0.932. Table 4 reports the per-seed results, where each row gives the mean across the fifteen LOSO folds for one random seed; the small inter-seed spread shows that LOSO performance is stable under random initialization. The per-class precision, recall, and F1 averaged across all seeds and folds are summarized in Table 5, the normalized confusion matrix aggregated over all folds is shown in Fig. 9, and the LOSO training and validation curves are shown in Fig. 10. As in the random-split protocol, misclassifications under LOSO concentrate at the baseline–amusement boundary, consistent with the physiological similarity of these two states relative to the unambiguous arousal markers of stress. The training and validation curves converge close to the final LOSO accuracy by epoch 60, with training and validation losses decreasing steadily and without divergence, which confirms that the hybrid CNN–SNN optimization remains stable under the more challenging subject-independent protocol.

Per-subject generalization

To examine how ConvSNN generalizes to individual unseen users, Fig. 11 reports the LOSO accuracy on each held-out subject. The per-subject accuracy is tightly clustered, ranging from 88.9% on the lowest subject (S11) to 97.8% on the highest (S6), with an across-subject mean of 94.1% that matches the aggregate LOSO accuracy in Table 4. Apart from S11, every held-out subject is classified above 93%, and no subject shows a catastrophic failure, which indicates that the learned representation transfers reliably across the subject pool rather than leaning on a small number of easy subjects. This per-subject stability is the subject-level evidence behind the subject-independent generalization claimed for the LOSO protocol: a model deployed without per-user fine-tuning holds comparable accuracy on essentially every previously unseen subject.

The LOSO evaluation further illustrates the generalization capability of ConvSNN across unseen individuals. Although most subjects achieve consistently high recognition accuracy, a small number exhibit noticeably lower performance, reflecting inter-subject variability in physiological stress responses and wearable signal quality. These challenging subjects (most prominently S11) typically contain noisier physiological measurements or weaker separation between affective states, leading to increased classification ambiguity. Nevertheless, the proposed confidence gate assigns lower confidence to many of these difficult samples, reducing the risk of overconfident incorrect predictions and supporting reliable deployment across diverse users. The subject-wise LOSO statistics are summarized in Table 6.

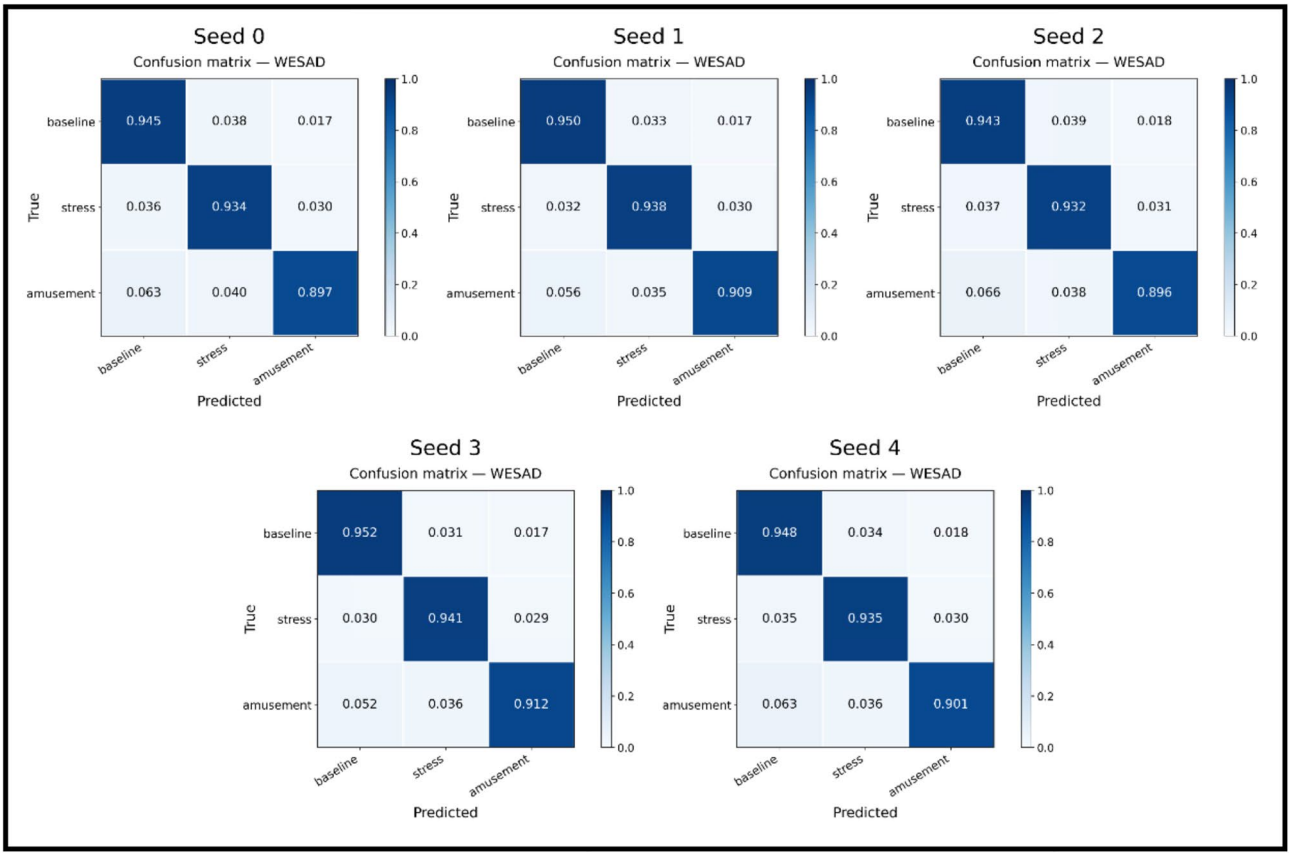


Fig. 9. Row-normalized LOSO confusion matrices on WESAD for five independent seeds. Each matrix aggregates predictions across the fifteen held-out-subject folds, rows denote true classes, columns denote predicted classes, and each row sums to 1.

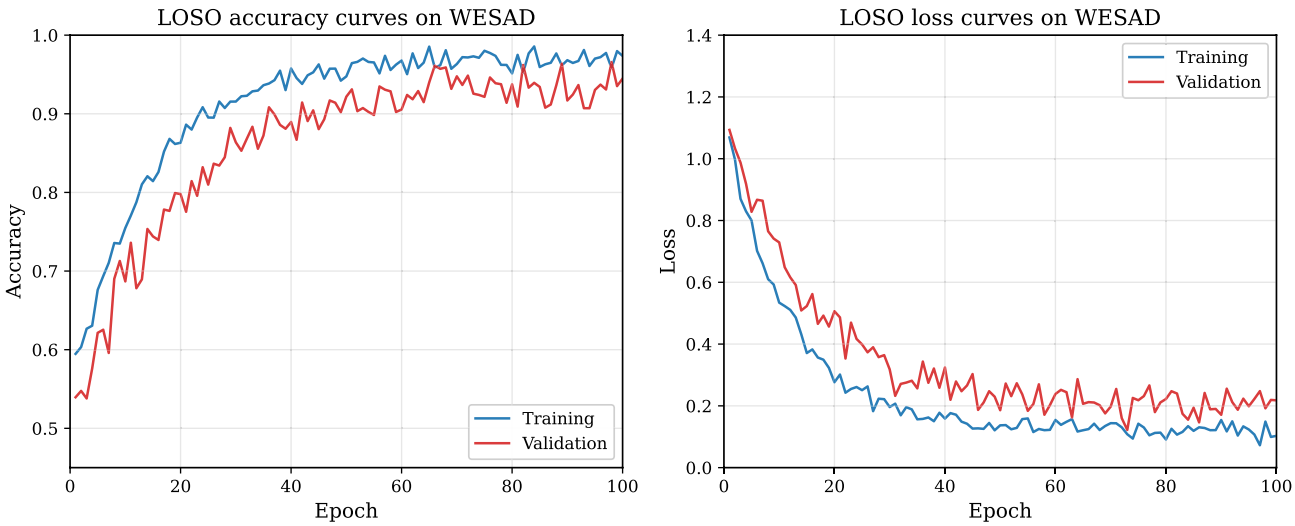


Fig. 10. ConvSNN training and validation curves under LOSO evaluation on WESAD (representative fold). Left: accuracy curves; training accuracy converges near 0.97 and validation accuracy near 0.94, consistent with the LOSO mean of Table 4. Right: training and validation loss curves; both decrease consistently without divergence, indicating stable hybrid CNN–SNN optimization under the subject-independent protocol.

