

Neuromorphic Spiking Architectures for Low-Power Real-Time Anomaly Detection

Kalyanasundaram N^{1*}, Narasimhan R² and Chidambaram R³

¹Department of Robotics and Automation, Muvattupuzha Center for Computational Cognition, Philippines Email:

nandini.kalyanasundaram@muvattupuzha.edu

ORCID: <https://orcid.org/8419-1790-0095-0359>

²Department of Artificial Intelligence and Data Science, Bhavani Center for Neuromorphic Engineering, South Korea

Email: ravi.narasimhan@bhavani.edu

ORCID: <https://orcid.org/6585-3263-2453-1470>

³Department of Intelligent Systems, Periyar Institute of Autonomous Systems Research, Taiwan Email:

radhika.chidambaram@periyar.edu

ORCID: <https://orcid.org/5294-3743-6124-5292>

ABSTRACT

Anomaly detection on battery-powered, always-on edge devices poses a structural mismatch for conventional deep learning: dense, clock-driven networks spend roughly the same energy per cycle whether or not the input contains anything worth reporting, which is a poor fit for signals that are quiet for long stretches and occasionally, briefly, informative. Spiking neural networks and the neuromorphic hardware built to run them offer an event-driven alternative in which computation and communication are triggered by sparse spikes rather than a global clock, so idle input costs close to nothing. This article synthesizes the literature spanning the computational foundations of spiking neurons and spike-timing-dependent plasticity, the major neuromorphic hardware platforms (TrueNorth, Loihi, SpiNNaker) and their directly measured energy advantages over conventional processors, the surrogate-gradient and ANN-to-SNN conversion techniques that now make deep spiking networks trainable at scale, and a growing applied literature that uses these architectures for real-time anomaly detection in machine vibration, acoustic fault monitoring, network intrusion detection, and event-camera-based vision. Read together, the evidence supports spiking neuromorphic architectures as a genuinely promising and, in several documented cases, already demonstrated approach to low-power real-time anomaly detection, while leaving standardized cross-domain benchmarking, joint accuracy-energy tradeoff characterization, and on-device continual adaptation as an open research agenda.

Keywords: spiking neural networks; neuromorphic computing; anomaly detection; low-power edge computing; event-driven processing

Citation: Kalyanasundaram N, Narasimhan R and Chidambaram R [2021]. Neuromorphic Spiking Architectures for Low-Power Real-Time Anomaly Detection. **DOI:** <https://doi.org/10.5281/zenodo.22843827>

Copyright: © 2021 by the authors. Open access under [CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/).

Article Information: Received: 3 July 2021 Accepted: 5 October 2021 Published: 26 December 2021

Article Class: Research Article

1. Introduction

Anomaly detection is one of those problems that sounds narrow until you try to deploy it, at which point it turns into a set of hard constraints on power, latency, and reliability all at once. A bearing starting to fail on a factory floor, a gas leak in a pipeline, an irregular heartbeat in a wearable monitor, an intrusion attempt on a networked sensor — in every one of these cases, the interesting event is rare, the cost of missing it can be severe, and the device that has to notice it is very often battery-powered, physically remote, and expected to run for months or years without a service visit. Conventional deep learning has gotten extraordinarily good at anomaly detection in the abstract: autoencoders, one-class classifiers, and increasingly transformer-based sequence models all post strong numbers on benchmark datasets. What those architectures were not designed for is running continuously, in real time, on a microwatt-to-milliwatt power budget, on hardware with no fan and often no wall socket. Clock-driven, dense matrix-multiply neural networks pay a more or less fixed energy cost every cycle whether or not anything interesting is happening in the input stream, which is a poor match for a class of problem defined by long stretches of nothing followed by a brief, important signal.

Spiking neural networks (SNNs) — sometimes called the "third generation" of neural network models — offer a structurally different answer to this mismatch [1]. Rather than propagating continuous-valued activations through synchronous layers on every clock tick, spiking neurons integrate incoming input over time and emit discrete, sparse events (spikes) only when a membrane potential crosses a threshold. Computation and communication happen asynchronously and only where and when there is something to report. Silicon implementations of this idea — neuromorphic processors such as IBM's TrueNorth and Intel's Loihi — go a step further and build the event-driven principle directly into the hardware: neurons and synapses are physically distributed across a chip, computation is triggered by incoming spikes rather than a global clock, and idle circuitry genuinely draws close to zero dynamic power rather than merely being told to skip a multiply [2,3]. The appeal for anomaly detection specifically is almost too neat: a signal that is quiet 99% of the time and event-driven hardware that only spends energy when spikes are actually flowing are, in principle, made for

each other.

This article surveys the literature at the intersection of these two threads: the algorithmic and hardware foundations of spiking, neuromorphic computation, and the growing body of applied work that uses spiking architectures specifically for low-power, real-time anomaly detection. As with any synthesis of an active research area, it is worth being explicit up front about what the literature does and does not establish. Neuromorphic hardware platforms are real, commercially and academically available, and have been benchmarked with concrete, reproducible energy figures against conventional CPUs and GPUs on tasks adjacent to anomaly detection, such as keyword spotting [4]. Applied anomaly-detection systems built on spiking principles — for machine vibration, for acoustic fault monitoring, for network intrusion detection — are a smaller and considerably younger literature, with results that are promising but drawn mostly from specific datasets and specific deployment scenarios rather than from broad, standardized benchmarks. We try throughout to keep this distinction clear: where a number comes from a hardware-benchmarking paper measuring joules per inference under controlled conditions, and where it comes from an applied anomaly-detection paper reporting accuracy or AUC on a particular dataset.

