HypoxSpike: Ternary Spiking Neural Network for Opioid Overdose Detection

Anush Lingamoorthy[1], Abhishek Kumar Mishra[1], Olumuyiwa Oni[2], Jacob S Brenner[3], Nagarajan Kandasamy[1], Amanda Watson[4]

[1]Drexel University
[2]AltruMed
[3]University of Pennsylvania
[4]University of Virginia
aln57@drexel.edu, am4862@drexel.edu, muyiwa@altru-med.com, jacob.brenner@pennmedicine.upenn.edu, nk78@drexel.edu, aawatson@virginia.edu

Abstract

Opioid overdose is a growing global health crisis that claims more than 120,000 lives annually, of which more than half use opioids alone, without access to bystander intervention. Fatal overdose events are marked by motionlessness, respiratory depression, and hypoxemia, yet current wearable systems often rely on a single biomarker, limiting detection speed and accuracy. We present HypoxSpike, a novel ternary spiking neural network designed for real-time, multi-biomarker overdose detection for low-power neuromorphic hardware, optimized for integration into shoulder-based wearables. HypoxSpike combines motion, respiration, and oxygen saturation signals, while accounting for skin tone and body physiology, thus addressing known racial bias in pulse oximetry. Our research uses an open-source shoulder-worn dataset from 19 patients experiencing sleep apnea, exploiting the shared physiological mechanisms underlying apnea and opioid overdose. This allows a direct comparison of our model with existing overdose detection approaches. HypoxSpike classifies three stages of hypoxemia with an average accuracy of 94%, outperforming state-of-the-art shoulder-based hypoxemia estimation while reducing false positive alert rates by 23.5%. By minimizing false positives, HypoxSpike supports accurate and power-efficient overdose detection, improving trust and usability for high-risk populations often overlooked by conventional systems.

Introduction

The opioid overdose epidemic is a serious public health crisis that has been exacerbated by the increase in synthetic opioids such as fentanyl and delays in timely intervention. Opioids target the regions of the brain responsible for regulating respiration, and fatal overdose leads to respiratory depression, causing people to stop breathing, resulting in neurological damage and death. Furthermore, the epidemic imposes a substantial financial burden on healthcare systems, with estimated annual costs of $11.3 billion in the United States alone (Inc. 2019; Velez et al. 2021). Almost 40% of overdose cases involve serious complications, including multiorgan failure due to prolonged oxygen deprivation, which requires expensive admissions to the intensive care unit. These severe outcomes highlight the urgent need for rapid intervention to mitigate life-threatening consequences. Although several technologies have been proposed to address this crisis (Lingamoorthy et al. 2024; Chan et al. 2021; Dhowan et al. 2019; Nandakumar, Gollakota, and Sunshine 2019; Huang et al. 2024; Lingamoorthy et al. 2023), including FDA-authorized devices (Knopf 2023), their adoption remains limited. Barriers such as form-factor stigma, smartphone dependency, and affordability restrict usability across socioeconomically vulnerable populations, particularly among people who use opioids (PWUOs) with unstable housing. This highlights a pressing need for accurate and power-efficient overdose monitoring systems that align with user preferences while enabling real-time detection and intervention.

Wearable sensors have gained attention as scalable tools for overdose detection. Masimo's Halo device monitors peripheral oxygen saturation via fingertip photoplethysmography (PPG), a form of optical spectroscopy that measures pulsatile changes in capillary blood volume through light absorption and reflection (Park et al. 2022; Watson et al. 2023). However, its visible form factor and the reliance on smartphones have hindered adoption (Knopf 2023). Other research systems, including sonar-based smartphones (Nandakumar, Gollakota, and Sunshine 2019), implantable naloxone injectors (Huang et al. 2024), and minimally invasive devices (Dhowan et al. 2019), face similar challenges related to user comfort, energy efficiency, and cost. Moreover, many of these approaches rely on a single biomarker, increasing false positive rates and limiting robustness in real-world deployment, which are major concerns expressed by PWUOs (Roth et al. 2025b,a). To address usability and adoption, Lingamoorthy et al. proposed a concealed shoulder-worn device for overdose monitoring based on PWUO willingness studies (Lingamoorthy et al. 2024; Kanter et al. 2021). Their capsule network based OxyCaps model was trained on obstructive sleep apnea (OSA) data as a proxy for opioid-induced hypoxemia. This approach has been used by FDA-approved overdose detectors due to the sensitive nature of collecting real-world opioid overdose data (Masimo Corporation 2023). The OxyCaps model demonstrated an accuracy of 92% in detecting hypoxemia, but was computationally intensive, requiring processing off-device on a smartphone.

Neuromorphic computing offers a promising solution for

on-body inference by mimicking the sparse and event-driven computation of the brain. Spiking Neural Networks (SNNs), the backbone of neuromorphic AI, have shown promise in applications such as ECG anomaly detection and PPG-based vital sign estimation (Bauer, Muir, and Indiveri 2019; Yang et al. 2024). However, traditional SNNs have limited information capacity, as they convey information using binary spikes $\{0, +1\}$. We propose HypoxSpike, a new ternary $\{-1, 0, +1\}$ SNN designed for real-time overdose detection using motion, respiration, and SpO_2 signals. The design improves information capacity and the model's robustness in processing time-series biosignals, while maintaining low energy costs. Our contributions are summarized as follows:

- We present HypoxSpike, the first ternary SNN-based architecture capable of running on neuromorphic hardware for real-time overdose detection using multi-biomarker input (SpO_2, respiration, and motion) from shoulder-mounted sensors.
- Our model achieves a classification accuracy of 94% on a 3-class shoulder-based hypoxemia task, outperforming the prior state-of-the-art OxyCaps model (92%) on the same dataset. Our model also decreases false positive alerts by 23.5% compared to OxyCaps.

Related Work

We discuss the technologies and algorithms surrounding the development of opioid overdose detectors, the applications of machine learning (ML) in hypoxemia estimation, and the applications of SNNs in healthcare and how it translates to overdose detection.

Overdose Detection Approaches

The ongoing opioid crisis has highlighted the critical role of timely real-time overdose detection and intervention, particularly as bystanders often do not recognize the symptoms of a fatal overdose, with 27.4% unable to do so (Ogeil et al. 2018; Dayton et al. 2019; Tas et al. 2023). Traditional bystander-reliant methods are limited by the need for a trained observer, leaving those who overdose alone particularly vulnerable. To address these challenges, a range of medical and wearable devices have been developed, such as Masimo Halo, an FDA-approved fingertip device that uses PPG sensors to monitor hypoxic states and provide tiered alerts. Other approaches have developed devices that use various sensors, including short-range active sonar, PPG, and accelerometers, to detect signs of overdose. These solutions often focus on canonical sites such as the fingertip, wrist, chest, and even implanted (Chan et al. 2021; Dhowan et al. 2019; Huang et al. 2024), which may be stigmatizing or unsuitable for many users.

