Research Square

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

An immunological synapse network for recognition of persistent weak anomalies

Sherif Tawfik

sherif.tawfic@gmail.com

Deakin University https://orcid.org/0000-0003-3592-1419

Hung Le

Deakin University

Svetha Venkatesh

Deakin University

Research Article

Keywords:

Posted Date: August 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10808625/v1

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

Additional Declarations: The authors declare no competing interests.

An immunological synapse network for recognition of persistent weak anomalies

Sherif Abdulkader Tawfik[1], Hung Le[1], and Svetha Venkatesh[1]

[1]Applied Artificial Intelligence Initiative

Abstract

Many time-series anomalies of practical consequence are not sharp spikes but weak deviations sustained across many consecutive samples. Standard deep-learning encoders, such as convolutional, recurrent and Transformer architectures, were not designed to distinguish whether faint evidence is transient or persistent. We introduce the Immunological Synapse Network (ISN), a differentiable architecture that maps four T-cell recognition mechanisms onto anomaly detection: a learned receptor scores local signal tokens by binding affinity; a kinetic proofreading cascade suppresses short-lived matches by requiring evidence to survive multiple sequential states; ITAM-like gates impose a learned threshold on each receptor; and a synapse-inspired spatial lattice concentrates surviving activations before classification. On a controlled synthetic benchmark whose label is recoverable only through temporal persistence, ISN is the only model that does not collapse to chance when its proofreading cascade is disabled, and it outperforms all baselines in full-model comparison. These results establish a clear operating boundary: immune-inspired persistence filtering earns its computational cost when anomalies are defined by temporal continuity, not by instantaneous magnitude.

Introduction

In many sensing domains, the most important anomalies are weak. An early arrhythmia, the cardiac signature of a sleep-apnea episode, an incipient machine fault or a spacecraft mode shift typically appears as low-amplitude evidence smeared across many samples rather than as a single decisive spike. Such patterns are hard to detect: they are rare, they sit in noise, and stronger transient fluctuations can mimic them over short windows. A useful detector therefore needs two things at once: sensitivity to subtle local structure, and a way to test whether that structure persists long enough to warrant an anomalous call.

Deep learning handles temporal classification and anomaly detection well, particularly with convolutional, recurrent and Transformer encoders [1–5]. None of these architectures, though, was designed with weak persistent anomalies in mind. A convolutional network looks for local, translation-invariant motifs [6, 7]; a recurrent network folds the past into a running hidden state [3, 4]; a Transformer weighs every pair of tokens against every other [5]. Each is powerful, yet, as our results will show, none has a built-in test for the property that defines this regime: whether weak evidence was sustained rather than fleeting. Classical matched filters and rule-based detectors [8] can encode persistence heuristics, but they do not learn task-specific features and lean on hand-tuned thresholds.

Adaptive immunity solves a recognition problem of the same shape. A T lymphocyte must pick out foreign peptide–major histocompatibility complex (pMHC) ligands from a vast background of self-peptide complexes that differ by only a few atoms, while avoiding both missed pathogens and attacks on healthy tissue [9, 10]. Several stages contribute to the decision, none of them decisive on its own. The T-cell receptor reads out binding affinity on first contact; a kinetic proofreading cascade [9] then discards short-lived, low-affinity contacts by demanding K sequential phosphorylation events that finish only while the receptor–ligand complex stays intact; activation thresholds set by immunoreceptor tyrosine-based activation motifs (ITAMs) keep weak signals from propagating; and the engaged receptors are organised spatially at the immunological synapse [10, 11], concentrating scattered evidence before the cell commits.

We map these four mechanisms onto a differentiable architecture, the Immunological Synapse Network (ISN). In ISN, (i) a learned receptor repertoire estimates the affinity of local signal tokens; (ii) a kinetic proofreading cascade suppresses transient or noise-driven activations by requiring evidence to survive several sequential states; (iii) ITAM-like gates impose a per-receptor activation threshold; and (iv) a synaptic lattice arranges the surviving receptor evidence into a spatial map before classification. The design keeps fast affinity estimation separate from the slower, threshold-gated accumulation of evidence, which is what suits ISN to anomalies defined by persistence rather than by instantaneous amplitude.