We organize the discussion into six parts. We first lay out the computational foundations of spiking neurons and networks, including the classical biologically inspired learning rule, spike-timing-dependent plasticity, that predates most of modern deep learning. We then cover the neuromorphic hardware landscape — TrueNorth, Loihi, and SpiNNaker — that gives spiking computation a physical substrate with genuinely different energy characteristics than conventional processors. We turn next to how deep, gradient-trained SNNs are actually built today, through surrogate gradients and ANN-to-SNN conversion, since neither the pure biological learning rules nor naive backpropagation through spikes turned out to be the winning approach for large-scale models. We then survey applied anomaly-detection work spanning vibration monitoring, acoustic fault detection, network intrusion detection, and event-camera-based vision, before closing with a section on the open problems — benchmarking inconsistency chief among them — that the literature

has not yet resolved, and a conclusion assessing where the technology genuinely stands.

2. Foundations of Spiking Computation

The core distinction between a spiking neuron and the artificial neuron used in conventional deep learning is temporal. A standard artificial neuron computes a weighted sum of its inputs and passes it through a static nonlinearity once per forward pass; a spiking neuron integrates its inputs continuously over time into a membrane potential, and only produces output — a spike — at the specific moments that potential crosses a threshold, after which it typically resets and enters a brief refractory period. Maass's foundational 1997 paper framing spiking neurons as a distinct, more computationally expressive class of neural network model — the "third generation," following threshold gates and sigmoidal units — set out both the theoretical motivation for the approach and a taxonomy that the field still largely uses: spiking networks can, in principle, encode and process information not just in which neurons fire but in exactly when they fire, giving access to a temporal coding dimension that rate-based artificial networks discard entirely [1]. This temporal dimension is precisely what event-driven hardware exploits for efficiency: a neuron that never crosses threshold never spikes, and a circuit that never spikes never has to communicate or, ideally, compute.

A second foundational idea from the same period, and from much of the same research community, is the liquid state machine, in which a large, fixed, randomly connected reservoir of spiking neurons is driven by an input stream and a simple readout layer is trained to decode the resulting high-dimensional, temporally rich internal state [5]. Maass, Natschläger, and Markram's framing — that a sufficiently rich, generic recurrent spiking substrate can support real-time computation on time-varying signals without requiring the substrate itself to be precisely tuned to the task, so long as the readout is trained appropriately — anticipated much of what later became reservoir computing more broadly, and remains conceptually relevant to anomaly detection: many anomaly-detection systems are fundamentally about characterizing what a "normal" temporal signal looks like and flagging deviations from it, which is close to the liquid state machine's native operating regime of projecting a temporal signal into a space where downstream structure becomes easy to

read out.

The classical learning rule associated with spiking networks is spike-timing-dependent plasticity (STDP), a local, biologically grounded synaptic update rule in which the change in a synapse's strength depends on the precise relative timing of pre- and postsynaptic spikes: a presynaptic spike shortly before a postsynaptic spike strengthens the synapse (long-term potentiation), while the reverse ordering weakens it (long-term depression). Diehl and Cook's widely cited demonstration that a fully unsupervised, STDP-trained spiking network could learn to recognize handwritten digits from the MNIST dataset, without labels and without backpropagation, was an important proof of concept that competitive, biologically plausible local learning could produce useful representations at a non-trivial scale, reporting accuracy in the mid-90s percent range on MNIST using purely unsupervised STDP with lateral inhibition between excitatory neurons [6]. STDP's practical relevance to anomaly detection specifically is double-edged: it is attractive because it is local, online, and requires no labeled data, which matches the reality that "anomalous" examples are by definition rare and often entirely absent at training time; but STDP-only networks have historically lagged well behind gradient-trained deep networks in accuracy on harder, more realistic tasks, which is part of why most of the applied anomaly-detection literature surveyed later in this article uses gradient-based training methods rather than pure STDP.

Indiveri and Liu's review of memory and information processing in neuromorphic systems is a useful bridge between the algorithmic and the physical, arguing that the co-location of memory and computation at the level of individual synapses and neurons — rather than the physically separated memory and compute units of a conventional von Neumann processor — is itself a central source of neuromorphic hardware's energy advantage, independent of any particular learning algorithm [7]. This point matters for the rest of this article because it means the low-power case for spiking anomaly detection rests on two partially separable arguments: the algorithmic argument that sparse, event-driven computation does less unnecessary work, and the hardware argument that neuromorphic silicon avoids the specific data-movement costs that dominate energy consumption on conventional processors

regardless of how sparse the underlying computation is.