Hypoxemia Detection Using ML

Hypoxemia is a critical indicator of Opioid-Induced Respiratory Depression (OIRD) and plays a vital role in the detection of overdoses. Lazazzera et al. developed a machine learning classifier for the detection of apnea and hypopnea based on PPG and SpO_2 signals, using fluctuations in the amplitude of PPG and oxygen desaturation (Lazazzera et al. 2021). Their Gaussian Support Vector Machine model achieved an accuracy of 75.1% in detecting apneas and hypopneas. Mahmud et al. proposed a Res-SE-ConvNet deep neural network (DNN) to classify three levels of hypoxemia severity: normal, moderate, and critical using only PPG signals (Mahmud, Imran, and Shahnaz 2022). Their solution achieved a state-of-the-art accuracy of 96.5% at a canonical site (fingertip). Among shoulder-based hypoxemia detection models, OxyCaps uses capsule networks to preserve temporal and spatial relationships in physiological signals such as SpO_2, respiration rate and motion (Lingamoorthy et al. 2024). Using OSA patient data as an analog, OxyCaps demonstrated state-of-the-art accuracy (92%) in hypoxemia detection on a non-canonical site (shoulder). Although capsule networks excel at maintaining temporal hierarchies, their computational complexity poses challenges for on-chip deployment and energy efficiency.

Estimating Biomarker Data Using SNNs

SNNs are artificial neural networks designed to mimic the way biological neurons communicate by transmitting information through discrete electrical pulses, or "spikes" rather than continuous signals. This spike-based communication enables SNNs to efficiently process temporal and spatial data, closely emulating the brain's natural information processing mechanisms. Neuromorphic hardware, such as IBM's TrueNorth and Intel's Loihi, as well as various FPGA-based designs (Liu et al. 2023; Carpegna, Savino, and Di Carlo 2022; Matinizadeh et al. 2024a,b,c) are designed to implement these networks directly in hardware, offering increased speed and energy efficiency. SNNs utilize binary discrete states $\{0,1\}$ for spike activations and have been applied in various healthcare domains. SNNs have been applied to biosignal data to decode muscle activity, classify ECG heartbeats, and monitor respiration rate (Kumarasinghe, Kasabov, and Taylor 2021; Burelo et al. 2022; Xing et al. 2022; Bauer, Muir, and Indiveri 2019). Recent advances have led to the exploration of ternary SNNs, which extend the binary framework by incorporating three discrete states in their spike activations (Guo et al. 2024). This enhancement addresses the information loss inherent in binary spike activations, increasing the network's information capacity. Ternary SNNs maintain the event-driven and multiplication-free operational advantages of binary SNNs while providing a richer representation of the data. In summary, while binary SNNs have demonstrated efficacy in various time-series applications, ternary SNNs offer enhanced information capacity and efficiency.

SNNs: Binary and Ternary Models

We discuss the foundations of SNNs, focusing on binary neurons for energy-efficient computation and ternary neurons for improved nformation capacity. These models provide the basis for the development of HypoxSpike.

Binary Spiking Neurons

SNNs encode and process information through discrete spikes, closely mimicking the event-driven behavior of biological neurons. In this framework, neurons are represented

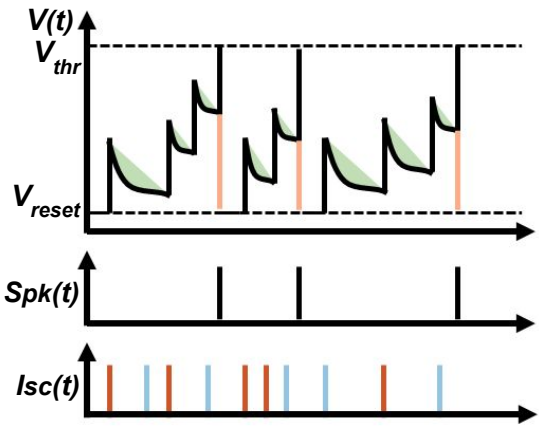

Figure 1: The update, fire, and reset cycle for an LIF neuron.

as vertices in a directed graph, with edges denoting synapses and weights indicating synaptic strength. This work uses the leaky integrate-and-fire (LIF) neuron model (Izhikevich 2004), where the membrane potential ${V_t}^L$ integrates input spikes and decays over time. When ${V_t}^L \geq V_{\text{thr}}$, a spike is emitted and the potential resets to V_{reset}.

$$ {Spk_t}^{L_j}, {V_t}^{L_j} = \begin{cases} 1, V_{reset}, & \text{if } {V_t}^{L_j} \geq V_{thr}; \\ 0, {V_t}^{L_j} & \text{if } {V_t}^{L_j} < V_{thr} \end{cases} \quad (1) $$

Algorithm 1 describes the process of updating the state of a neuron at time t. The Heaviside step function H ensures that the membrane potential resets to V_{reset} after spiking. The step function is defined as $H(z) = 1$ for $z > 0$ and $H(z) = 0$ otherwise. If a neuron fires a spike at time $t-1$, its membrane potential at the subsequent time t is reset, effectively removing the decayed component $W_{vd}^L \cdot {V_{t-1}}^L$ through the term $(1 - |Spk_{t-1}^L|)$, ensuring that the potential is cleared for the next cycle of activity. This mechanism, which governs state transitions and spike generation, depends on decay factors W_{scd}^L, W_{vd}^L and a trainable parameter W_{fv}, as shown in Figure 1. The progression to the next state of the neuron, denoted as $State_t^L$, follows a systematic cycle of updating, firing, and resetting, as governed by (1) and (2).

$$ \begin{aligned} Isc_t^L &= W_{scd}^L \cdot Isc_{t-1}^L + W_{fv} \cdot Spk_t^{L-1} \text{ and} \\ {V_t}^L &= W_{vd}^L \cdot V_{t-1}^L + Isc_t^L \end{aligned} \quad (2) $$

SNNs are trained using direct spike-based methods or conversion-based DNN-to-SNN approaches. Direct methods use gradient descent with surrogate gradients for supervised learning (Lee et al. 2020; Mishra et al. 2024), while unsupervised learning relies on STDP (Diehl and Cook 2015). Conversion methods approximate firing rates or replace activations (Cao, Chen, and Khosla 2015). Our approach employs direct supervised training tailored for efficient spike-based learning.