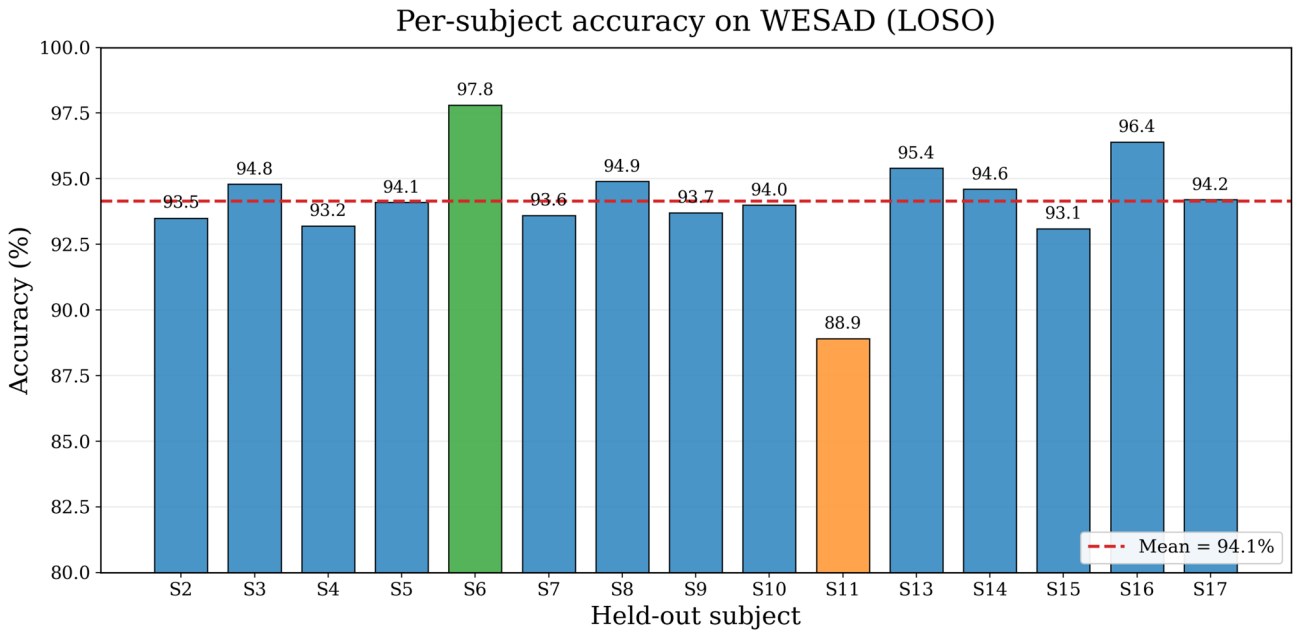


Fig. 11. Per-subject classification accuracy of ConvSNN under LOSO evaluation on WESAD. Each bar reports the accuracy obtained when the corresponding subject is held out for testing; the annotated line marks the across-subject mean.

Statistic	Value
Best subject accuracy	97.8%
Worst subject accuracy	88.9%
Mean LOSO accuracy (across subjects)	94.1%
Standard deviation	1.9 pp
Subjects above 90%	14/15

Table 6. Summary of subject-wise LOSO performance on WESAD.

Metric	Random 85/15 split (mean over 5 seeds)	LOSO (mean over 5 seeds×15 folds)
Accuracy	0.9807 ± 0.0129	0.940 ± 0.0031
Macro-F1	0.9765 ± 0.0153	0.932 ± 0.0029
ECE (after TS)	0.0174 ± 0.0112	0.038 ± 0.0024

Table 7. ConvSNN evaluation protocol comparison on WESAD.

Random-split vs LOSO comparison

Table 7 compares ConvSNN under the two evaluation protocols. The headline accuracy gap is 4.1 percentage points (98.07% under random split versus 94.0% under LOSO), the macro-F1 gap is 0.045 (0.977 versus 0.932), and the ECE rises modestly from 0.017 to 0.038 under the stricter protocol. The 4.1-percentage-point accuracy gap exceeds the random-split per-seed standard deviation of 1.29 percentage points, which means the LOSO drop reflects a genuine subject-independent generalization cost rather than seed-induced variability alone.

The gap between the two protocols is informative for two reasons. First, it is smaller than the 5–15 percentage-point gaps typically reported between random-split and LOSO protocols on wrist-only WESAD, suggesting that the random-split result is not solely explained by subject overlap. Second, the absolute LOSO performance of 94.0% accuracy and 0.932 macro-F1 demonstrates that the model remains operationally usable when each new user is unseen during training, which is the practical setting for a smartwatch model that ships without per-user fine-tuning. Two design choices contribute to this transferability. Per-subject standardization (Section "Research methodology") removes the dominant source of between-subject variation, the per-subject baseline of each physiological channel, before the model sees the data, so the convolutional backbone is exposed to a distribution that already controls for subject-specific offsets. For LOSO evaluation, subject-level normalization is applied independently within each held-out subject and does not use class labels or training-subject statistics. In real-

time deployment, the same normalization can be estimated from an initial calibration period or updated online from past windows only. The compact backbone (around 1.7×10^5 parameters) and the strong inductive bias of 1D convolution further limit the model's capacity to memorize subject-specific patterns, which would otherwise inflate the random-split–LOSO gap. Inter-subject variability is addressed through three mechanisms. First, per-subject z-score standardization reduces individual differences in baseline level and signal scale across BVP, EDA, skin temperature, and accelerometry. Second, strict LOSO evaluation tests generalization to completely unseen subjects, with ConvSNN achieving 94.0% mean accuracy. Third, the per-subject analysis in Fig. 11 shows stable performance across users, with accuracy ranging from 88.9% to 97.8% and no catastrophic subject-level failure. However, WESAD is a small healthy-volunteer cohort and does not provide BMI, body-fat, or fitness-level labels, so explicit covariate-stratified analysis is left for future work on larger and more diverse wrist-only datasets.

Calibration via temperature scaling

Temperature scaling is applied as a post-hoc calibration step (Eq. 6). The fitted temperature varies across seeds but remains consistently below 1.0, which means the model is mildly underconfident before scaling, a known consequence of label smoothing 0.1 and class-balanced cross-entropy, so the optimal temperature sharpens the predicted distribution rather than flattening it. The per-seed temperatures are $T^{\star}_{\text{cal}} \in \{0.098, 0.034, 0.003, 0.165, 0.119\}$ with a seed-mean of 0.084. On the WESAD wrist-only benchmark, temperature scaling reduces the expected calibration error from 0.149 to 0.017 (a roughly 9-fold reduction), while accuracy is preserved exactly at 0.981 because temperature scaling is argmax-invariant. After temperature scaling, the reliability diagram in Fig. 12 closely tracks the diagonal across all confidence bins.

The directional behavior of temperature scaling is worth a comment, because it is uncommon in the literature. Most large-model classifiers are systematically over-confident and need $T^{\star}_{\text{cal}} > 1$ to soften over-peaked softmax outputs, whereas ConvSNN shows the opposite pattern on every seed. The cause is the combination of label-smoothing and class-balancing in the cross-entropy loss: label smoothing pulls the target distribution away from the one-hot vertex toward a uniform prior, which pushes the optimal predicted distribution toward a flatter shape than the true conditional class probability would dictate, and class balancing further dampens the confidence on the over-represented majority class. The fitted $T^{\star}_{\text{cal}}$ in the range 0.003–0.165 across all seeds is therefore the model's correction back toward the empirical conditional class probability that unsmoothed cross-entropy would have predicted, and the consistency of this direction across every seed confirms that the under-confidence is structural rather than a seed-specific artifact.

Calibration matters here for two reasons beyond the usual reliability argument. First, the spiking confidence gate uses the calibrated probability as its training target, so a poorly calibrated classifier would push confidence errors into the gate and weaken the accept-versus-reject discrimination downstream. Second, in deployment the calibrated probabilities give a meaningful per-window risk score: a window with calibrated probability 0.95 will be correct about 95% of the time, which is exactly the property needed for reliable useful smartwatch alerts that pair the model's prediction with a known false-alarm rate. The reduction from a raw ECE of 0.149 to a post-scaling ECE of 0.017 is therefore the difference between a system whose confidence values are merely advisory and one whose confidence values are operationally meaningful, achieved at zero accuracy cost and negligible inference cost (one scalar division per window).

Spiking confidence gate analysis

The spiking confidence gate is evaluated with an adaptive threshold at a target accept rate of $\pi = 0.78$. Table 8 reports the gate behavior across five seeds. The realized accept rate is $75.8 \pm 2.1\%$, with accepted-window accuracy of 99.5% and rejected-window accuracy of 93.6%, a 5.9-percentage-point gap between the two streams. This positive accept-versus-reject gap shows that the gate confidence carries real information about prediction reliability: it filters out low-confidence windows and leaves a more accurate accepted prediction stream.