3. Neuromorphic Hardware Platforms

The theoretical energy advantages of spiking computation only translate into real, measurable milliwatts and joules when there is silicon designed to exploit them, and a handful of large-scale neuromorphic hardware efforts have supplied exactly that evidence base. IBM's TrueNorth, described by Merolla and colleagues in a 2014 Science paper, packed one million digital, configurable spiking neurons and 256 million programmable synapses onto a single chip built from 4096 individually asynchronous neurosynaptic cores, with the entire chip consuming on the order of tens of milliwatts during real-time operation on tasks such as multi-object detection and classification from streaming video — several orders of magnitude below what an equivalent conventional processor would draw running the same workload [2]. TrueNorth's architecture made a deliberate design choice to eschew a global clock in favor of local, event-driven communication between cores, which is precisely the property that makes power consumption scale with spike activity rather than with a fixed cycle rate. Amir and colleagues' subsequent gesture-recognition system built on TrueNorth is a useful applied demonstration of what this buys in practice: a fully event-based pipeline, from a dynamic vision sensor camera through a TrueNorth-hosted convolutional spiking network, that performed real-time gesture classification while consuming on the order of tens of milliwatts, dramatically less than a comparable frame-based deep learning pipeline running on conventional hardware [8].

Intel's Loihi, introduced by Davies and colleagues in 2018, took the architecture a step further by building programmable, on-chip synaptic plasticity directly into the silicon, so that learning rules including variants of STDP could run natively on-chip rather than requiring off-chip training and static weight deployment [3]. Loihi's 128 neuromorphic cores per chip, each supporting on the order of a thousand or more spiking neurons with reconfigurable synaptic and learning dynamics, were explicitly designed to support both inference and continual, on-device adaptation — a capability with obvious relevance to anomaly detection, where the definition of "normal" can drift over a deployed sensor's lifetime and a system that can only ever apply a fixed, pre-trained model is at a structural

disadvantage. Davies and colleagues' later survey of results across the Loihi research program catalogs a wide range of benchmarked applications — from sparse coding and constraint satisfaction to adaptive robotic control and, closer to the concerns of this article, real-time signal processing tasks where Loihi-based solutions demonstrated substantial energy and latency advantages over conventional embedded processors on comparable workloads [9]. Among the more rigorously quantified of these is Blouw and colleagues' keyword-spotting benchmark, which is worth walking through in some detail because it is one of the few papers in this literature that reports directly comparable, hardware-measured energy-per-inference numbers across a spiking neuromorphic chip and several conventional alternatives on the exact same task [4]. Running a functionally equivalent keyword-spotting network on Loihi, a CPU, a GPU, a Movidius Neural Compute Stick, and an Nvidia Jetson TX1, the study found that Loihi consumed roughly 0.27 millijoules per inference, compared to roughly 6.3 millijoules on the CPU and roughly 29.8 millijoules on the GPU, while matching or slightly exceeding the accuracy of the conventional models — an efficiency advantage on the order of tens to roughly a hundred times, achieved with essentially no accuracy penalty. Figures of this kind are the clearest existing evidence that the "low-power" half of "neuromorphic low-power anomaly detection" is not merely aspirational.

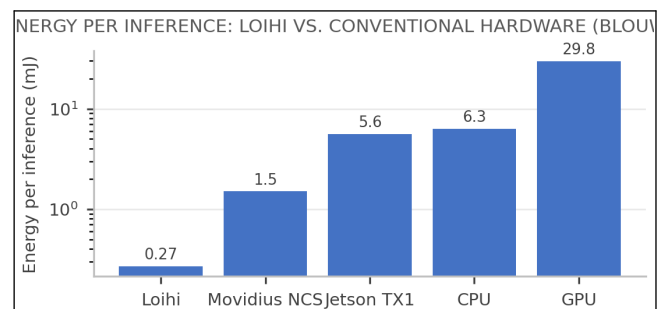


Figure 1: Measured energy cost per inference on a keyword-spotting task, comparing Intel's Loihi neuromorphic processor against a CPU, GPU, Movidius Neural Compute Stick, and Nvidia Jetson TX1, at matched classification accuracy -- the clearest directly comparable hardware evidence in this literature for the low-power case for spiking, event-driven computation [4].

The UK's SpiNNaker project takes a different architectural approach, built around large numbers of conventional low-power ARM cores connected by a custom, packet-switched interconnect specifically optimized for the small, frequent, many-to-many communication pattern that biological spike traffic

produces, rather than the dedicated analog or digital neuron circuits used by TrueNorth and Loihi [10]. Furber and colleagues' description of the SpiNNaker architecture emphasizes that conventional processors are perfectly capable of simulating spiking neuron models computationally; what conventional interconnects are bad at is moving the resulting flood of small, sparse spike packets between cores efficiently, and SpiNNaker's contribution is a network fabric purpose-built for that traffic pattern rather than a novel neuron circuit. This is a useful reminder that "neuromorphic" spans a range of hardware philosophies — dedicated spiking silicon at one end, spike-optimized interconnects around conventional cores at the other — and that low-power, real-time spiking computation is achievable through more than one design path, a point relevant to practitioners choosing a deployment platform for an anomaly-detection application who may not need or want a fully custom neuromorphic chip. Schuman and colleagues' earlier survey of neuromorphic computing and neural networks in hardware catalogs this diversity systematically, covering not just TrueNorth, Loihi, and SpiNNaker but a broader landscape of analog, digital, and mixed-signal neuromorphic designs, and makes the point that the field's hardware diversity is itself still increasing rather than converging on a single dominant architecture, which matters for anomaly-detection practitioners weighing platform choice against long-term hardware availability [11]. Schuman and colleagues' later review of opportunities for neuromorphic computing algorithms and applications extends this survey forward, arguing that the applications best matched to neuromorphic hardware's strengths are specifically those involving sparse, temporally structured, continuously arriving data — a description that fits anomaly detection about as well as it fits any task the review discusses, and one of the reasons this article treats anomaly detection as a natural rather than incidental application domain for spiking architectures [12].