Ternary Spiking Neuron

Since SNNs convey information using binary spikes ({0,1}), they achieve high energy efficiency by replacing multiplications with additions. However, binary spike activation maps have limited information capacity compared to full-precision DNN activations, as they quantize membrane potentials to the same spike values, leading to information loss and reduced precision (Sun et al. 2022; Guo et al. 2022; Wang, Zhang, and Zhang 2023; Guo et al. 2024). To address this, we propose a model for a ternary spiking neuron that increases information capacity by transmitting spikes with values {-1,0,1}, while retaining the event-driven and multiplication-free benefits of binary neurons. The ternary spiking neuron is defined as follows:

Algorithm 1: LIF neuron state update at time t in layer L

Input:
Spk_{t-1}^L: previous spike, Isc_{t-1}^L: synaptic current
V_{t-1}^L: membrane potential, Spk_t^{L-1}: current spike from previous layer, W_{scd}^L: synaptic current-decay weight, W_{vd}^L: voltage-decay weight, V_{thr}: threshold voltage , W_{fv}: postsynaptic potential parameters, H: heaviside step function

function LIFSTATEUPDATE($Spk_t^{L-1}, State_{t-1}^L, W_{scd}^L, W_{vd}^L, V_{thr}, W_{fv}, H$)
  $Spk_{t-1}^L, Isc_{t-1}^L, V_{t-1}^L \leftarrow State_{t-1}^L$
  $Isc_t^L \leftarrow W_{scd}^L \cdot Isc_{t-1}^L + W_{fv} \cdot Spk_{t-1}^L$
  ${V_t}^L \leftarrow W_{vd}^L \cdot {V_{t-1}}^L \cdot (1 - |Spk_{t-1}^L|) + Isc_t^L$
  $Spk_t^L \leftarrow H({V_t}^L - V_{thr})$
  $State_t^L \leftarrow (Spk_t^L, Isc_t^L, {V_t}^L)$
  return $Spk_t^L, State_t^L$
end function

$$ \overset{+}{Spk}_t^{L-1} = (Spk_t^{L-1} == 1) * Spk_t^{L-1}, \quad (3) $$

$$ \overset{-}{Spk}_t^{L-1} = (Spk_t^{L-1} == -1) * Spk_t^{L-1}, $$

$$ Isc_t^L = W_{scd}^L \cdot Isc_{t-1}^L + \overset{+}{W}_{fv} \cdot \overset{+}{Spk}_t^{L-1} + \overset{-}{W}_{fv} \cdot \overset{-}{Spk}_t^{L-1}, \quad (4) $$

$$ {V_t}^L = W_{vd}^L \cdot V_{t-1}^L + Isc_t^L, $$

$$ {Spk_t}^{L_j}, {V_t}^{L_j} = \begin{cases} 1, V_{reset} & \text{if } {V_t}^{L_j} \geq V_{thr}; \\ -1, V_{reset} & \text{if } {V_t}^{L_j} \leq -V_{thr}; \\ 0, {V_t}^{L_j} & \text{if } -V_{thr} < {V_t}^{L_j} < V_{thr} \end{cases} \quad (5) $$

Here, $\overset{+}{Spk}_t^{L-1} \in \{0, 1\}$ and $\overset{-}{Spk}_t^{L-1} \in \{-1, 0\}$ represent positive and negative spikes, with $\overset{+}{W}_{fv}$ and $\overset{-}{W}_{fv}$ as their respective trainable weights. By applying distinct weights, the neuron processes excitatory and inhibitory activations separately, improving stability and control over membrane potential changes, thereby enhancing the model's representational capacity to capture complex features.

Unlike previous works that use $\{0, 1, 2\}$ spikes, requiring multiplications (Sun et al. 2022), our approach achieves multiplication-free inference while distinguishing excitatory and inhibitory activations. Compared to Guo et al., our model's use of distinct weights for positive and negative

spikes further improves stability and representational capacity (Guo et al. 2024).

Development of HypoxSpike

We propose the HypoxSpike architecture, shown in Fig. 2, which comprises a convolutional spike encoding layer for extracting low-level features as spikes from the input, followed by multiple spike-based convolutional layers to capture spatio-temporal features. The architecture ends with a fully connected spiking layer and an output layer. Each component of the architecture, along with the training procedure, is detailed in the following sections.

Convolutional Spike Encoding Layer

Given the input features, we employ spike rate-based encoding to transform normalized continuous values from optical and motion sensors into Poisson spike trains, where feature values determine the mean of a binomial distribution (Tavanaei et al. 2019; Subbulakshmi Radhakrishnan et al. 2021). This biologically inspired method ensures that stronger inputs produce higher spike rates, while weaker inputs result in fewer spikes (Neftci, Mostafa, and Zenke 2019; Rathi and Roy 2020; Tavanaei et al. 2019). After normalization, input features are resized to $(1, 18, 18)$ and processed through a convolutional layer with a (3×3) kernel (stride of 1, without padding), producing 64 feature maps that capture spatial activation patterns. For a feature map P^{c_i}, the value at position (x, y) is computed as

$$\left(\sum_d \sum_j \sum_k IW_d(x+j, y+k) \times {W_d^{enc}}_{c_i}(j,k) \right) + b_{c_i}^{enc}, \tag{6}$$

where IW and ${W_d^{enc}}_{c_i}$ are the input volume and kernel, $b_{c_i}^{enc}$ is the bias, and d iterates over the depth dimensions. The resulting output $P^{c_i} \in (batch\_size, 64, 8, 8)$ serves as the potential function f_v for postsynaptic encoding.

Each feature map undergoes the LIF operation cycle described in (1) and (2), converting potentials into binary spike events $P^{c_i}(Spk_t)(x, y)$ while preserving spatial relationships. This trainable encoding process eliminates manual parameter tuning, improving efficiency over traditional heuristic methods (Schrauwen and Van Campenhout 2003; Petro, Kasabov, and Kiss 2019). Figure 3 illustrates the operation for two time steps.

In summary, spikes generated by the convolutional spike encoding layer indicate the presence or absence of spike events at each feature map location, seamlessly transitioning from encoding to further processing within the SNN.