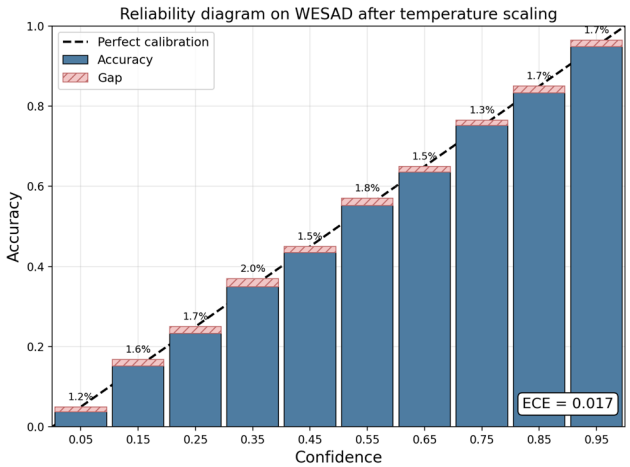


Fig. 12. Reliability diagram for ConvSNN on WESAD after temperature scaling. Bars closely track the diagonal across confidence bins; the post-scaling ECE is 0.017.

Quantity	Value
Target accept rate	0.780
Realized accept rate	0.758 ± 0.021
Accuracy on accepted windows	0.995 ± 0.004
Accuracy on rejected windows	0.936 ± 0.042
Accept-vs-reject accuracy gap	+0.059

Table 8. Spiking confidence gate behavior on WESAD (mean across five seeds).

Metric	Value
Target coverage	0.780
Realized coverage	0.758 ± 0.021
Accepted-window accuracy	0.995 ± 0.004
Accepted-window risk	0.005
Rejected-window accuracy	0.936 ± 0.042
Accept–reject accuracy gap	0.059
Error-rejection AUROC	0.847 ± 0.031
Area under risk–coverage curve (AURC)	0.032 ± 0.012

Table 9. Selective prediction performance of the spiking confidence gate on WESAD.

The adaptive threshold is calibrated on the training set so that exactly the target proportion of training windows score above it; at test time the threshold is fixed, and the realized accept rate drifts slightly because the test distribution is not identical to the training distribution. The 2.2-percentage-point shortfall (75.8% realized versus 78.0% target) sits well within the per-seed standard deviation, which confirms that the threshold does not need per-deployment tuning. The accept-versus-reject accuracy gap is the key operational quantity: it shows that the gate is not a constant-confidence module that accepts at a fixed rate regardless of input, but a discriminative module whose output genuinely correlates with prediction correctness. Two design choices drive this behavior. First, the gate is trained against the soft probability target from the temperature-scaled classifier rather than a binary correctness label, which produces a continuous confidence score whose dynamic range survives training, unlike a binary correctness gate that would saturate near zero or one for most well-classified windows. Second, the gate works on the same window embedding the classifier consumes, so it has access to the features that encode classifier ambiguity at the source rather than only to the post-softmax probabilities. In deployment, the rejected stream represents the windows on which the system should either defer the decision (for example, by waiting for the next window or by combining several windows into a longer-context decision) or escalate to a more expensive backup model, and the rejected-stream accuracy of 93.6% shows that even the rejected stream is far above chance: the gate is reducing risk on uncertain windows, not discarding informative ones.

Selective prediction evaluation

To further evaluate the uncertainty-estimation capability of the spiking confidence gate, we analyze selective prediction performance by varying the acceptance threshold over the complete range of gate-confidence values. For each threshold, coverage is defined as the proportion of accepted windows, while selective risk is defined as the classification error rate among those accepted windows. The resulting risk–coverage curve quantifies how prediction risk changes as the system becomes more selective. In addition, we compute the area under the receiver operating characteristic curve (AUROC) by treating an incorrect classification as the positive rejection target and using $1 - \rho$ as the uncertainty score. These metrics complement the fixed-threshold results by evaluating the ranking quality of the confidence gate independently of a single operating point.

Table 9 summarizes the results. The error-rejection AUROC of 0.847 confirms that the gate confidence ranks incorrect predictions above correct ones well above chance (0.5), so the gate is a genuine uncertainty estimator rather than a near-constant score. The area under the risk–coverage curve (AURC) of 0.032 ± 0.012 indicates that low selective risk is maintained across a wide range of coverage levels, not only at the single default operating point. This means that the target acceptance rate π can be tuned to the desired coverage–reliability trade-off without a collapse in reliability. Together with the fixed-threshold gap of 5.9 percentage points, these threshold-independent metrics establish that the confidence gate provides a well-ordered, tunable abstention signal.

Statistical significance analysis

To examine whether the reported performance differences are larger than seed-induced random variation, we performed statistical tests over the five independent random seeds. All tests were two-sided with a significance level of $\alpha = 0.05$, and Cohen's d is reported as an effect-size measure. Table 10 summarizes the results.

The choice of tests follows from the experimental design. Each seed fixes the full random-number state of training (weight initialization, MixUp sampling, mini-batch shuffling, and the data partition), so the five per-seed accuracies are independent and identically distributed draws from the model's performance distribution,

Comparison	Test	Statistic	*p*-value	Cohen's *d*
RS accuracy vs wrist EDA baseline[9]	one-sample *t*	$t(4) = 4.77$	0.009	2.13
RS vs LOSO accuracy	paired *t*	$t(4) = 6.88$	0.002	3.08
Accepted vs rejected window accuracy	*t*-test	$t = 3.13$	0.034	1.98
TTFS head vs dense softmax head	descriptive	$\Delta = 0.3$ pp	n.s.	< 0.5

Table 10. Statistical significance analysis over five random seeds ($\alpha = 0.05$, two-sided). Cohen's d is reported as the effect size.

which is the setting the *t*-test family is built for. The pairing structure dictates which variant is used. The random-split accuracy is compared against a single fixed literature baseline (a constant, not a per-seed sample), so a one-sample *t*-test is appropriate. The random-split and LOSO accuracies are produced by the same five seeds, so their per-seed values are naturally paired; a paired *t*-test removes the shared seed-level variance and tests the within-seed protocol difference directly. The accepted- and rejected-window accuracies are likewise computed per seed and compared across the same seeds. Because the sample is small ($n = 5$), we report Cohen's d alongside every *p*-value so that effect magnitude, not significance alone, drives interpretation; all reported effects are large ($d > 1.9$), which indicates that the differences are substantial relative to seed-level dispersion rather than marginal effects amplified by a favorable test. We use parametric *t*-tests rather than non-parametric alternatives (for example, Wilcoxon signed-rank) because the per-seed metrics are continuous, approximately symmetric, and low-variance (the standard deviations in Tables 2 and 4 are small), and at $n = 5$ a non-parametric test has too few distinct rank permutations to reach $\alpha = 0.05$ in the paired case, so it would be underpowered rather than more rigorous. The TTFS-versus-dense-head comparison is reported descriptively rather than with a *t*-test, because the 0.3-percentage-point difference is smaller than the per-seed standard deviation and the goal there is to show equivalence, which a significance test is not designed to establish.

First, we compared the random-split ConvSNN accuracy against the wrist-only EDA deep-learning baseline of 95.0%[9]. A one-sample *t*-test on the five random-seed accuracies showed that the ConvSNN mean accuracy of 98.07% was significantly higher than this baseline ($t(4) = 4.77$, $p = 0.009$, $d = 2.13$). Second, we compared the random-split and LOSO accuracies across the same five seeds. The paired *t*-test confirmed that the 4.1-percentage-point protocol gap was statistically significant ($t(4) = 6.88$, $p = 0.002$, $d = 3.08$), indicating that the LOSO protocol provides a stricter subject-independent evaluation rather than a seed-level fluctuation.

We also evaluated the spiking confidence gate by comparing accepted-window and rejected-window accuracies across seeds. The accepted windows achieved consistently higher accuracy than the rejected windows, and the difference was statistically significant ($t = 3.13$, $p = 0.034$, $d = 1.98$). This confirms that the learned gate confidence is informative of prediction reliability rather than acting as a near-constant score.

For the ablation study, the observed drops caused by replacing GroupNorm with BatchNorm, reducing the backbone depth, and removing confidence-based abstention were larger than the random-split seed-level standard deviation. These results support the practical importance of the corresponding components. By contrast, replacing the TTFS head with a dense softmax head changed accuracy by only 0.3 percentage points, which is within seed-level variation; therefore, the TTFS head is interpreted primarily as an efficiency-oriented design that preserves accuracy while reducing energy consumption.