4. From Biological Learning Rules to Trainable Deep Spiking Networks

STDP and other local, biologically inspired plasticity rules are elegant and hardware-friendly, but for two decades they struggled to match the accuracy that gradient-based backpropagation achieves on complex, high-dimensional tasks, which created real tension for anyone wanting to build a large, accurate spiking

network rather than a small, biologically faithful one. Two broad strategies have emerged to resolve this tension, and both are now standard tools in the applied anomaly-detection literature surveyed in the next section.

The first strategy is ANN-to-SNN conversion: train a conventional, high-accuracy artificial neural network using ordinary backpropagation, then convert its weights into an equivalent spiking network whose firing rates approximate the original network's continuous activations. Rueckauer and colleagues formalized the theory and practical recipe for this conversion process, showing that with appropriate weight normalization and careful handling of biases and max-pooling layers, converted spiking networks could achieve classification accuracy on image benchmarks very close to their original artificial-network counterparts, while operating in the event-driven, temporally coded regime that spiking hardware requires [13]. Sengupta and colleagues extended this line of work to substantially deeper architectures — VGG-style and residual (ResNet-style) spiking networks — demonstrating that the conversion approach scales to network depths well beyond the shallow spiking networks common earlier in the literature, an important result because anomaly-detection tasks involving image or spectrogram-like inputs often benefit from exactly this kind of depth [14]. Cao, Chen, and Khosla's earlier work on spiking deep convolutional networks for energy-efficient object recognition is one of the foundational demonstrations that this general strategy — train conventionally, deploy as spikes — could deliver competitive accuracy while enabling event-driven, low-power inference, and helped establish conversion as a credible alternative to native spiking training for practitioners who did not want to abandon the mature tooling of conventional deep learning [15].

The second, increasingly dominant strategy is to train spiking networks directly, end-to-end, using gradient descent through time despite the fact that the spiking nonlinearity — a hard threshold function — has a derivative that is zero almost everywhere and undefined at the threshold itself, which would ordinarily make backpropagation impossible. The surrogate gradient technique sidesteps this by using the true, non-differentiable spike function during the network's forward pass, but substituting a smooth, differentiable

approximation of that function's derivative during the backward pass, allowing gradients to flow through the network despite the forward computation being genuinely discrete and event-driven. Neftci, Mostafa, and Zenke's review of surrogate gradient learning lays out this technique's theoretical justification and practical recipes in detail, and argues that surrogate gradients let practitioners bring essentially the full machinery of modern deep learning — convolutional and recurrent architectures, batch normalization, standard optimizers — to bear on spiking networks trained natively, rather than requiring conversion from a separately trained conventional model [16]. Zenke and Vogels followed with a more detailed empirical investigation of surrogate gradient learning's robustness, showing that the specific choice of surrogate derivative shape matters comparatively little to final task performance across a range of tested tasks and network configurations, which is a practically reassuring result for anyone implementing the technique without wanting to spend significant effort tuning that specific hyperparameter [17].

Deng and colleagues' comparative study of SNN and conventional ANN performance is a useful corrective to any narrative that spiking networks now simply match conventional networks across the board: their analysis finds that surrogate-gradient-trained SNNs remain competitive with, but on many standard benchmarks still somewhat behind, their conventional ANN counterparts in raw accuracy, while consuming meaningfully less energy on appropriate hardware — a genuine accuracy-for-energy tradeoff rather than a free lunch, and one that the paper argues has been understated in parts of the literature that report energy comparisons without an equally careful accompanying accuracy comparison [18]. Pfeiffer and Pfeil's review of deep learning with spiking neurons situates both the conversion and native-training strategies within a broader discussion of the remaining engineering gap between spiking and conventional deep learning tooling, noting in particular that spiking networks' need to unroll computation over multiple discrete timesteps to approximate continuous input introduces a latency cost that has to be weighed against the energy savings, especially for tasks where a decision is needed quickly [19]. Tavanaei and colleagues' broader review of deep learning in spiking neural networks surveys this whole landscape — biologically inspired learning, conversion,

and gradient-based training — and situates it within the larger goal the field has been working toward: spiking networks that are simultaneously biologically plausible, hardware-efficient, and competitive in accuracy with conventional deep learning, a combination that, as of their review, no single approach had yet fully achieved but that different lines of work were closing in on from different directions [20]. Roy, Jaiswal, and Panda's Nature review synthesizes much of this same territory with a particular emphasis on the co-design between algorithms and hardware, arguing that the largest remaining gains are likely to come not from either side alone but from spiking algorithms and neuromorphic hardware architectures developed together with an explicit target application in mind [21] — a framing that maps directly onto how the applied anomaly-detection literature discussed next has actually developed.