Spiking-Based Convolutional Layer

This layer processes the dynamic spike patterns generated by the spike encoding layer. Unlike the previous layer, which encodes normalized input data into spikes, this convolutional layer operates entirely on spike inputs produced by LIF neurons. By mimicking LIF behavior, it processes spikes through convolutions adjusted by synaptic weights, enabling the extraction of higher-level features by capturing complex spatial dependencies. The convolution is defined as

$$f_{conv}^{c_i}(P^{c_i}(Spk_t))(x, y) = \left(\sum_d \sum_j \sum_k P_d^{c_i}(Spk_t)(x+j, y+k) \cdot W_d^{L\,c_i}(j,k) \right) + b^{L\,c_i}. \tag{7}$$

where $f_{conv}^{c_i}$ computes the postsynaptic potential at location (x, y) within the ${c_i}^{\text{th}}$ feature map. This layer uses learnable kernel weights $W_d^{L\,c_i}$ and biases $b^{L\,c_i}$ to optimize responses in the L^{th} layer. Our architecture extends this approach with up to two subsequent spiking convolutional layers, which further process the spike outputs, enabling intricate spatial feature extraction. LIF operations in each convolutional layer generate spike events for further processing.

Spiking-Based Fully Connected Layer

The fully connected layer in HypoxSpike synthesizes the spatial features extracted by the preceding spiking-based convolutional layers through a linear transformation, enabling the network to identify complex patterns and better approximate the target function. The linear transformation is defined as

$$f_{fcv}(Spk_t^{L-1}) = W_{fc}^L \cdot Spk_t^{L-1} + b_{fc}^L, \tag{8}$$

where f_{fcv} represents the postsynaptic potential in the fully connected layer L, and Spk_t^{L-1} denotes the spike inputs from the last spiking-based convolutional layer. This layer utilizes learnable parameters, including synaptic weights W_{fc}^L and bias vector b_{fc}^L, to optimize its behavior.

Non-Spiking-Based Fully Connected Layer

The non-spiking fully connected layer serves as a decoder, analogous to the output layer in a DNN, translating the spike-based representation into a vector of class probabilities for each time step. This translation is performed as

$$prob_t = W_{nspk}^L \cdot Spk_t^{L-1} + b_{nspk}^L \text{ and}$$
$$Prob_{class} = \sum_t W_t \cdot prob_t, \tag{9}$$

where W_{nspk}^L and b_{nspk}^L are the learnable parameters of the non-spiking layer, and Spk_t^{L-1} represents the incoming spikes from the preceding spiking fully connected layer. Equation (9) calculates the time-specific vector of the class probabilities $prob_t$ using time-dependent weights W_t, followed by a weighted sum across all time steps to calculate the final class probabilities $Prob_{class}$. In summary, this layer consolidates and interprets the features learned by previous layers, mapping spike-based representations into understandable output predictions.

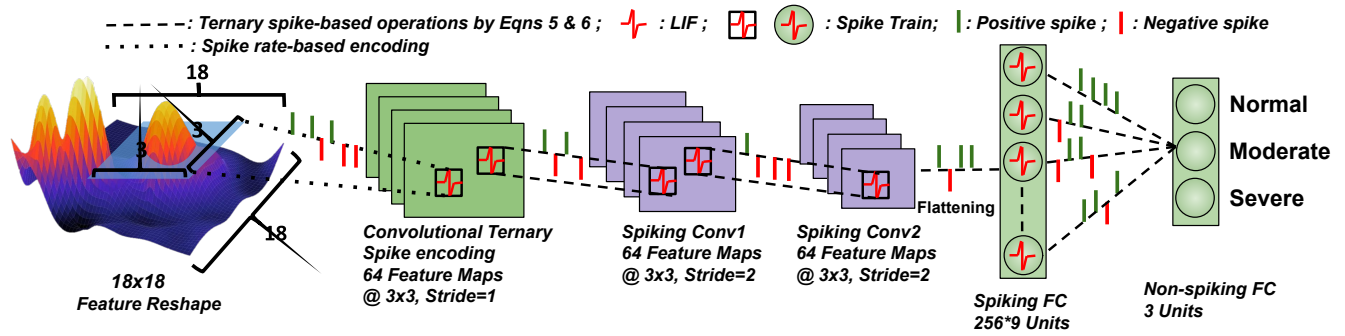

Figure 2: HypoxSpike Architecture: convolutional ternary spike encoding layer, spike-based convolutional layers, a fully connected spiking layer, and the non-spiking output layer.

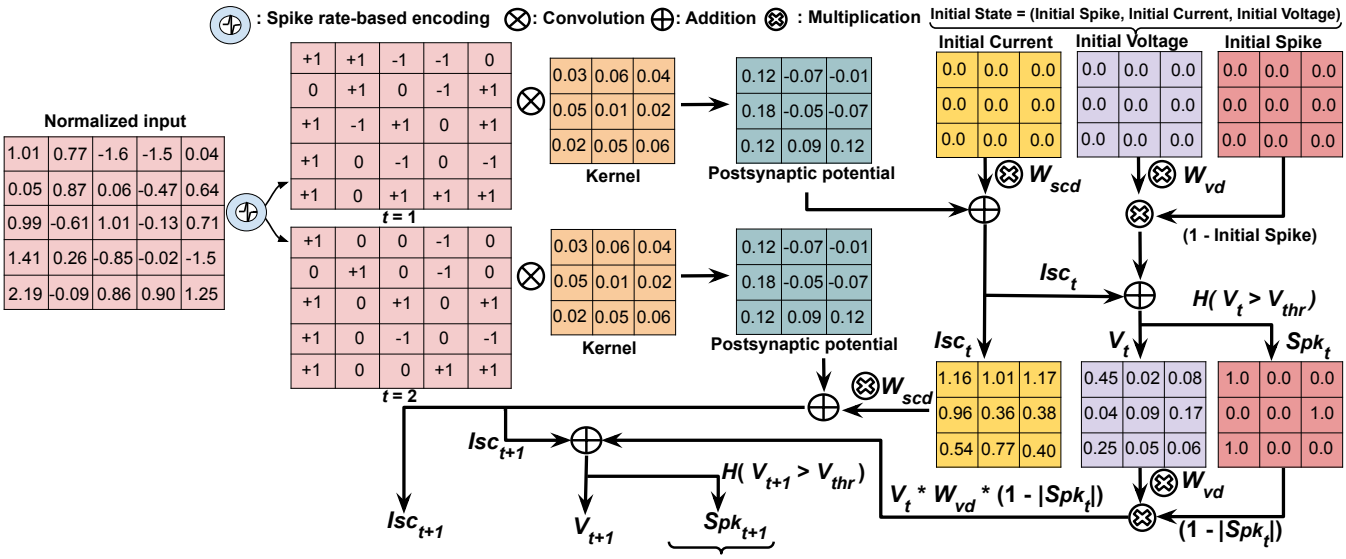

Figure 3: Convolutional spike encoding operation for the first two time steps.