We present ISN and evaluate it broadly. We benchmark it against convolutional, recurrent and Transformer encoders on a synthetic persistence task and on seven real-world datasets spanning cardiac, respiratory, neuromuscular, acoustic and spacecraft-telemetry signals, and we read off from the pattern of results when persistence filtering helps. ISN gave the best F1, AUROC and AUPRC on the synthetic test and on four of the seven real-world benchmarks: all four ECG arrhythmia tasks, photoplethysmography event detection, sleep apnea detection and spacecraft telemetry, with margins that held across random seeds and held-out splits. It was level with the best baseline on electromyography, and it fell behind on the two datasets whose discriminative signal is already condensed into per-sample statistics (network intrusion) or coarse acoustic summaries (machine sound),

where raw-token persistence is not what carries the information. Together these results show ISN to be a strong model for weak persistent anomalies and mark out the conditions under which its immune-inspired prior carries to a new domain.

Results

The Immunological Synapse Network

ISN treats a one-dimensional signal window as a stream of antigens to be screened. A multiscale convolutional front end turns the raw window into a short sequence of feature vectors, or tokens, each summarising the local waveform at three temporal scales (Fig. 1a). Every token is then matched against a repertoire of R learned receptor prototypes, and the match is scored as a binding affinity, by analogy with how a T-cell receptor reads the affinity of a peptide–MHC contact.

The affinities do not act directly. They drive a differentiable kinetic proofreading cascade of $K+1$ states (Fig. 1b), modelled on the mechanism by which T cells reject brief, low-affinity contacts. Binding mass enters the first state and either advances toward the terminal state or drains back to the start; only tokens whose affinity stays high for long enough push mass all the way through, while transient or noisy matches bleed out first. A learned gate blends this slow, persistence-filtered signal with the fast direct-affinity estimate, so the network can set its own balance between sensitivity and specificity for each task. A per-receptor ITAM-like threshold then silences receptors whose accumulated evidence stays below a learned level.

Finally, the surviving receptor activity is laid out on a learned two-dimensional lattice that imitates the immunological synapse: when overall activation is high, token positions contract toward the lattice centre, echoing the central supramolecular activation cluster, so strong evidence concentrates rather than disperses. The lattice, together with a parallel residual convolutional branch, feeds a compact classification head. Full equations and the biological correspondence for each stage are given in Methods.