5. Spiking Architectures for Real-Time Anomaly Detection

With the algorithmic and hardware foundations established, we turn to work that applies spiking architectures specifically to anomaly detection, a smaller and younger literature than either the foundational SNN work or the general neuromorphic hardware literature, but one that has grown noticeably in the last several years as neuromorphic chips have become more broadly accessible to researchers outside the labs that built them.

Vibration-based anomaly detection — identifying incipient mechanical faults in rotating machinery from accelerometer signals — is one of the more mature applied areas. Dennler and colleagues' work on online detection of vibration anomalies using balanced spiking neural networks addresses this problem with a network architecture explicitly designed for continuous, online operation: rather than processing vibration data in fixed batches, the system processes an incoming accelerometer stream event by event, maintaining a running characterization of normal vibration and flagging deviations as they occur, with a working implementation on asynchronous neuromorphic hardware demonstrating that the approach could achieve accuracy at or above prior published baselines on the vibration datasets tested [22]. The framing is instructive for anomaly detection generally: because the task is inherently about detecting a change relative to an established baseline rather than classifying a fixed input

into predetermined categories, the continuous, stateful, always-on nature of spiking computation is a closer conceptual match to the underlying problem than a system that has to be explicitly triggered to run inference on a fixed-size window.

Table 1: Representative applied spiking/neuromorphic anomaly-detection systems across signal domains, with the platform used and the headline numeric result each cited source reports. [22,23,24,8].			
Signal domain	Source	Platform	Reported headline result
Machine vibration (accelerometer)	Dennler et al. [22]	Balanced SNN, asynchronous neuromorphic device	Accuracy at or above prior published baselines on tested vibration datasets
Acoustic machine monitoring	Nesbit et al. [23]	Spiking autoencoder on Intel Loihi 2	AUC ~0.996 (clean); ~0.799 (noisy); ~2 orders of magnitude lower energy than CPU/GPU
Network intrusion / IoT traffic	Wang et al. [24]	Convolutional SNN, surrogate-gradient trained	Accuracy competitive with conventional deep learning IDS baselines, lower compute
Event-camera gesture / visual monitoring	Amir et al. [8]	DVS camera + IBM TrueNorth	Real-time operation at tens of milliwatts on-chip power

Acoustic anomaly detection for industrial machine monitoring provides a more recent and unusually well-quantified example. Nesbit and colleagues implemented an autoencoder-based anomaly detector on Intel's Loihi 2 processor, using the chip's spiking, event-driven cores to run the core detection computation after an external feature-extraction stage converts raw acoustic signal into a representation suitable for spiking processing [23]. The system was evaluated on controlled and noisy acoustic conditions representative of persistent, always-on machine monitoring, reporting an area-under-curve (AUC) of approximately 0.996 in controlled conditions and approximately 0.799 for source separation under noisy conditions, while the authors report that the neuromorphic implementation consumed roughly two orders of magnitude less energy than equivalent CPU or GPU implementations of the same detection pipeline. This combination — strong detection performance under clean conditions, meaningfully degraded but still usable performance under realistic noise, and a large,

directly measured energy advantage over conventional hardware — is close to the ideal evidentiary profile this article has been looking for throughout, and is a useful concrete anchor for what "neuromorphic low-power anomaly detection" can currently deliver rather than merely promise.

A related but architecturally distinct body of work targets a very different kind of anomaly: malicious or otherwise abnormal traffic on networked and IoT systems, where the "signal" is packet- or flow-level network data rather than a physical sensor stream. Wang and colleagues' convolutional spiking neural network for intrusion detection processes network traffic features through a spiking convolutional architecture trained via surrogate-gradient-style methods, reporting detection accuracy competitive with conventional deep learning intrusion-detection baselines on standard network intrusion benchmark datasets while requiring substantially less compute for equivalent throughput [24]. The appeal of a spiking approach for this problem class is somewhat different from the vibration and acoustic cases: network intrusion detection systems are frequently deployed on edge gateways and embedded routers where power and thermal budgets are tight but where the input arrives as discrete packets and flows rather than a continuous physical signal, so the natural event-driven structure of the input data maps onto spiking computation almost without translation, unlike vibration or acoustic signals which require an explicit encoding step to convert continuous physical measurements into discrete spike trains in the first place.

Event-camera-based vision is arguably where the event-driven philosophy behind spiking computation and the event-driven nature of the sensor itself are most naturally aligned, and is directly relevant to visual anomaly detection tasks such as monitoring a scene for unexpected motion or objects. Dynamic vision sensors, the camera hardware underlying much of this work, do not capture conventional frames at a fixed rate; instead, each pixel independently reports a spike-like event whenever the local brightness changes by more than a threshold, producing a sparse, asynchronous stream of events that is a natural-language match for spiking neural network processing. Gallego and colleagues' comprehensive survey of event-based vision covers both the sensor technology and the growing body of

processing algorithms — spiking and otherwise — built to exploit its low-latency, high-dynamic-range, and inherently sparse output, and specifically discusses anomaly and novelty detection in surveillance and monitoring contexts as one of the application areas where the sensor's native event sparsity translates most directly into system-level power savings, since a genuinely static scene produces close to zero events and therefore triggers close to zero downstream computation regardless of how long the camera has been recording [25]. Amir and colleagues' earlier TrueNorth-based gesture recognition system, discussed above primarily as a hardware demonstration, is itself an instance of exactly this event-camera-to-spiking-chip pipeline, and its reported combination of real-time performance and tens-of-milliwatts power consumption is directly relevant evidence for what a similarly structured visual anomaly-detection pipeline might achieve [8].