SNN Training Procedure

HypoxSpike is trained using a cross-entropy loss function

$$\mathcal{L}_{ce}(y, Prob_{class}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic} \cdot \log(Prob_{class_{ic}}) \tag{10}$$

to classify C different hypoxemia classes where y_{ic} is an indicator of the correct class label c for observation i, $Prob_{class_{ic}}$ is the predicted probability for class c and N is the total number of training events. Given the temporal dependencies in the propagation of spikes between layers, gradient-based training incorporates both spatial and temporal derivatives (Wu et al. 2018). Synaptic current decay parameters W_{scd} and voltage decay parameters W_{vd} are optimized using a spatio-temporal backpropagation method.

To address the non-differentiability of the Heaviside step function H, we use a pseudo-gradient approach with the rectangular approximation

$$G_{ps}(V) = \begin{cases} C_g & \text{if } (|V + V_{thr}| + |V - V_{thr}|) < C_w, \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where C_g is the gain factor, V_{thr} is the threshold voltage and C_w is the threshold window of the gradient. The pseudo-gradient enables effective optimization of the ternary spike activation. Loss gradients are calculated with respect to neuronal states, synaptic weights, and biases as

$$\frac{\partial \mathcal{L}_{ce}}{\partial V_t^L} = G_{ps}(V_t^L) \cdot \frac{\partial \mathcal{L}_{ce}}{\partial Spk_t^L} + W_{vd}^L \cdot (1 - |Spk_t^L|) \cdot \frac{\partial \mathcal{L}_{ce}}{\partial V_{t+1}^L},$$
$$\frac{\partial \mathcal{L}_{ce}}{\partial Isc_t^L} = \frac{\partial \mathcal{L}_{ce}}{\partial V_{t+1}^L} + W_{scd}^L \cdot \frac{\partial \mathcal{L}_{ce}}{\partial Isc_{t+1}^L}, \frac{\partial \mathcal{L}_{ce}}{\partial Spk_t^{L-1}} = W'^L \cdot \frac{\partial \mathcal{L}_{ce}}{\partial Isc_t^L},$$
$$\frac{\partial \mathcal{L}_{ce}}{\partial W^L} = \sum_{t=1}^{T} Spk_t^{L-1} \cdot \frac{\partial \mathcal{L}_{ce}}{\partial Isc_t^L} \text{ and } \frac{\partial \mathcal{L}_{ce}}{\partial b^L} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}_{ce}}{\partial Isc_t^L}$$
$$\frac{\partial \mathcal{L}_{ce}}{\partial {W_{scd}}^L} = \sum_{t=1}^{T} Isc_{t-1}^L \cdot \frac{\partial \mathcal{L}_{ce}}{\partial Isc_t^L}$$
$$\frac{\partial \mathcal{L}_{ce}}{\partial W_{vd}^L} = \sum_{t=1}^{T} V_{t-1}^L \cdot (1 - |Spk_{t-1}^L|) \cdot \frac{\partial \mathcal{L}_{ce}}{\partial V_t^L} \tag{12}$$

Algorithm 2 summarizes the training process, combining forward propagation through ternary spike-based layers and backward propagation using gradients calculated by the equations described above. The loss is calculated across time

Algorithm 2: Training Procedure

Input:
$Train$: training set $\{(IW^i, C_i)\}_1^N$
W: weights, Spk: spikes, V_{thr}: threshold voltage
Isc: synaptic current, W_{scd}: current-decay weight
V: membrane potential, W_{vd}: voltage decay weight
W_{fv}: postsynaptic function weight, n_{epochs}: epochs
η: learning rate, T: timesteps, W_t: time-dependent weight

L^{Total} = [Convolutional Spike Encoding (·), Spike-based Conv1 (·), Spike-based Conv2 (·), Spiking-based fully connected (·)]
L_{nspk} = Non-spiking-based fully connected (·)
for each mini-batch B in $Train$ **do**
    Forward pass
    Out_prob = [ ] /* List to store output probabilities */
    for $t = 1, \ldots, T$ **do**
        for L in L^{Total} **do**
            $Spk_t^L, State_t^L = \text{LIFStateUpdate}(Spk_t^{L-1}, State_{t-1}^L, W_{scd}^L, W_{vd}^L, V_{thr}, W_{fv}, H)$
        end for
        Out_prob += $L_{nspk}(Spk_t^L) \cdot W_t$
    end for
    Calculate the loss
    Out_prob = $\sum_{t=1}^{T}$ Out_prob$[t, :, :]$
    $\mathcal{L}_{ce}(y_B, \text{Out\_prob})$ /* Out_prob $\in$ (B, $|C|$) */
    Backward pass
    Calculate gradient of the loss $\mathcal{L}_{ce}$ with respect to weights using (12)
    Update the model weights as $W := W - \eta \frac{\partial \mathcal{L}_{ce}}{\partial W}$
end for

steps and the weights are updated to minimize it using a learning rate η. The final class probabilities $Prob_{class}$ are derived by adding time-weighted probabilities $Prob_t$ over all time steps.

We train HypoxSpike using a batch size of 256, learning rate $\eta = 10^{-4}$, threshold voltage $V_{thr} = 0.1$, gain factor $C_g = 0.3$, and gradient threshold window $C_w = 1.5$. The synaptic current and membrane voltage decay parameters were set to 0.1, with a simulation time step $T = 6$. The rectangular approximation function is a simple yet flexible mechanism to modulate both the steepness and spread of a neuron's sensitivity to membrane potential, enabling tunable gating behavior for the spiking dynamics. Collectively, these settings ensured stable training and efficient spike-based processing throughout the simulation.

Results and Evaluation

We describe the open-source dataset used (Lingamoorthy et al. 2024) and assess the performance of HypoxSpike as a hypoxemia classifier against the current SoTA, as detailed in the accompanying open-source repository (https://github.com/abhishekkumarm98/HypoxSpike). The performance of the classifier is evaluated using Precision, Recall, F1-Score, and Accuracy metrics.