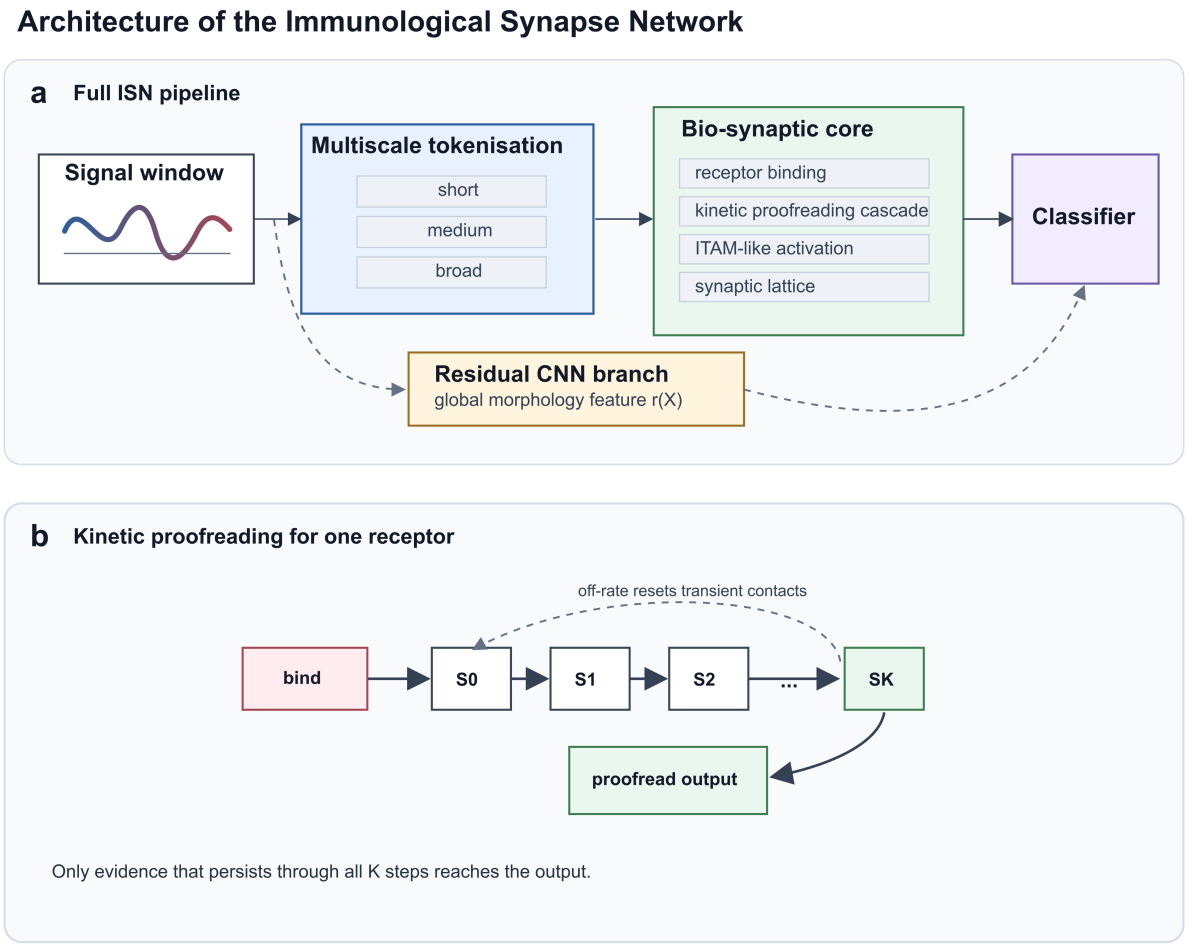


Figure 1: **Architecture of the Immunological Synapse Network. a**, Overview of the ISN pipeline. A signal window undergoes multiscale tokenisation, followed by the bio-synaptic core (receptor binding, kinetic proofreading cascade, ITAM gating, synaptic lattice), a parallel residual CNN branch and a classification head. **b**, Detail of the kinetic proofreading cascade for a single receptor. Binding mass from the token stream enters the first cascade state; only mass that persists across all K phosphorylation steps reaches the proofread output u^{KPR}. Transient contacts are returned to the entry state by the off-rate k^{off}.

The seven real-world tasks are ECG arrhythmia classification [12, 13], photoplethysmography [14], surface electromyography [15], sleep-apnea detection from raw ECG [16], network intrusion [17], machine-sound anomaly detection [18] and spacecraft telemetry [19]. They span enough modalities to show whether the persistence-filtering prior travels beyond any single signal type, and where it breaks down.

We compared ISN with four baselines: a residual convolutional network (CNN), a gated recurrent unit (GRU) [4], a long short-term memory network (LSTM) [3] and a Transformer encoder [5]. All were trained under an identical protocol, using the same preprocessing, splits, optimiser, loss and five random seeds, and all operate on the standardised input window directly.

Table 1: **Synthetic weak-evidence stress test.** Models were evaluated on fixed synthetic splits over three seeds. Positive windows contain a longer uninterrupted weak-motif train; negative windows contain the same weak-event budget split into short reset-separated bursts. ISN no-KPR disables the KPR cascade and uses direct binding alone.

Model / condition	F1	AUROC	AUPRC
CNN	0.799±0.004	0.864±0.008	0.857±0.012
GRU	0.679±0.000	0.493±0.007	0.515±0.009
Transformer	0.679±0.001	0.514±0.002	0.522±0.003
ISN no-KPR	0.812±0.093	0.793±0.214	0.802±0.203
ISN KPR	**0.849±0.003**	**0.926±0.004**	**0.928±0.004**

Persistent weak anomalies and kinetic proofreading

We first built a synthetic task aimed at the regime kinetic proofreading (KPR) is meant for. The task has two classes, both of which contain the same number of weak motifs and the same large transient distractors. What differs is timing: positive windows hold one long, uninterrupted train of weak motifs, while negative windows split the same weak-event budget into short bursts separated by reset gaps. On this test ISN with KPR was both the most accurate and the most stable model (AUPRC 0.928±0.004, AUROC 0.926±0.004, F1 0.849±0.003; Table 1). Switching KPR off, with the splits and all other ISN settings held fixed, dropped mean AUPRC to 0.802±0.203 and made the result swing widely from seed to seed. CNN picked up part of the local burst structure (AUPRC 0.857±0.012) but stayed below ISN with KPR. GRU and Transformer mostly collapsed to predicting the positive class, which inflates their F1 relative to their near-chance AUROC and AUPRC. This is exactly the behaviour KPR is built to produce: keep weak evidence that persists, and reject strong evidence that only flickers or arrives in reset-separated bursts.