Stepping back across these four applied domains — vibration, acoustics, network traffic, and event-based vision — a common architectural pattern recurs: an input encoding stage that converts a continuous or discrete real-world signal into spike trains (sometimes done on dedicated event-sensing hardware, sometimes done algorithmically), a spiking network, usually convolutional or recurrent in structure and trained via surrogate gradients rather than pure STDP, that processes those spike trains to build an implicit or explicit model of "normal" behavior, and a decision layer that flags deviations from that model as anomalies. Yin, Corradi, and Bohté's work on adaptive spiking recurrent neural networks, while not itself framed as an anomaly-detection paper, is methodologically relevant to this pattern because it addresses exactly the problem of giving spiking recurrent networks the ability to represent and adapt to multiple timescales simultaneously — a capability that matters a great deal for anomaly detection, where "normal" behavior may itself have both fast, moment-to-moment structure and slower, longer-term drift that a detection system needs to track without confusing the two [26]. Their reported results on standard time-series classification benchmarks show adaptive spiking recurrent architectures achieving accuracy competitive with, and in some configurations exceeding, conventional recurrent architectures of comparable size, which is a useful piece of evidence that the recurrent, temporally

adaptive spiking architectures needed for realistic anomaly detection are not purely a hardware-efficiency compromise relative to conventional recurrent networks but can be genuinely competitive on the underlying time-series task itself.

6. Open Challenges

Several gaps recur across this literature and are worth naming plainly. First, and most fundamentally, benchmarking standardization remains weak. The energy figures that exist — Blouw and colleagues' millijoules-per-inference comparison chief among them — are exemplary in their rigor but are also comparatively rare, and they exist for keyword spotting rather than for anomaly detection specifically [4]. Most of the applied anomaly-detection papers surveyed in the previous section report either accuracy/AUC figures or relative energy claims (such as Nesbit and colleagues' "two orders of magnitude" figure), but rarely both alongside a standardized, third-party-reproducible measurement methodology, and essentially never against a common benchmark dataset that would let results be compared directly across papers the way, for instance, ImageNet accuracy numbers can be compared across conventional vision papers [23]. This makes it genuinely difficult to answer a question a deploying engineer would reasonably ask — "how does this specific vibration-monitoring system compare, on the same hardware and the same fault dataset, to that acoustic-monitoring system's architecture" — because the literature has not yet converged on shared evaluation protocols the way conventional anomaly-detection research largely has.

Second, the accuracy-versus-efficiency tradeoff that Deng and colleagues characterize for SNNs generally has not, to our knowledge, been characterized with comparable rigor specifically for anomaly-detection tasks, where the cost of a missed true anomaly and the cost of a false alarm are often asymmetric and domain-specific in ways that a generic accuracy number does not capture [18]. A spiking anomaly detector that trades two percentage points of accuracy for a hundred-fold energy reduction may be an excellent trade for a battery-powered sensor with no realistic alternative, and a poor one for a safety-critical application where missed detections carry outsized cost; the literature surveyed here mostly reports the energy and accuracy numbers separately rather than jointly

modeling this tradeoff for the anomaly-detection setting specifically.

Third, input encoding — converting a continuous physical signal into spike trains — remains a comparatively under-theorized step relative to the network architectures that consume the resulting spikes. Several of the applied papers discussed above rely on an external, non-spiking feature-extraction or encoding stage before the spiking network itself begins processing, including Nesbit and colleagues' acoustic system, which performs feature extraction outside the neuromorphic core [23]. This is a pragmatic engineering choice, but it also means that a nontrivial share of the overall pipeline's energy budget can sit outside the part of the system actually being measured as "neuromorphic," and papers vary in how transparently they account for this. A more end-to-end treatment of encoding, ideally itself running on low-power event-generating hardware such as the dynamic vision sensors discussed by Gallego and colleagues, would strengthen the overall energy case considerably [25].

Fourth, most of the applied anomaly-detection systems reviewed here are evaluated on a single dataset or a small number of closely related datasets specific to their domain, and cross-domain generalization — whether a spiking architecture and training recipe that works well for vibration anomalies transfers with minimal adaptation to acoustic or network-traffic anomalies — is largely unexplored. Given that the underlying architectural pattern (spike encoding, temporally adaptive spiking network, deviation-based decision layer) recurs across the domains surveyed, this seems like a natural and currently open research direction rather than a fundamental obstacle.

Fifth, continual adaptation to drifting notions of "normal" is a capability the hardware substrate — particularly Loihi's on-chip plasticity — is explicitly designed to support, but which the applied anomaly-detection literature surveyed here has, with partial exceptions, not yet exploited fully; most of the systems described above are trained once, offline or in a training-then-deployment split, rather than continuing to adapt their internal model of normal behavior on-device after deployment [3,9]. Given that a core motivation for neuromorphic anomaly detection is long-duration, low-maintenance deployment, closing this gap between what the hardware supports and what

deployed systems actually use seems like one of the more consequential open items for the field.