Model	Precision	Recall	F1-Score	Accuracy
RandomForest^C	0.88	0.88	0.88	0.88
ExtraTrees^C	0.88	0.88	0.88	0.88
LightGBM^C	0.84	0.84	0.84	0.84
XGBoost^C	0.88	0.88	0.88	0.88
CatBoost^C	0.85	0.85	0.85	0.86
HistGradient^C	0.89	0.89	0.89	0.90
3-Layers ANN^C	0.89	0.89	0.89	0.89
3-Layers CNN	0.90	0.90	0.90	0.90
OxyCaps^C	0.92	0.92	0.92	0.92
HypoxSpike	**0.93** $\pm$**0.010**	**0.93** $\pm$**0.011**	**0.93** $\pm$**0.007**	**0.94** $\pm$**0.005**

Table 1: Ablation study 5-Fold cross-validation and all the results (^{C}) are taken from (Lingamoorthy et al. 2024).

Dataset

The dataset used was ethically collected from 19 sleep apnea patients during overnight studies, exploiting the physiological similarities between apneic events and opioid overdoses. This approach has been used by FDA-approved overdose detectors due to the sensitive nature of collecting real-world opioid overdose data (Masimo Corporation 2023). Furthermore, FDA guidelines for overdose detection devices require only 12 subjects with various skin tones (FDA 2013). Our 19-subject dataset exceeds this, both in size and diversity.

The custom shoulder-mounted device recorded raw PPG and motion signals at 25Hz, alongside a commercial fingertip pulse oximeter to label hypoxemic states. Hypoxemia levels were classified into three stages based on SpO_2 levels: Normal (96%–100%), Moderate (92%–95%), and Severe (88%–91%). The dataset provided biomarker features such as oxygen saturation, respiration rate, and motion levels. The device dynamically normalized the PPG signal across skin tones by adjusting the drive current of each LED to compensate for variations in melanin absorption. The resulting LED current values, which reflect the optical power required for consistent signal acquisition, are stored as part of the dataset. In our model, the currents drawn by the red and infrared LEDs serve as proxy features for skin tone and individual physiology, allowing correction of skin tone–related errors commonly observed in optical measurements. This dataset enables the development of models for overdose detection while addressing challenges such as lower perfusion and increased skin thickness at the shoulder, compared to traditional sensing locations such as the wrist and fingertip.

Cross-Validation

To validate HypoxSpike, we used a 5-fold cross-validation approach, similar to OxyCaps (Lingamoorthy et al. 2024) and other works (Lazazzera et al. 2021; Mahmud, Imran, and Shahnaz 2022). The results of all 5-folds are averaged to provide a comprehensive evaluation of the model's performance compared to OxyCaps.

The results reported in Table 1 demonstrate that HypoxSpike outperforms previous models, achieving an accuracy of 94% and surpassing the SoTA OxyCaps algorithm (92%). Although a 2% improvement may appear modest, it

	HypoxSpike			OxyCaps^C		
Metric	Normal	Mod.	Sev.	Normal	Mod.	Sev.
Precision	0.932	0.904	0.957	0.911	0.893	0.957
Recall	0.923	0.906	0.963	0.906	0.881	0.975
F1-score	0.928	0.905	0.960	0.909	0.887	0.966
Sensitivity	0.923	0.906	0.963	0.906	0.881	0.975
Specificity	0.966	0.952	0.978	0.956	0.948	0.978

Table 2: Class-wise comparison of HypoxSpike and OxyCaps^C. Metrics are computed on the same test set.

	Predicted Labels		
True Labels	Severe	Moderate	Normal
Severe	13002	420	80
Moderate	433	12071	821
Normal	153	867	12300

Table 3: Confusion Matrix

translates to substantial gains in real-world applications: Table 2 shows that HypoxSpike improves specificity for the Normal class by 1.39% and reduces false positive alerts by 23.5%. Incorrectly flagging overdoses creates a major barrier to adoption and trust among PWUOs (Roth et al. 2025b). Reducing false alerts is both statistically significant and socially impactful. We also included a conventional 3-layer CNN baseline widely used in time-series classification involving biosignals, and with a model structure similar in complexity to HypoxSpike. As reported in Table 1, HypoxSpike achieves superior accuracy, confirming the advantage of the ternary spiking architecture over the neuron activation functions used in the CNN.

Referring to Table 2, HypoxSpike demonstrates better class-wise balance than OxyCaps with improved F1-scores in both Normal and Moderate categories and near-identical performance in Severe. In particular, HypoxSpike reduced false alarms in the Normal class by 23.5%, a critical advancement for real-world trust and usability. Although the recall for the Severe class is slightly lower, this can be attributed to the tighter decision boundaries of the network in multi-class settings, which prioritize minimizing false alarms in Normal cases without sacrificing critical sensitivity elsewhere. Sensitivity and specificity metrics show that HypoxSpike achieves a more stable classification between classes, with a general variation drop of 38% in F1-score, 20% in precision, and 93% in recall compared to OxyCaps. HypoxSpike shows reduced class imbalance effects compared to OxyCaps, maintaining consistently high performance across all classes rather than peaking for Severe. Table 3 shows that misclassifications in the Moderate class are reduced by 3%, showing improved temporal discrimination of gradual hypoxemic decline.

Table 4 presents an ablation study that compares binary and ternary SNN models across varying time-step windows. The ternary model consistently outperforms its binary counterpart, demonstrating its enhanced capacity to encode richer temporal dynamics in physiological signals. Although performance improves with longer windows up to 6 steps, gains taper off beyond that, indicating diminishing returns. These results confirm the benefits of using ternary spikes over traditional binary spikes for time-series classification.

Method	**Time-steps**	**F1-score**
Binary spike	4	0.915
Ternary spike	4	0.927
Binary spike	6	0.926
Ternary spike	6	0.934
Binary spike	8	0.924
Ternary spike	8	0.931

Table 4: Performance of binary and ternary spikes.

Discussion

Traditionally, SpO_2 measurements have relied on high-perfusion sites such as the fingertip. However, accurate hypoxemia detection with low false-postive rates from non-canonical sites such as the shoulder is critical to advancing wearable, patient-centric solutions to detect opioid overdose. Our work considered brain-inspired computing to develop SNNs for this purpose, since: (1) they process time-series data such as PPG signals with high accuracy; (2) are robust to noisy input data, which is expected from shoulder-mounted PPG sensor; and (3) provide a simpler computational model compared to a DNN, allowing for power-efficient implementations on neuromorphic hardware.

We have shown that HypoxSpike is very competitive with the other proposed approaches, achieving an accuracy of 94%. HypoxSpike demonstrates robustness to noise, using raw PPG signals without any denoising. SNN also offers a much simpler computational model compared to a traditional DNN due to two reasons: (1) spike activity within the SNN is typically sparse, with many neurons inactive at any given moment, leading to fewer active calculations; and (2) addition operations dominate over multiplication operations.