In the full model, where a parallel residual branch can also fit the synthetic timing rule, kinetic proofreading had a smaller but consistent marginal effect. Varying proofreading depth (Table 2) increased AUPRC from 0.956±0.007 with the cascade disabled to a shallow optimum of 0.962±0.004 at $K = 3$, after which performance plateaued. A companion component ablation on the same full-model setting (Table 3) showed the same pattern: removing KPR alone reduced AUPRC by 0.005, whereas removing the residual branch or all selected support components caused much larger drops. Thus the strict KPR-isolation test in Table 1 asks whether KPR alone can solve the persistence rule; the full-model ablations ask how much KPR adds once residual and direct-binding fallback routes are also available.

Table 2: **Kinetic-proofreading depth on the synthetic stress test.** Full ISN with a varying number of proofreading layers K (K=off disables the cascade), on fixed synthetic splits over ten seeds. Bold marks the best mean.

K	F1	AUROC	AUPRC
off	0.892±0.007	0.962±0.005	0.956±0.007
1	0.893±0.009	0.964±0.004	0.960±0.005
2	**0.897±0.004**	0.965±0.004	0.961±0.004
3	0.897±0.003	**0.966±0.003**	**0.962±0.004**
4	0.895±0.004	0.965±0.003	0.960±0.003
5	0.896±0.008	0.965±0.002	0.961±0.003

Table 3: **Component ablation on the synthetic stress test.** Each row removes one mechanism from the full ISN; fixed splits, ten seeds. This full-model analysis complements the strict KPR-isolation comparison in Table 1.

Variant	F1	AUROC	AUPRC
Full ISN (K=3)	**0.897±0.003**	**0.966±0.003**	**0.962±0.004**
No KPR	0.892±0.007	0.962±0.005	0.956±0.007
No KPR/direct mixture	0.889±0.009	0.960±0.004	0.954±0.005
No temporal readout	0.866±0.071	0.915±0.136	0.906±0.139
No residual branch	0.830±0.090	0.868±0.170	0.861±0.174
No support (mix+temporal+res)	0.653±0.000	0.522±0.018	0.510±0.014