7. Conclusion

The literature reviewed here supports a fairly specific and, we think, defensible claim: spiking neuromorphic architectures are a genuinely promising, and in several concrete instances already demonstrated, approach to low-power, real-time anomaly detection, grounded in both an algorithmic argument — sparse, event-driven computation naturally matches a problem defined by rare events against a quiet baseline — and a hardware argument, evidenced most rigorously by direct energy measurements such as Blouw and colleagues' roughly hundredfold energy-per-inference advantage for Loihi over a GPU on keyword spotting, that neuromorphic silicon delivers real, not merely theoretical, power savings for appropriately structured workloads [4]. Applied systems built on this foundation — Dennler and colleagues' vibration-anomaly detector, Nesbit and colleagues' acoustic fault monitor, Wang and colleagues' network intrusion detector — demonstrate that the pattern generalizes across physically very different signal domains, using a broadly consistent architecture of spike encoding, a temporally adaptive spiking network trained via surrogate gradients or conversion from a conventionally trained model, and a deviation-based decision layer [22,23,24].

What the literature does not yet support is a claim of broad, benchmarked maturity. The energy figures that exist are compelling but sparse and task-specific; the applied anomaly-detection systems are evaluated largely in isolation from one another, on their own datasets, without the shared benchmarks that would let a practitioner compare architectures directly; and the on-chip continual-learning capability that neuromorphic hardware like Loihi was explicitly built to support remains under-exploited in the anomaly-detection systems that have actually been built and evaluated to date. None of these gaps strike us as fundamental obstacles so much as an accurate description of a genuinely young applied literature sitting on top of a considerably more mature foundational one — spiking computation theory dating to the 1990s, neuromorphic hardware with a decade or more of iterative silicon development behind it, and training techniques that have only in the last several years closed most of the accuracy gap with conventional

deep learning. Given that trajectory, and the consistency with which independent groups working on physically unrelated signal types have converged on a similar architectural pattern, we read the evidence as supporting continued, deployment-focused investment in this direction rather than either premature enthusiasm or premature dismissal.