We do not consider issues related to the deployment of SNNs on neuromorphic hardware in this work, although this is the focus of our ongoing and future work. Deployment on a target hardware requires custom modifications tailored to the SNN's dynamics. SNNs incur some latency during inference, as T simulation steps are needed to stabilize the spiking activity in the network. (HypoxSpike incurs an inference time of 0.06 ms in our simulation setup.) This latency may pose challenges for real-time applications, but can be mitigated by using custom hardware (BrainChip 2023).

Our work marks a step towards wearable health technologies for underserved populations. HypoxSpike not only improves the accuracy of hypoxemia detection but also significantly reduces false positive alerts, a key factor for adoption among PWUOs. The success of HypoxSpike also demonstrates the broader potential of neuromorphic computing for low-power, real-time monitoring in domains such as sleep apnea, high-altitude hypoxia, and chronic obstructive pulmonary disease.

References

Bauer, F. C.; Muir, D. R.; and Indiveri, G. 2019. Real-time ultra-low power ECG anomaly detection using an event-driven neuromorphic processor. *IEEE transactions on biomedical circuits and systems*, 13(6): 1575–1582.

BrainChip. 2023. Akida Neuromorphic System-on-Chip.

Burelo, K.; Ramantani, G.; Indiveri, G.; et al. 2022. A neuromorphic spiking neural network detects epileptic high frequency oscillations in the scalp EEG. *Scientific Reports*, 12: 1798.

Cao, Y.; Chen, Y.; and Khosla, D. 2015. Spiking deep convolutional neural networks for energy-efficient object recognition. *Int'l J. Computer Vision*, 113: 54–66.

Carpegna, A.; Savino, A.; and Di Carlo, S. 2022. Spiker: An FPGA-optimized hardware accelerator for spiking neural networks. In *IEEE Computer Society Annual Symp. VLSI (ISVLSI)*, 14–19.

Chan, J.; Iyer, V.; Wang, A.; Lyness, A.; Kooner, P.; Sunshine, J.; and Gollakota, S. 2021. Closed-loop wearable naloxone injector system. *Scientific reports*, 11(1): 1–13.

Dayton, L.; Gicquelais, R. E.; Tobin, K.; Davey-Rothwell, M.; Falade-Nwulia, O.; Kong, X.; Fingerhood, M.; Jones, A. A.; and Latkin, C. 2019. More than just availability: Who has access and who administers take-home naloxone in Baltimore, MD. *PloS one*, 14(11): e0224686.

Dhowan, B.; Lim, J.; MacLean, M. D.; Berman, A. G.; Kim, M. K.; Yang, Q.; Linnes, J.; Lee, C. H.; Goergen, C. J.; and Lee, H. 2019. Simple minimally-invasive automatic antidote delivery device (A2D2) towards closed-loop reversal of opioid overdose. *Journal of Controlled Release*, 306: 130–137.

Diehl, P. U.; and Cook, M. 2015. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in computational neuroscience*, 9: 99.

FDA. 2013. Pulse Oximeters—Premarket Notification Submissions [510(k)s]: Guidance for Industry and Food and Drug Administration.

Guo, Y.; Chen, Y.; Liu, X.; Peng, W.; Zhang, Y.; Huang, X.; and Ma, Z. 2024. Ternary spike: Learning ternary spikes for spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12244–12252.

Guo, Y.; Tong, X.; Chen, Y.; Zhang, L.; Liu, X.; Ma, Z.; and Huang, X. 2022. Recdis-snn: Rectifying membrane potential distribution for directly training spiking neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 326–335.

Huang, H.-W.; Chai, P. R.; Lee, S.; Kerssemakers, T.; Imani, A.; Chen, J.; Heim, M.; Bo, J. Y.; Wentworth, A.; Sanoudos-Dramaliotis, F. T.; et al. 2024. An implantable system for opioid safety. *Device*, 2(10).

Inc., P. 2019. Opioid Overdoses Costing U.S. Hospitals an Estimated $11 Billion Annually. https://premierinc.com/newsroom/press-releases/opioid-overdoses-costing-u-s-hospitals-an-estimated-11-billion-annually. Accessed: 2024-11-01.

Izhikevich, E. M. 2004. Which model to use for cortical spiking neurons? *IEEE Trans. Neural Networks*, 15(5): 1063–1070.

Kanter, K.; Gallagher, R.; Eweje, F.; Lee, A.; Gordon, D.; Landy, S.; Gasior, J.; Soto-Calderon, H.; Cronholm, P. F.; Cocchiaro, B.; et al. 2021. Willingness to use a wearable device capable of detecting and reversing overdose among people who use opioids in Philadelphia. *Harm Reduction Journal*, 18(1): 1–14.

Knopf, A. 2023. Masimo "Halo" gets FDA go-ahead for OTC and Rx. *Alcoholism & Drug Abuse Weekly*, 35(16): 4–5.

Kumarasinghe, K.; Kasabov, N.; and Taylor, D. 2021. Brain-inspired spiking neural networks for decoding and understanding muscle activity and kinematics from electroencephalography signals during hand movements. *Scientific Reports*, 11(1): 2486.

Lazazzera, R.; Deviaene, M.; Varon, C.; Buyse, B.; Testelmans, D.; Laguna, P.; Gil, E.; and Carrault, G. 2021. Detection and Classification of Sleep Apnea and Hypopnea Using PPG and SpO_2 Signals. *IEEE Transactions on Biomedical Engineering*, 68(5): 1496–1506.

Lee, C.; Sarwar, S. S.; Panda, P.; Srinivasan, G.; and Roy, K. 2020. Enabling spike-based backpropagation for training deep neural network architectures. *Frontiers in neuroscience*, 14: 497482.

Lingamoorthy, A.; Watson, A.; Henderson, K.; Mandal, A.; Gordon, D.; Ma, X.; Weimer, J.; Kandasamy, N.; and Brenner, J. S. 2023. Dove: Shoulder-Based Opioid Overdose Detection and Reversal Device. In *2023 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 56–67. IEEE.

Lingamoorthy, A. N.; Mishra, A. K.; Kumar, S.; Gordon, D.; Brenner, J.; Kandasamy, N.; and Watson, A. 2024. Drug Overdose Vital-Signs Evaluator Using Machine Learning. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 7358–7366. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Liu, H.; Chen, Y.; Zeng, Z.; Zhang, M.; et al. 2023. A Low Power and Low Latency FPGA-Based Spiking Neural Network Accelerator. In *IEEE Int'l Joint Conf. Neural Networks (IJCNN)*, 1–8.