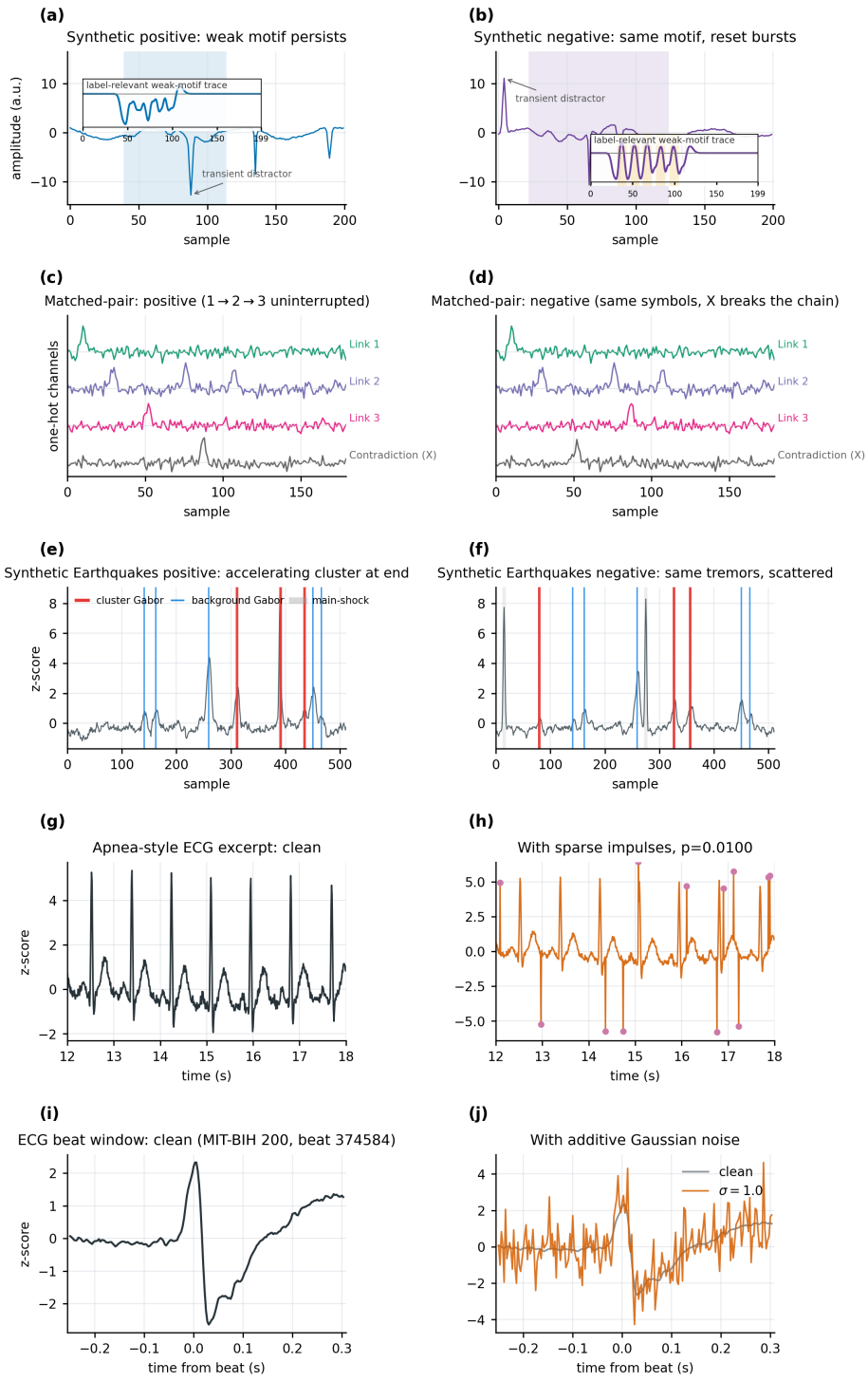


Figure 2: **Representative signal examples for the robustness analyses.** (a,b) Synthetic weak-evidence task; inset shows only the label-relevant weak-motif trace with the background, decoys and distractors removed. (c,d) Chain-control task ($K = 3$); the four one-hot symbol channels (Link 1, 2, 3, contradiction X) are vertically offset for visualisation; cf. Table 4. (e,f) Synthetic foreshock-cluster task; red bars mark the $K = 3$ cluster Gabor tremors, blue bars the matched-count background tremors, grey shading the matched main-shock-like distractor spikes; cf. Table 6. (g,h) Apnea-style ECG window before/after sparse impulse noise at $p = 0.0100$; cf. Table 7. (i,j) MIT-BIH ECG beat window before/after additive Gaussian noise at $\sigma = 1.0$; cf. Table 9.

Table 4: **Component ablation on the chain-control task (K=3).** Per-symbol counts, pairwise co-occurrences and motif budget are matched across classes; only order-K contiguity differs. Ten seeds, 8,000 train / 2,000 val / 2,000 test.

Variant	F1	AUROC	AUPRC
Full ISN, KPR K=1	0.831±0.169	0.757±0.243	0.752±0.248
Full ISN, KPR K=2	0.798±0.165	0.706±0.240	0.701±0.244
Full ISN, KPR K=3	0.797±0.165	0.707±0.240	0.702±0.244
Full ISN, KPR K=4	0.798±0.165	0.707±0.240	0.701±0.244
Full ISN, KPR K=5	0.798±0.165	0.708±0.239	0.702±0.244
No KPR	0.663±0.000	0.505±0.013	0.498±0.011
No KPR/direct mixture	0.899±0.154	0.848±0.232	0.847±0.234
No temporal readout	0.662±0.001	0.505±0.011	0.498±0.011
No residual branch	**0.966±0.101**	**0.964±0.109**	**0.963±0.112**
No support (mix+temporal+res)	0.663±0.000	0.745±0.103	0.710±0.101

Chain-control task

The weak-evidence task above mixes two confounds: positives and negatives differ both in motif density and in the order in which motifs appear. To isolate order from density we defined the chain-control task. Two symbol sequences over the alphabet $\{1, \ldots, K, X\}$ are constructed so that, for every $m \leq K-1$, they carry *identical* m-gram counts; only one difference remains, an order-K property: the positive sequence contains the contiguous chain $1 \to 2 \to \cdots \to K$ uninterrupted, whereas the negative one inserts a contradiction symbol X between two consecutive links so that the longest reset-free run is at most $K-1$. Each symbol is rendered as a unit-height Gaussian bump on its own one-hot channel under per-channel Gaussian noise σ=0.15, spaced by gaps much larger than the receptive field of the multiscale tokeniser, so detecting single motifs is trivial but composing them across the window is not. By construction, any order-$\leq (K-1)$ feature (per-symbol counts, ordered pair co-occurrence) is at chance and only the order-K contiguity discriminates. We used K=3, ten seeds, and disabled the residual CNN branch so the score reflects what the immune core alone composes.