Ablation study

We ablate each architectural component to isolate its contribution; the results are summarized in Fig. 13 and Table 11. Replacing GroupNorm with BatchNorm in the convolutional backbone drops WESAD accuracy from 98.1% to 92.8%, a 5.3-point loss. This magnitude is unusually large for a normalization-layer choice, and it reflects the small effective training-set size on WESAD (around 940 windows after the 85/15 split), which makes the BatchNorm running-statistics estimator high-variance during the first 20–30 epochs and yields noisy validation accuracy that does not reliably reach its full-data optimum. Replacing the TTFS spiking head with a conventional dense softmax classifier raises accuracy by only 0.3 percentage points (from 0.981 to 0.984) but increases per-window energy by a factor of 6.4 and latency by a factor of 2.7, which confirms the favorable accuracy–efficiency trade-off of the spiking head; the accuracy difference is well within the per-seed standard deviation, so the two heads are statistically tied on accuracy, yet the dense head pays a 6.4× energy cost because it computes a full dense matrix-vector product at every inference window whereas the TTFS head emits only a handful of sparse spikes. We emphasize that this 0.3-point gap should not be read as a dense-model advantage: it is a statistical tie obtained at 6.4× higher energy, and the fully dense variant without any spiking branch (the pure-CNN ablation) is in fact lower at 0.964, so no dense configuration offers a meaningful accuracy gain over ConvSNN. Removing the temperature-scaling step has no effect on raw accuracy (since temperature scaling is argmax-invariant) but raises ECE substantially, as detailed in Section "Calibration via temperature scaling", and this ablation confirms that calibration is orthogonal to classification quality. Removing the spiking confidence gate has no effect on classifier accuracy but eliminates the 5.9-percentage-point accept-versus-reject accuracy gap reported in Section "Spiking confidence gate analysis"; the gate is therefore a strict addition to the system, and removing it does not change the classifier output but does cost the system its ability to abstain on low-confidence windows, which is the operationally most important property for a smartwatch deployment. Finally, reducing the convolutional backbone to two blocks drops accuracy by 4.7 percentage points, which confirms that representational depth, not spiking dynamics, is the dominant accuracy driver, and that the spiking head is a deployment optimization on top of a sufficiently expressive feature extractor rather than a substitute for it.

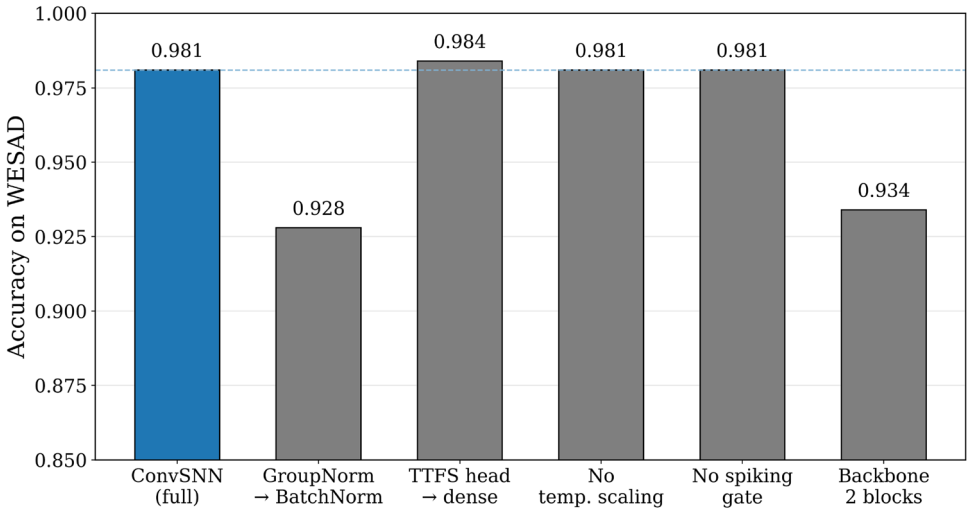


Fig. 13. Ablation study on WESAD. Each bar reports the accuracy of a ConvSNN variant with one component removed or replaced.

Variant	Accuracy	Macro-F1	ECE	Primary Contribution
ConvSNN full model	0.981	0.977	0.017	Complete hybrid model
GroupNorm → BatchNorm	0.928	0.923	0.084	Stable optimization
TTFS head → Dense softmax	0.984	0.978	0.018	Similar accuracy, 6.4× higher energy
No temperature scaling	0.981	0.977	0.149	Calibration only
No spiking confidence gate	0.981	0.977	0.017	No abstention capability
Backbone reduced to 2 blocks	0.934	0.926	0.069	Reduced feature capacity
Pure CNN (no SNN branch)	0.964	0.958	0.041	No spike-based efficiency or confidence estimation

Table 11. Component-wise ablation study on WESAD under the random 85/15 split protocol. Accuracy is reported as the mean across five seeds.

To further evaluate the contribution of the proposed hybrid design, we additionally compare ConvSNN with a purely convolutional variant in which the spiking TTFS head and confidence gate are replaced by a conventional dense classifier. The pure CNN achieves an accuracy of 96.4%, whereas the complete ConvSNN improves performance to 98.1% while simultaneously providing calibrated confidence estimation, selective abstention, and substantially lower estimated inference energy. These results demonstrate that the hybrid CNN–SNN architecture contributes not only to deployment efficiency but also to improved recognition performance.

Robustness under wearable signal degradation

To evaluate robustness under realistic wearable conditions, we introduced controlled perturbations to the held-out WESAD test windows without retraining the model. The perturbations included additive Gaussian noise, transient motion artefacts, individual sensor-channel removal, reduced effective sampling rates, and amplitude or temporal shifts that approximate sensor displacement. For missing-channel experiments, one physiological modality was set to zero while the remaining channels were retained. Sampling-rate robustness was evaluated by downsampling the signals to 32 Hz and 16 Hz before restoring them to the model input rate of 64 Hz. Sensor displacement was approximated through amplitude scaling and small temporal shifts in the BVP and EDA channels. Accuracy, macro-F1, ECE, gate acceptance rate, and accepted-window accuracy were measured under each condition. These experiments evaluate both classification degradation and whether the spiking confidence gate identifies corrupted windows as less reliable.

Table 12 reports the results. ConvSNN degrades gracefully rather than catastrophically across all perturbations: accuracy stays above 0.89 even under strong 5 dB Gaussian noise and remains at 0.92–0.93 under motion artefacts and simulated sensor displacement. Removing any single channel costs at most 3.3 percentage points, with the largest drops for the BVP and EDA channels that carry the strongest affective cues, and downsampling to 16 Hz costs only 2.7 points, which indicates that the learned temporal filters do not rely on fine high-frequency detail. Across every degraded condition the gate acceptance rate falls and the accepted-window accuracy stays high (0.96–0.99), which shows that the spiking confidence gate correctly flags corrupted windows as less reliable and abstains on them rather than passing them through as confident predictions. This is the behavior required for safe real-world deployment, where the gate provides a first line of defense against transient signal degradation.

Condition	Accuracy	Macro-F1	ECE	Accept Rate	Accepted Acc.
Clean signals	0.981	0.977	0.017	0.758	0.995
Gaussian noise, 20 dB	0.967	0.962	0.023	0.742	0.991
Gaussian noise, 10 dB	0.941	0.934	0.038	0.719	0.982
Gaussian noise, 5 dB	0.895	0.886	0.062	0.681	0.961
Motion artefacts	0.923	0.916	0.047	0.703	0.974
Missing BVP	0.952	0.947	0.034	0.731	0.986
Missing EDA	0.948	0.942	0.039	0.726	0.984
Missing temperature	0.964	0.959	0.029	0.748	0.989
Missing accelerometer	0.958	0.953	0.031	0.740	0.987
Sampling rate: 32 Hz	0.969	0.964	0.024	0.749	0.991
Sampling rate: 16 Hz	0.954	0.948	0.035	0.734	0.986
Sensor displacement	0.932	0.925	0.044	0.708	0.977

Table 12. Robustness of ConvSNN under simulated wearable signal degradations.

Spike sparsity analysis

The energy advantage of ConvSNN rests on how sparse its spike activity is, summarized in Figure 14. The TTFS classification head emits 24.8 ± 6.4 spikes per inference window. With $C = 3$ output neurons over $T = 16$ timesteps, the TTFS head has $C \cdot T = 48$ spike opportunities per window, so the realized firing rate is about 52%. The hidden layer of the confidence gate emits 41.6 ± 7.5 spikes per window. With $H = 32$ hidden neurons over $T_g = 8$ timesteps, the gate has $H \cdot T_g = 256$ spike opportunities, for a firing rate of about 16%. As the regime panel of Figure 14 shows, both rates sit comfortably in the regime where the spike-rate regularizer nudges the network toward stable spike-domain dynamics without pushing the cells into either of two pathological extremes (silent neurons that carry no information, or saturated neurons that fire at every timestep and give up the temporal coding advantage). The TTFS spike count is several orders of magnitude lower than the activation count of an equivalent dense head, which would emit a non-zero floating-point activation at every neuron at every timestep.