8. References

- Maass, W. (1997) Networks of Spiking Neurons: The Third Generation of Neural Network Models. *Neural Networks*, 10(9): 1659-1671. DOI: [10.1016/S0893-6080\(97\)00011-7](https://doi.org/10.1016/S0893-6080(97)00011-7)
- Merolla, P.A., Arthur, J.V., Alvarez-Icaza, R., Cassidy, A.S., Sawada, J., Akopyan, F., et al. (2014) A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface. *Science*, 345(6197): 668-673. DOI: [10.1126/science.1254642](https://doi.org/10.1126/science.1254642)
- Davies, M., Srinivasa, N., Lin, T.H., Chinya, G., Cao, Y., Choday, S.H., et al. (2018) Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. *IEEE Micro*, 38(1): 82-99. DOI: [10.1109/MM.2018.112130359](https://doi.org/10.1109/MM.2018.112130359)
- Blouw, P., Choo, X., Hunsberger, E., Eliasmith, C. (2019) Benchmarking Keyword Spotting Efficiency on Neuromorphic Hardware. *arXiv preprint*. DOI: [10.48550/arXiv.1812.01739](https://doi.org/10.48550/arXiv.1812.01739)
- Maass, W., Natschläger, T., Markram, H. (2002) Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations. *Neural Computation*, 14(11): 2531-2560. DOI: [10.1162/089976602760407955](https://doi.org/10.1162/089976602760407955)
- Diehl, P.U., Cook, M. (2015) Unsupervised Learning of Digit Recognition Using Spike-Timing-Dependent Plasticity. *Frontiers in Computational Neuroscience*, 9: 99. DOI: [10.3389/fncom.2015.00099](https://doi.org/10.3389/fncom.2015.00099)
- Indiveri, G., Liu, S.C. (2015) Memory and Information Processing in Neuromorphic Systems. *Proceedings of the IEEE*, 103(8): 1379-1397. DOI: [10.1109/JPROC.2015.2444094](https://doi.org/10.1109/JPROC.2015.2444094)
- Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Di Nolfo, C., et al. (2017) A Low Power, Fully Event-Based Gesture Recognition System. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 7243-7252. DOI: [10.1109/CVPR.2017.781](https://doi.org/10.1109/CVPR.2017.781)
- Davies, M., Wild, A., Orchard, G., Sandamirskaya, Y., Guerra, G.A.F., Joshi, P., Plank, P., Risbud, S.R. (2021) Advancing Neuromorphic Computing with Loihi: A Survey of Results and Outlook. *Proceedings of the IEEE*, 109(5): 911-934. DOI: [10.1109/JPROC.2021.3067593](https://doi.org/10.1109/JPROC.2021.3067593)
- Furber, S.B., Galluppi, F., Temple, S., Plana, L.A. (2014) The SpiNNaker Project. *Proceedings of the IEEE*, 102(5): 652-665. DOI: [10.1109/JPROC.2014.2304638](https://doi.org/10.1109/JPROC.2014.2304638)
- Schuman, C.D., Potok, T.E., Patton, R.M., Birdwell, J.D., Dean, M.E., Rose, G.S., Plank, J.S. (2017) A Survey of Neuromorphic Computing and Neural Networks in Hardware. *arXiv preprint*. DOI: [10.48550/arXiv.1705.06963](https://doi.org/10.48550/arXiv.1705.06963)
- Schuman, C.D., Kulkarni, S.R., Parsa, M., Mitchell, J.P., Date, P., Kay, B. (2022) Opportunities for Neuromorphic Computing Algorithms and Applications. *Nature Computational Science*, 2: 10-19. DOI: [10.1038/s43588-021-00184-y](https://doi.org/10.1038/s43588-021-00184-y)
- Rueckauer, B., Lungu, I.A., Hu, Y., Pfeiffer, M., Liu, S.C. (2017) Conversion of Continuous-Valued Deep Networks to Efficient Event-Driven Networks for Image Classification. *Frontiers in Neuroscience*, 11: 682. DOI: [10.3389/fnins.2017.00682](https://doi.org/10.3389/fnins.2017.00682)
- Sengupta, A., Ye, Y., Wang, R., Liu, C., Roy, K. (2019) Going Deeper in Spiking Neural Networks: VGG and Residual Architectures. *Frontiers in Neuroscience*, 13: 95. DOI: [10.3389/fnins.2019.00095](https://doi.org/10.3389/fnins.2019.00095)
- Cao, Y., Chen, Y., Khosla, D. (2015) Spiking Deep Convolutional Neural Networks for Energy-Efficient Object Recognition. *International Journal of Computer Vision*, 113(1): 54-66. DOI: [10.1007/s11263-014-0788-3](https://doi.org/10.1007/s11263-014-0788-3)
- Neftci, E.O., Mostafa, H., Zenke, F. (2019) Surrogate Gradient Learning in Spiking Neural Networks. *IEEE Signal Processing Magazine*, 36(6): 51-63. DOI: [10.1109/MSP.2019.2931595](https://doi.org/10.1109/MSP.2019.2931595)
- Zenke, F., Vogels, T.P. (2021) The Remarkable Robustness of Surrogate Gradient Learning for Instilling Complex Function in Spiking Neural Networks. *Neural Computation*, 33(4): 899-925. DOI: [10.1162/neco_a_01367](https://doi.org/10.1162/neco_a_01367)
- Deng, L., Wu, Y., Hu, X., Liang, L., Ding, Y., Li, G., Zhao, G., Li, P., Xie, Y. (2020) Rethinking the Performance Comparison Between SNNs and ANNs. *Neural Networks*, 121: 294-307. DOI: [10.1016/j.neunet.2019.09.005](https://doi.org/10.1016/j.neunet.2019.09.005)
- Pfeiffer, M., Pfeil, T. (2018) Deep Learning with Spiking Neurons: Opportunities and Challenges. *Frontiers in Neuroscience*, 12: 774. DOI: [10.3389/fnins.2018.00774](https://doi.org/10.3389/fnins.2018.00774)
- Tavanaei, A., Ghodrati, M., Kheradpisheh, S.R., Masquelier, T., Maida, A. (2019) Deep Learning in Spiking Neural Networks. *Neural Networks*, 111: 47-63. DOI: [10.1016/j.neunet.2018.12.002](https://doi.org/10.1016/j.neunet.2018.12.002)
- Roy, K., Jaiswal, A., Panda, P. (2019) Towards Spike-Based Machine Intelligence with Neuromorphic Computing. *Nature*, 575: 607-617. DOI: [10.1038/s41586-019-1677-2](https://doi.org/10.1038/s41586-019-1677-2)
- Dennler, N., Haessig, G., Cartiglia, M., Indiveri, G. (2021) Online Detection of Vibration Anomalies Using Balanced

Spiking Neural Networks. arXiv preprint. DOI: [10.48550/arXiv.2106.00687](https://doi.org/10.48550/arXiv.2106.00687)

23. Nesbit, S.C., Vergara, V.M., Felix, M.A., Kain, E.T., García Carrillo, L.R., Kunde, G.J., Sornborger, A.T. (2026) Low-Power, Neuromorphic, Acoustic Anomaly Detection for Persistent Machine Monitoring. arXiv preprint. DOI: [10.48550/arXiv.2608.18341](https://doi.org/10.48550/arXiv.2608.18341)
24. Wang, Z., Ghaleb, F.A., Zainal, A., Md Siraj, M., Lu, X. (2024) An Efficient Intrusion Detection Model Based on Convolutional Spiking Neural Network. Scientific Reports, 14: 7054. DOI: [10.1038/s41598-024-57691-x](https://doi.org/10.1038/s41598-024-57691-x)
25. Gallego, G., Delbruck, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., et al. (2022) Event-Based Vision: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(1): 154-180. DOI: [10.1109/TPAMI.2020.3008413](https://doi.org/10.1109/TPAMI.2020.3008413)
26. Yin, B., Corradi, F., Bohté, S.M. (2021) Accurate and Efficient Time-Domain Classification with Adaptive Spiking Recurrent Neural Networks. Nature Machine Intelligence, 3: 905-913. DOI: [10.1038/s42256-021-00397-w](https://doi.org/10.1038/s42256-021-00397-w)

9. Declarations

Funding

This literature synthesis received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that they have no known competing interests.

Data Availability Statement

This article is a literature synthesis and did not generate new primary data; all quantitative figures discussed are drawn from, and attributed to, the cited sources, whose own data and code availability statements should be consulted directly.

Ethical Approval

Not applicable. This work did not involve human participants, animal subjects, or the collection of any primary experimental data.