Mahmud, T. I.; Imran, S. A.; and Shahnaz, C. 2022. Res-SE-ConvNet: A Deep Neural Network for Hypoxemia Severity Prediction for Hospital In-Patients Using Photoplethysmograph Signal. *IEEE Journal of Translational Engineering in Health and Medicine*, 10: 1–9.

Masimo Corporation. 2023. De Novo Classification Request for Masimo Corporation SafetyNet Opioid System. Technical Report DEN200011, U.S. Food and Drug Administration, Center for Devices and Radiological Health (CDRH). Accessed: 2025-08-02.

Matinizadeh, S.; Mohammadhassani, A.; Pacik-Nelson, N.; Polykretis, I.; Tishbi, K.; Kumar, S.; Mishra, A. K.; Kandasamy, N.; Shackleford, J.; et al. 2024a. Neuromorphic

computing for the masses. In *2024 International Conference on Neuromorphic Systems (ICONS)*, 39–46. IEEE.

Matinizadeh, S.; Mohammadhassani, A.; Pacik-Nelson, N.; Polykretisl, I.; Mishra, A.; Shackleford, J.; Kandasamy, N.; Gallo, E.; and Das, A. 2024b. A Fully-Configurable Digital Spiking Neuromorphic Hardware Design with Variable Quantization and Mixed Precision. In *2024 IEEE 67th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 937–941.

Matinizadeh, S.; Pacik-Nelson, N.; Polykretis, I.; Tishbi, K.; Kumar, S.; Mohammadhassani, A.; Mishra, A.; Kandasamy, N.; Shackleford, J.; et al. 2024c. A fully-configurable open-source software-defined digital quantized spiking neural core architecture. *arXiv preprint arXiv:2404.02248*.

Mishra, A.; Kumar, S.; Lingamoorthy, A.; Das, A.; and Kandasamy, N. 2024. Wafer2Spike: Spiking Neural Network for Wafer Map Pattern Classification. In *2024 IEEE International Test Conference (ITC)*, 16–20. IEEE.

Nandakumar, R.; Gollakota, S.; and Sunshine, J. E. 2019. Opioid overdose detection using smartphones. *Science translational medicine*, 11(474): eaau8914.

Neftci, E. O.; Mostafa, H.; and Zenke, F. 2019. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6): 51–63.

Ogeil, R. P.; Dwyer, J.; Bugeja, L.; Heilbronn, C.; Lubman, D. I.; and Lloyd, B. 2018. Pharmaceutical opioid overdose deaths and the presence of witnesses. *International Journal of Drug Policy*, 55: 8–13.

Park, J.; Seok, H. S.; Kim, S.-S.; and Shin, H. 2022. Photoplethysmogram analysis and applications: an integrative review. *Frontiers in physiology*, 12: 808451.

Petro, B.; Kasabov, N.; and Kiss, R. M. 2019. Selection and optimization of temporal spike encoding methods for spiking neural networks. *IEEE transactions on neural networks and learning systems*, 31(2): 358–370.

Rathi, N.; and Roy, K. 2020. Diet-snn: Direct input encoding with leakage and threshold optimization in deep spiking neural networks. *arXiv preprint arXiv:2008.03658*.

Roth, A.; Laurano, R.; D'Angelo, A.; Lingamoorthy, A.; Salzman, M.; Oni, O.; Brenner, J.; Cocchiaro, B.; Lankenau, S.; and Baston, C. 2025a. Building a Sensor to Detect and Respond to Opioid Overdose: Experiences on a Wearable Prototype From People who Inject Drugs. *Drug and alcohol dependence*, 267: 118–118.

Roth, A. M.; D'Angelo, A. K.; Gordon, D.; Cocchiaro, B.; Lingamoorthy, A.; Laurano, R.; Salzman, M.; Brenner, J. S.; and Baston, C. 2025b. Use of a Shoulder-Mounted Wearable Sensor Prototype Designed to Detect Opioid-Related Overdose: A Qualitative User Experience Study. *J Med Internet Res*, 27: e73566.

Schrauwen, B.; and Van Campenhout, J. 2003. BSA, a fast and accurate spike train encoding scheme. In *Proc. Int'l Joint Conf. Neural Networks (IJCNN)*, volume 4, 2825–2830.

Subbulakshmi Radhakrishnan, S.; Sebastian, A.; Oberoi, A.; Das, S.; and Das, S. 2021. A biomimetic neural encoder for spiking neural network. *Nature communications*, 12(1): 2143.

Sun, C.; Chen, Q.; Fu, Y.; and Li, L. 2022. Deep spiking neural network with ternary spikes. In *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 251–254. IEEE.

Tas, B.; Lawn, W.; Traykova, E. V.; Evans, R. A.; Murvai, B.; Walker, H.; and Strang, J. 2023. A scoping review of mHealth technologies for opioid overdose prevention, detection and response. *Drug and Alcohol Review*, 42(4): 748–764.

Tavanaei, A.; Ghodrati, M.; Kheradpisheh, S. R.; Masquelier, T.; and Maida, A. 2019. Deep learning in spiking neural networks. *Neural networks*, 111: 47–63.

Velez, F. F.; Colman, S.; Kauffman, L.; Ruetsch, C.; and Anastassopoulos, K. 2021. Real-world reduction in healthcare resource utilization following treatment of opioid use disorder with reSET-O, a novel prescription digital therapeutic. *Expert Review of Pharmacoeconomics & Outcomes Research*, 21(1): 69–76.

Wang, X.; Zhang, Y.; and Zhang, Y. 2023. MT-SNN: enhance spiking neural network with multiple thresholds. *arXiv preprint arXiv:2303.11127*.

Watson, A.; Kendell, C.; Lingamoorthy, A.; Lee, I.; and Weimer, J. 2023. Lumos: An Open-Source Device for Wearable Spectroscopy Research. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4): 1–24.

Wu, Y.; Deng, L.; Li, G.; and Shi, L. 2018. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience*, 12: 323875.

Xing, Y.; Zhang, L.; Hou, Z.; Li, X.; Shi, Y.; Yuan, Y.; Zhang, F.; Liang, S.; Li, Z.; and Yan, L. 2022. Accurate ECG Classification Based on Spiking Neural Network and Attentional Mechanism for Real-Time Implementation on Personal Portable Devices. *Electronics*, 11(12): 1889.

Yang, G.; Kang, Y.; Charlton, P. H.; Kyriacou, P. A.; Kim, K. K.; Li, L.; and Park, C. 2024. Energy-efficient PPG-based respiratory rate estimation using spiking neural networks. *Sensors*, 24(12): 3980.