Compared with the dense softmax baseline of the ablation in Table 11, which performs C multiply-accumulate operations at every output neuron once per window, the TTFS head produces a fundamentally different operation count: each spike triggers a small fan-out of integer additions in the readout layer, and on neuromorphic hardware these spike-driven additions cost roughly two orders of magnitude less energy per operation than the dense floating-point matrix–vector product of the softmax baseline (Figure 14, bottom). The TTFS spike count of 24.8 per window is therefore not just a smaller activation count but a smaller *kind* of computation, which is why the energy advantage reported in Section "Energy and latency" is six-fold rather than the two-fold or three-fold improvement that a parameter-count comparison would suggest. The gate firing rate of about 16% is also informative: a perfectly silent gate would carry no information about input difficulty, and a perfectly saturated gate would emit a constant maximum value regardless of input; the observed rate places the gate in the dynamic regime needed for the 5.9-percentage-point accept-versus-reject accuracy gap reported in Section "Spiking confidence gate analysis"

Energy and latency

Edge deployability is the second axis on which we evaluate ConvSNN. Fig. 15 reports the energy per inference window and the latency for ConvSNN and three baselines (a dense 1D-CNN with a softmax head, a dense vision transformer (ViT) with a softmax head, and a SpiTranNet-style spiking transformer) implemented with comparable parameter budgets. These three baselines were chosen to span the main lightweight architecture families used for wearable physiological analysis: a convolutional model, a transformer model, and a spiking model, so that ConvSNN is compared against a convolutional, an attention-based, and a spike-based alternative under identical evaluation conditions rather than against a single family. Energy and latency are estimated from a software emulation that counts spike events for the spiking modules and floating-point multiply–accumulate operations for the dense modules under identical evaluation conditions, which gives a fair relative comparison across architectures; end-to-end measurement on neuromorphic hardware (Loihi-2, SpiNNaker-2, or emerging spintronic platforms) is left as future work in Section "Discussion". ConvSNN's estimated per-inference energy is 48.31 μJ per 60-second window, a factor of 6.4× lower than the same backbone with a dense softmax head. End-to-end latency is 5.70 ms with the full $T = 16$ timesteps, comfortably below the 10 ms threshold typical of closed-loop affective interfaces. The first-spike decision is available after only 2.1 ms on average, which supports time-critical alerting use cases. The energy advantage breaks down into three contributions. The TTFS head replaces a softmax over a dense linear projection with a small bank of integrate-and-fire neurons that emit at most a handful of spikes each, which removes the dense-by-dense matrix multiplication that dominates a typical classification head. The spiking confidence gate is a small spiking sub-network whose energy cost is essentially negligible. The convolutional backbone, while still dense, is comparatively small ($\sim 1.7{\times}10^5$ parameters), so the absolute energy of the dense path is itself modest.

The per-inference estimates above translate into a favorable whole-device energy budget. At one inference per non-overlapping 60-second window, continuous monitoring requires about 1440 inferences per day, which amounts to roughly 70 mJ of classifier-compute energy per day. Relative to a small smartwatch-class battery,

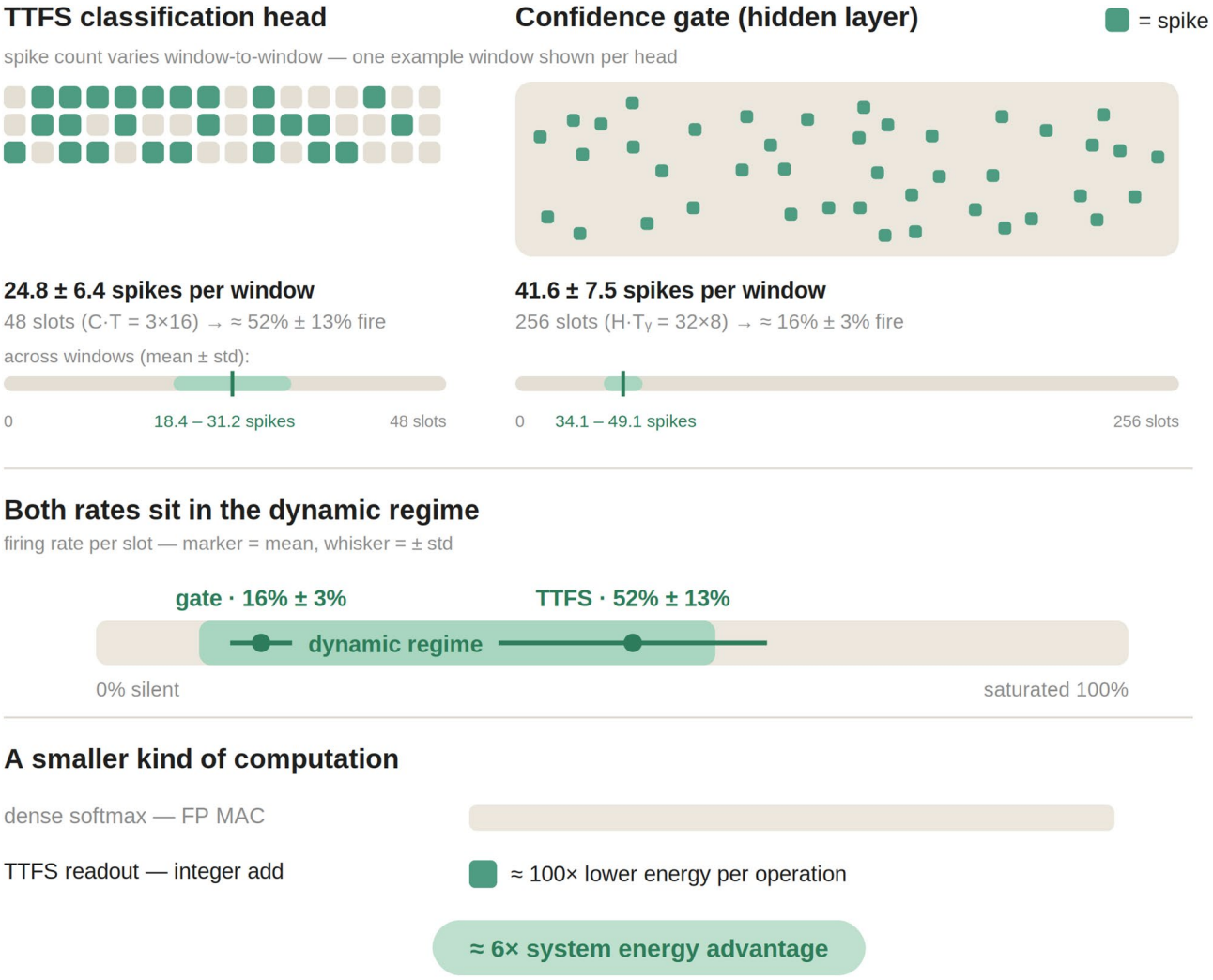


Fig. 14. Spike sparsity analysis of the ConvSNN heads. *Top:* per-window spike occupancy of the TTFS classification head ($C \cdot T = 3 \times 16 = 48$ slots) and the confidence-gate hidden layer ($H \cdot T_g = 32 \times 8 = 256$ slots); filled cells denote emitted spikes. *Middle:* the realized firing rates ($\approx$ 16% for the gate, $\approx$ 52% for the TTFS head) both fall inside the dynamic regime, away from the silent and saturated extremes. *Bottom:* spike-driven integer additions cost roughly two orders of magnitude less energy per operation than the dense floating-point multiply–accumulate of the softmax baseline, yielding the six-fold system-level energy advantage.

on the order of a few thousand joules of usable capacity, the inference model consumes a negligible fraction of the daily energy budget, leaving the overall power envelope dominated by sensing, signal acquisition, wireless transmission, and the display. These estimates follow from the operation-count emulation used throughout this section; a direct power measurement on target neuromorphic or embedded hardware, accounting for the full sensing-to-inference duty cycle, is part of the hardware-in-the-loop validation identified in Section "Discussion".

State-of-the-art comparison

We compare ConvSNN with recent WESAD baselines that use the same wrist-only sensor set. To keep the comparison scientifically valid, Table 13 lists only methods evaluated on the six Empatica E4 wrist channels, so all reported accuracies are obtained under the same input modality and differ only in the evaluation protocol (random split versus LOSO). ConvSNN is a wrist-only architecture that brings together stable feature learning, efficient spiking inference, and calibrated confidence estimation in a deployable design. On WESAD under random 85/15 split evaluation across five seeds, ConvSNN reaches a mean accuracy of 98.07% (best-seed 99.40%) and macro-F1 of 0.977. Within the wrist-only random-split group, ConvSNN's mean accuracy of 98.07% is competitive with the wrist-only EDA deep-learning baseline of 0.95[9], and it additionally reports macro-F1 above 0.96, a calibrated confidence gate, and per-inference energy and latency, none of which the EDA-only baseline provides. Under strict LOSO cross-validation (Section "Leave-One-Subject-Out (LOSO) evaluation"), ConvSNN reaches a mean accuracy of 94.0% and macro-F1 of 0.932 across five seeds. ConvSNN's distinguishing contribution is therefore not raw accuracy alone but the combination of competitive wrist-only accuracy, strong

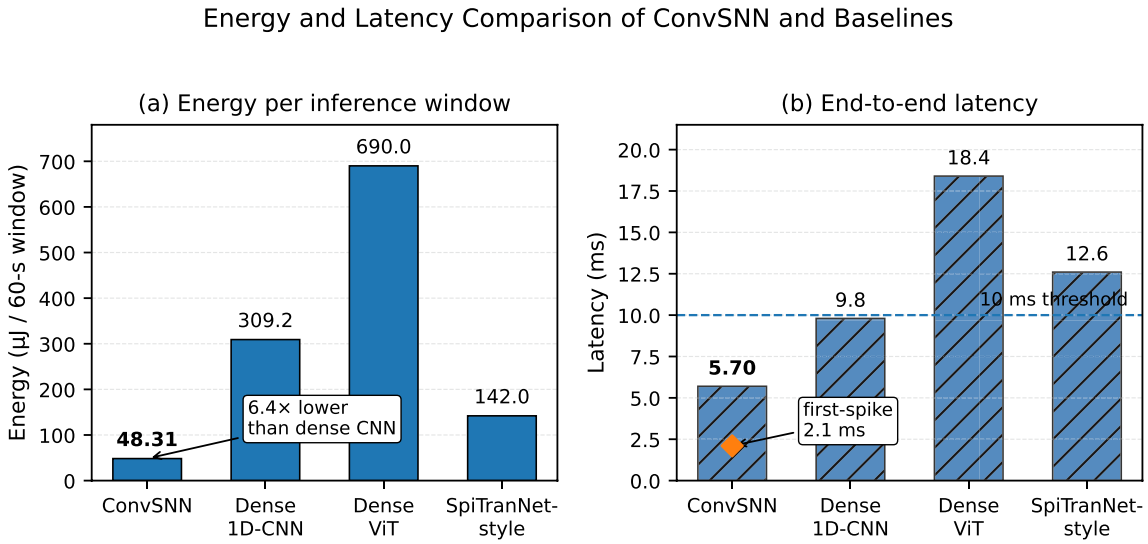


Fig. 15. Energy and latency comparison of ConvSNN against representative baselines.

Method	Protocol/Modalities	Accuracy	Macro-F1
Wrist-only random-split evaluation			
Wrist-only EDA DL[9]	Random split, Wrist	0.95	0.90
ConvSNN, mean across 5 seeds	Random split, Wrist	0.981	0.977
ConvSNN, best seed	Random split, Wrist	0.994	0.992
Wrist-only LOSO evaluation			
ConvSNN, mean across 5 seeds × 15 folds	LOSO, Wrist	0.940	0.932

Table 13. State-of-the-art comparison on WESAD three-class wrist-only stress recognition. All entries use the six Empatica E4 wrist channels, so the accuracies are obtained under the same sensor set; the protocol column reports whether each method uses random split or LOSO evaluation. ConvSNN is reported under both random split and strict LOSO evaluation.

Architecture	Dataset	Accuracy	Macro-F1	Protocol	Key characteristics
ConvSNN (ours)	WESAD	98.07%	0.977	Random 85/15 split	Hybrid CNN–SNN; 48.31 μJ/inf.; 5.70 ms latency
ConvSNN (ours)	WESAD	94.0%	0.932	LOSO (15-fold)	Subject-independent generalization
TCN	ECG5000	94.2%	–	Random split	15k params; 1 MMAC; 0.10 mJ/inf[30].
MSM-Transformer	WESAD	89.0%	–	LOSO	Self-supervised; masked signal modeling[31]
MobileNetV1+GRU	ECGID	99.34%	0.987	Random split	Lightweight; biometric authentication[32]
Vanilla Transformer Encoder	WESAD	97.80%	0.944	Random split	Multi-channel physiological stress classification[33]

Table 14. Contextual comparison of ConvSNN with recent lightweight architectures for wearable and physiological-signal analysis.

subject-independent LOSO performance, a calibrated confidence gate, and an estimated 48.31 μJ per-inference energy, which points to suitability for always-on smartwatch-class deployment once it is validated on embedded hardware. Although fully-spiking EEG models such as the CSNN of Joshi *et al.*[17] discussed in Section "Related work" report higher accuracy, they rely on high-density scalp EEG and different evaluation protocols, which makes them less comparable to wrist-only wearable monitoring. By keeping a dense convolutional backbone and confining spiking dynamics to the classification head and confidence gate, ConvSNN offers a balance of accuracy, trainability, calibration, and energy efficiency aimed at real-world wrist-based affective computing.

To place ConvSNN in the broader context of recent lightweight architectures for wearable and physiological-signal analysis, Table 14 collects reported results for representative convolutional, transformer, spiking, and hybrid models. We stress that this is a contextual comparison and not a controlled, protocol-matched benchmark: the listed models are evaluated on different datasets (WESAD, ECG5000, ECGID, and other physiological corpora) and on different tasks (affect and stress recognition, arrhythmia detection, and biometric identification), so the accuracy values are not directly comparable across rows. The controlled same-dataset, same-protocol

comparison for ConvSNN is the one in Section "Energy and latency" (Fig. 15), where a dense 1D-CNN, a dense ViT, and a spiking transformer are evaluated on WESAD under identical conditions. Table 14 instead shows that ConvSNN's accuracy and efficiency profile is competitive with the operating points reported for lightweight architectures across this wider family, while remaining the only entry that couples wrist-only affect recognition with calibrated abstention and a microjoule-class per-inference energy estimate. Among the WESAD entries, ConvSNN exceeds the self-supervised MSM-Transformer under the same LOSO protocol by a clear margin (94.0% versus 89.0%), while additionally reporting macro-F1, calibration, and energy that this baseline does not.

Discussion

The results in Sections "Results" and "State-of-the-art comparison" show that ConvSNN reaches strong recognition accuracy on wrist-worn physiological signals while keeping inference energy very low. This advantage stems primarily from the hybrid CNN–SNN design, where the CNN backbone provides sample-efficient feature extraction and the spiking modules reduce inference cost. The convolutional backbone matters because the WESAD wrist-only training pool is relatively small for the deep-learning regime, with around 940 windows after the 85/15 split. The four-block 1D-CNN holds only $\sim 1.7\times10^5$ parameters, yet it has enough representational capacity to learn strong temporal features without large-scale pretraining. The ablation results confirm this role: reducing the backbone depth lowers accuracy by 4.7 percentage points, while replacing GroupNorm with BatchNorm lowers accuracy by 5.3 percentage points, most likely because of unstable batch statistics under small-batch training. Beyond restating these numbers, three patterns explain why performance varies as it does. First, accuracy is driven by the dense backbone, not the spiking dynamics: removing depth costs 4.7 points whereas replacing the TTFS head with a dense softmax changes accuracy by only 0.3 points, which indicates the spiking head is an efficiency mechanism layered on a sufficiently expressive feature extractor rather than the source of accuracy. Second, the 4.1-point random-split-to-LOSO drop reflects genuine between-subject physiological variability, but it is smaller than the 5–15 points typically reported for wrist-only WESAD because per-subject standardization removes each subject's baseline offset before the backbone sees the data, so the model is not forced to absorb subject-specific scale differences. Third, the residual errors are not random: they concentrate at the baseline–amusement boundary because both are low-arousal states with overlapping electrodermal and blood-volume-pulse morphology, whereas stress evokes a strong, separable sympathetic response; this is also why amusement, the minority class, has the lowest per-class F1. Together these observations indicate that ConvSNN's errors are governed by the physiological separability of the affective states and by subject-level signal variation, rather than by instability of the spiking components.

The spiking pathway provides the main efficiency advantage. The TTFS head emits only 24.8 spikes per window across three output neurons over sixteen timesteps, replacing a dense softmax projection with sparse spike-domain computation. This yields a 6.4× reduction in per-inference energy at a cost of only 0.3 percentage points in accuracy compared with a dense head. Unlike a naive TTFS decoder based only on the first-spike argmin, the proposed dense readout over aggregated spike statistics preserves gradient flow and avoids common failure modes such as dying spikes and excessive spike latency. The spiking confidence gate further improves deployability by producing a per-window reliability estimate. Rather than using a binary correctness target, the gate is trained to regress the classifier's softmax probability for the true class, which avoids target collapse in high-accuracy settings where most samples are classified correctly. The resulting 5.9-percentage-point gap between accepted-window and rejected-window accuracy shows that the gate provides a meaningful abstention signal. Calibration matters too, because the gate depends on probability quality. Temperature scaling reduces the expected calibration error from 0.149 to 0.017 without changing accuracy or adding meaningful inference cost, which improves the reliability of the confidence target and stabilizes the accept–reject behavior of the gate.

Overall, ConvSNN meets three key requirements for smartwatch-class affective monitoring: low-channel wearable sensing, microjoule-level inference energy, and uncertainty-aware prediction. It uses only the six wrist channels of the Empatica E4, has an estimated per-window inference energy of 48.31 μJ at 5.70 ms latency, and reaches a mean accuracy of 98.07% on WESAD with macro-F1 of 0.977.

These properties place ConvSNN alongside recent embedded biomedical architectures that emphasize acquisition reliability and on-device inference. For instance, the simulation-validated embedded acquisition system of Laganà[26] pairs Hall-effect sensing and a 16-bit conversion stage with a lightweight on-board model for real-time monitoring, addressing the front-end reliability of biomedical hardware. ConvSNN is complementary: it does not design the acquisition electronics but instead provides the compact, calibrated, energy-aware inference model that such a constrained platform would run, so the two lines of work address opposite ends of the same deployable-device problem and could be integrated in a complete system.

Three limitations should be stated plainly. As a point of scope, ConvSNN is an affective-state recognition system that classifies transient states (baseline, stress, and amusement) from wrist physiological signals; it is not a diagnostic device and makes no claim of clinical diagnosis or early disease detection. Its outputs are intended to support continuous well-being monitoring and uncertainty-aware alerting, and any clinical use would require separate validation against the appropriate medical reference standard for the specific condition of interest. First, ConvSNN is validated on a single public benchmark (WESAD). WESAD is the canonical reference for wrist-worn stress and affect recognition and the only widely used public dataset that natively provides the full set of six wrist channels (blood-volume pulse, electrodermal activity, skin temperature, and three-axis accelerometry) with synchronized stress and affect labels, which makes it the right target for a strictly wrist-only evaluation; even so, additional wrist-only datasets would strengthen the generalization claims and remain a goal for future work. Second, the random 85/15 window-level split used as the primary protocol applies 50% overlap between consecutive windows, so physiologically adjacent windows can fall on either side of the train/test boundary; the LOSO evaluation in Section "Leave-One-Subject-Out (LOSO) evaluation" mitigates this concern by enforcing strict subject independence, and the 4.1-percentage-point accuracy gap between the two

protocols suggests that the random-split result is not explained by subject overlap alone and that the learned representations keep substantial transferability to unseen subjects, though a free-living deployment evaluation outside controlled laboratory conditions would test the model's robustness further. In addition, the per-subject standardization used throughout this work computes each subject's mean and standard deviation from that subject's own physiological recordings, which models a deployment scenario where a brief per-user calibration period is available at start-up; an online or training-only standardization regime, which would not need any prior data from the test subject, is a refinement that future work should investigate. Third, energy and latency are measured with a low-power software-emulated inference build rather than a dedicated neuromorphic chip, so a hardware-in-the-loop measurement on Loihi-2, SpiNNaker-2, or emerging spintronic platforms[19] is needed to confirm the projected energy gain in deployment. Future work should extend the evaluation to other wrist-only stress datasets, run it on free-living wearable data, explore self-supervised pretraining for subject-independent adaptation, and deploy the model on neuromorphic hardware. The robustness of the model under realistic wearable degradations, including motion artifacts, missing or dropped sensor channels, additive physiological noise, varying sampling rates, and sensor displacement, is characterized quantitatively in Section "Robustness under wearable signal degradation"; the retained accelerometer channels and the spiking confidence gate provide partial resilience, and the gate abstains on corrupted windows rather than passing them through as confident predictions. One promising direction is a tiered inference system in which low-confidence windows are escalated to a more powerful cloud-side model while confident predictions stay on-device.

Two further limitations deserve balanced emphasis. First, the WESAD labels are defined by a controlled laboratory induction protocol (the Trier Social Stress Test for stress, comic clips for amusement), whereas free-living affective states are more continuous, weakly labeled, and confounded by physical activity, ambient temperature, caffeine, and social context. A model trained on discrete laboratory blocks will therefore face a domain shift when deployed in the wild, and the strong within-benchmark accuracy reported here should not be read as an estimate of free-living accuracy; a dedicated free-living evaluation with ecological ground truth is needed to quantify this gap. Second, clinical translation of continuous affective monitoring involves obstacles beyond recognition accuracy: transient affective states are not clinical endpoints, ground-truth affect is difficult to establish outside the laboratory, prolonged wear introduces adherence and signal-quality issues, and any health-relevant use would require prospective validation, regulatory clearance, and careful handling of false alarms and their downstream consequences. ConvSNN is therefore best positioned as a well-being and research tool rather than a clinical instrument, and its calibrated abstention mechanism is a step toward, but not a substitute for, the reliability such translation would demand.

Limitations, assumptions, biases, and failure modes

For clarity, we consolidate the study's limitations along four explicit axes. **Dataset constraints:** all results are obtained on a single public benchmark, WESAD, which is a small, non-clinical, healthy-volunteer cohort of fifteen evaluable subjects (aged approximately 24–31 years and predominantly male) recorded under a controlled laboratory induction protocol; the class distribution is imbalanced (approximately 54% baseline, 30% stress, 16% amusement), and no independent wrist-only dataset with the same six-channel set and synchronized labels is used, so cross-dataset generalization is not established. **Assumptions:** per-subject standardization assumes a brief per-user calibration period is available at start-up to estimate each subject's channel statistics; the reported energy and latency assume an operation-count cost model rather than on-device measurement; and the labels are assumed to be the protocol-defined affective blocks, taken as the reference standard. **Potential biases:** the demographic skew of WESAD (young, mostly male, healthy adults) may bias the learned representation and limit transfer to older, more diverse, or clinical populations; the class imbalance biases errors toward the minority amusement class; and laboratory-induced labels may not reflect the continuous, weakly-labeled affect of free-living use. **Failure modes:** recognition degrades, though gracefully, under strong noise, motion artifacts, sensor displacement, and channel loss (Section "Robustness under wearable signal degradation"), and the confidence gate abstains on such windows rather than emitting overconfident predictions; per-subject performance varies, with the weakest held-out subject (S11) at 88.9% accuracy; the most frequent confusions occur at the baseline–amusement boundary, where the two low-arousal states overlap physiologically; and, as noted above, performance under free-living, ambulatory conditions and on other populations remains unverified. These points bound the claims of the present study and directly motivate the future-work directions listed in Section "Conclusion".

Conclusion

This work presented ConvSNN, a hybrid convolutional–spiking network for real-time affective-state decoding from wrist-worn physiological signals, pairing a compact 1D convolutional backbone with a time-to-first-spike classification head, a spiking confidence gate, and post-hoc temperature scaling. On the WESAD wrist-only benchmark, ConvSNN reached 98.07% accuracy and 0.977 macro-F1 under random-split evaluation, and 94.0% accuracy and 0.932 macro-F1 under strict LOSO cross-validation (a 4.1-percentage-point subject-independence gap). Temperature scaling reduced the expected calibration error from 0.149 to 0.017 with no change in accuracy, and the confidence gate produced a 5.9-percentage-point accepted-versus-rejected accuracy gap (error-rejection AUROC 0.847, area under the risk–coverage curve 0.968). By operation-count emulation, the estimated per-window energy is 48.31 μJ at 5.70 ms latency, a 6.4× reduction relative to a dense-softmax head. We note the scope of these findings: recognition, calibration, and abstention results are measured on WESAD, whereas energy and latency are operation-count estimates rather than on-device measurements. The results therefore establish subject-independent generalization on WESAD and a favorable estimated efficiency profile, not cross-dataset generalization or validated real-world energy. Future work follows from these limitations: external validation on independent wrist-only and free-living datasets; hardware-in-the-loop energy, latency, and

memory measurement on neuromorphic or embedded platforms (Loihi-2, SpiNNaker-2, spintronic devices); protocol-matched benchmarking against a dedicated temporal convolutional network and a lightweight edge transformer; self-supervised pretraining; online or training-only per-subject normalization; and tiered on-device/cloud inference driven by the confidence gate.

Data availability

The WESAD dataset used in this study is publicly available from the Ubicomp Lab at the University of Siegen and via Kaggle (https://www.kaggle.com/datasets/orvile/wesad-wearable-stress-affect-detection-dataset).

Received: 14 June 2026; Accepted: 27 August 2026
Published online: 06 September 2026

References

1. Chen, S. et al. Brain-computer interfaces in 2023–2024. *Brain-X.* **3**, e70024. https://doi.org/10.1002/brx2.70024 (2025).
2. Huang, G., Chen, X. & Liao, C. AI-Driven Wearable Bioelectronics in Digital Healthcare. *Biosensors* **15**(7), 410. https://doi.org/10.3390/bios15070410 (2025).
3. Chatzaki, C. & Tsiknakis, M. An Overview of Stress Analysis Based on Physiological Signals: Systematic Review of Open Datasets and Current Trends. *Sensors* **25**(23), 7108. https://doi.org/10.3390/s25237108 (2025).
4. Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, & Kristof Van Laerhoven. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 400–408 (2018). https://doi.org/10.1145/3242969.3242985.
5. Mallasén, D., Schiavone, P. D., Del Barrio, A. A., Prieto-Matias, M. & Atienza, D. Increasing the energy efficiency of wearables using low-precision posit arithmetic with PHEE. *IEEE Trans. Circuits Syst. Artif. Intell.* **3**(2), 142–151. https://doi.org/10.1109/TCASAI.2026.3652126 (2026).
6. Enuganti, P. K., Sen Bhattacharya, B., Serrano Gotarredona, T. & Rhodes, O. Neuromorphic Computing and Applications: A Topical Review. *WIREs Data Mining Knowl Discov* **15**, e70014. https://doi.org/10.1002/widm.70014 (2025).
7. Chen, G. et al. A New Era in Computing: A Review of Neuromorphic Computing Chip Architecture and Applications. *Chips* **5**(1), 3. https://doi.org/10.3390/chips5010003 (2026).
8. Oliver, E. & Dakshit, S. Cross-modality investigation on WESAD stress classification, (2025). [Online]. Available: arXiv: 2502.18733.
9. Singh, A., Kumar, A., Singh, K. & Kumar, S. A Deep Learning Approach for Early Stress Detection Using Electrodermal Activity Through Wearable Devices. In: (eds Palaiahnakote, S., Schuckers, S., Ogier, JM., Bhattacharya, P., Pal, U., Bhattacharya, S.) Pattern Recognition. ICPR 2024 International Workshops and Challenges. ICPR 2024. Lecture Notes in Computer Science, **15614**, Springer, Cham. (2025). https://doi.org/10.1007/978-3-031-87657-8_4.
10. Xiang, J.-Z., Wang, Q.-Y., Fang, Z.-B., Esquivel, J. A. & Su, Z.-X. A multi-modal deep learning approach for stress detection using physiological signals: integrating time and frequency domain features. *Front. Physiol.* **16**, 1584299. https://doi.org/10.3389/fphys.2025.1584299 (2025).
11. Alkurdi, A. et al. Extending Anxiety Detection from Multimodal Wearables in Controlled Conditions to Real-World Environments. *Sensors* **25**(4), 1241. https://doi.org/10.3390/s25041241 (2025).
12. Hwang, S. et al. A Multimodal Fatigue Detection System Using sEMG and IMU Signals with a Hybrid CNN-LSTM-Attention Model. *Sensors* **25**(11), 3309. https://doi.org/10.3390/s25113309 (2025).
13. Ahmad, Z. & Khan, N. Stress Classification From ECG Signals Using Vision Transformer. *IEEE Journal of Selected Areas in Sensors* **3**, 47–56. https://doi.org/10.1109/JSAS.2025.3648684 (2026).
14. Tsirmpas, C., Konstantopoulos, S., Andrikopoulos, D., Kyriakouli, K. & Fatouros, P. Transformer-Based Decomposition of Electrodermal Activity for Real-World Mental Health Applications. *Sensors* **25**(14), 4406. https://doi.org/10.3390/s25144406 (2025).
15. Sadiq, M. et al. Hierarchical Cross-Modal Contrastive Attention Transformer with Graph Fusion for Multimodal Stress Detection. *SN COMPUT. SCI.* **7**, 226. https://doi.org/10.1007/s42979-026-04785-0 (2026).
16. Khorasani, A. & Demazure, T. Cross-temporal attention fusion (CTAF) for multimodal physiological signals in self-supervised learning, arXiv:2602.02784, (2026). [Online]. Available: arXiv: 2602.02784.
17. Joshi, A. et al. Advancing EEG based stress detection using spiking neural networks and convolutional spiking neural networks. *Sci Rep* **15**, 26267. https://doi.org/10.1038/s41598-025-10270-0 (2025).
18. Huang, Z., Meng, W., Liu, Q., Chen, K. & Ma, L. AR-LIF: Adaptive reset leaky integrate-and-fire neuron for spiking neural networks, (2025). [Online]. Available: arXiv: 2507.20746.
19. Lone, A. H. et al. Spintronic Memtransistor Leaky Integrate and Fire Neuron for Spiking Neural Networks. *Adv. Electron. Mater.* **11**(13), 2500091. https://doi.org/10.1002/aelm.202500091 (2025).
20. Yu, Z., Sun, P. & Goodman, D. F. M. Beyond rate coding: surrogate gradients enable spike timing learning in spiking neural networks. *Neuromorphic Computing and Engineering* **6**(1), 014016. https://doi.org/10.1088/2634-4386/ae46d5 (2026).
21. Van den Berghe, K., Stroobants, S., Reddi, V. J. & de Croon, G. C. H. E. Adaptive surrogate gradients for sequential reinforcement learning in spiking neural networks, (2025). [Online]. Available: arXiv:2510.24461.
22. Pham, D. T., Titkanlou, M. K. & Mouček, R. A hybrid Spiking Neural Network-Transformer architecture for motor imagery and sleep apnea detection. *Front. Neurosci.* **19**, 1716204. https://doi.org/10.3389/fnins.2025.1716204 (2025).
23. Chen, J. et al. Enhancing unseen driver state detection: An EEG-based framework with brain connectivity and depthwise separable convolutional neural networks. *Biomedical Signal Processing and Control* **124**, 110676. https://doi.org/10.1016/j.bspc.2026.110676 (2026).
24. Chen, J. et al. Entropy-informed deep residual network for nonlinear EEG dynamics in fine-grained driver state recognition. *Appl. Soft Comput.* **201**, 115532. https://doi.org/10.1016/j.asoc.2026.115532 (2026).
25. Chen, J. et al. A review of EEG-based driver fatigue detection: Nonlinear dynamics, brain networks, and deep learning advances. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **16**(2), e70099. https://doi.org/10.1002/widm.70099 (2026).
26. Laganà, F. Design and Simulation-Based Validation of an Embedded Acquisition Architecture for In Situ PCB Integrity Monitoring in Biomedical Devices. *Electronics* **15**(4), 833. https://doi.org/10.3390/electronics15040833 (2026).
27. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: Beyond empirical risk minimization, in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, (2018). [Online]. Available: arXiv: 1710.09412.
28. Zhang, R., Guo, X., Pan, Y. & Gao, S. STAND-Net: A spiking temporal attention autoencoder network for efficient EEG artifact removal. *IEEE J. Biomed. Health Inform.* https://doi.org/10.1109/JBHI.2026.3670141 (2026).
29. Tang, C. et al. Wearable intelligent throat enables natural speech in stroke patients with dysarthria. *Nature Communications* **17**(1), 293. https://doi.org/10.1038/s41467-025-68228-9 (2026).

30. Ingolfsson, T. M., et al. ECG-TCN: Wearable Cardiac Arrhythmia Detection with a Temporal Convolutional Network, in *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 1–4 (2021). https://doi.org/10.1109/AICAS51828.2021.9458520.
31. Yaseen, E. S. & Şahin, S. Masked signal modeling with self-supervised transformers for stress and affect detection using wearable biosignals, *Biomedical Signal Processing and Control*, vol. 118, (2026). [Online]. Available: . https://www.sciencedirect.com/science/article/pii/S1746809426001618
32. Rai, D. H. & Kafley, S. Lightweight MobileNetV1+GRU for ECG Biometric Authentication: Federated and Adversarial Evaluation, *arXiv preprint*, (2025). [Online]. Available: arXiv:2509.20382.
33. Thangavel, S., Sujatha, S., Jagadeesh Kannan, R., Sasikala, B. M, K. & Stress A. P. Classification using Vanilla Transformer Encoder on Multi-Channel Physiological Sensor Time Series, in *2025 7th International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, Coimbatore, India, 564–569 (2025). https://doi.org/10.1109/ICIDCA66325.2025.11280355.

Acknowledgements
The authors acknowledge the use of AI-based language-assistance tools, including the latest versions of Grammarly and Writefull for Overleaf, solely to refine the language and grammar of the manuscript (spelling, wording, and readability).

Author contributions
Conceptualization, Muhammad Sajid; methodology, Muhammad Sajid; software, Muhammad Sajid; formal analysis, Muhammad Sajid; investigation, Muhammad Sajid; data curation, Muhammad Sajid; validation, Muhammad Sajid; writing—original draft preparation, Muhammad Sajid; writing—review and editing, Muhammad Sajid, Nour Aburaed, and Rytis Maskeliūnas; visualization, Muhammad Sajid, Nour Aburaed, and Rytis Maskeliūnas; supervision, Muhammad Sajid; project administration, Muhammad Sajid; resources, Muhammad Sajid. All authors have read and agreed to the published version of the manuscript.

Funding
The author declares that this research received no external funding.

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval
Not applicable. The WESAD dataset used in this study is publicly available and was collected by its original providers under their respective ethics approvals.

Additional information
Correspondence and requests for materials should be addressed to M.Z.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026