KPR was the single load-bearing mechanism on this task (Table 4). Disabling the cascade collapsed ISN to chance (AUPRC 0.498±0.011); removing the temporal readout did the same. Removing the residual branch was *the most beneficial* ablation (AUPRC 0.963±0.112), because the residual branch supplies a non-cascading shortcut that competes with the immune core for the order-K decision. Varying KPR depth $K_{\text{model}} \in \{1, \ldots, 5\}$ with the residual branch removed produced no end-to-end depth peak (all means within ±0.005 of 0.96), matching the prediction that a sufficiently expressive readout absorbs the depth burden regardless of the cascade's nominal length.

Repeating the depth sweep with the residual branch always disabled (Table 5) confirmed the absence of an end-to-end optimum: every depth $K_{\text{model}} \in \{1, 2, 3, 4, 5\}$ landed within ±0.003 of 0.96 AUPRC, with the high seed-level variance (s.d. $\approx$ 0.11) characteristic of an

Table 5: **KPR depth sweep on the chain-control task with the residual branch disabled.** Closing the residual leak is necessary to expose the cascade's contribution but is not sufficient to produce a depth-K peak: every depth tied within seed noise. Ten seeds. The K=3 row reproduces the *No residual branch* entry of Table 4.

K_{model}	F1	AUROC	AUPRC
1	0.966±0.101	0.963±0.112	0.964±0.110
2	**0.966±0.101**	**0.965±0.105**	**0.964±0.108**
3	0.966±0.101	0.964±0.109	0.963±0.112
4	0.966±0.101	0.962±0.114	0.961±0.117
5	0.966±0.101	0.959±0.122	0.959±0.124

under-determined readout that solves the task at any depth on the good seeds and not at all on the bad ones. End-to-end the cascade therefore matters as a mechanism (no-KPR collapses) but its depth does not, because a sufficiently expressive readout absorbs the depth burden regardless of the cascade's nominal length. Channel noise alone (raised from 0.15 to 0.50) did not recover the peak either: all depths $K \geq 1$ landed at ≈ 0.83 AUPRC, because single-bump detection on dedicated channels remains tractable enough that one cascade step saturates.

Synthetic foreshock-cluster analogue of UCR Earthquakes

The UCR Earthquakes benchmark asks whether a major seismic event will follow within the next 24 h given a 512-sample window of single-channel hourly seismicity. The predictive structure, well established in foreshock studies, is a temporally accelerating cluster of small precursor tremors arriving at shrinking inter-event times in the hours before a main shock; isolated tremors, distant aftershocks and class-independent large spikes are not predictive. We constructed a synthetic analogue that preserves this structure without the noise of a real catalogue. Each window contains a balanced budget of small Gabor pulses drawn from a log-normal amplitude distribution, plus one to two class-independent main-shock-like spikes ($5-8\sigma$). Both classes carry identical counts, identical amplitudes and identical per-window mean, max and standard deviation after per-sequence z-scoring. They differ only in the temporal arrangement of K=3 cluster pulses: positives place them as a forward-decreasing-gap cluster ($\tau = 80$, ratio $r = 1.8$, so gaps $80, 44$) ending at $0.85 \cdot T$; negatives place the same three pulses at random scattered positions. An order-K oracle (longest reset-free acceleration combined with end-anchoring) scores AUC ≈ 1.0, while single-point statistics (max-abs, mean-abs, std, count) score ≈ 0.5.

Trained end-to-end with the residual branch disabled and per-class balanced positives (three seeds), the model produced a clear $\cap$-shape across KPR depths (Table 6). Removing KPR brought the model essentially to chance (AUPRC 0.541±0.018). Cascade depths K=2 and K=3 tied as the best at ≈ 0.62 AUPRC; K=1, K=4 and K=5 all sat near 0.60. The peak being at K=2 rather than K=3 is consistent with the cascade needing

Table 6: **KPR depth sweep on the synthetic foreshock-cluster task.** All variants disable the residual branch. Three seeds, 4,000 train / 1,000 val / 1,000 test, $T = 512$, balanced positives.

K_{model}	F1	AUROC	AUPRC
off (no KPR)	0.681±0.000	0.522±0.020	0.541±0.018
1	0.681±0.001	0.598±0.079	0.604±0.067
2	**0.701±0.028**	**0.620±0.108**	**0.624±0.096**
3	0.693±0.017	0.609±0.093	0.615±0.083
4	0.686±0.007	0.598±0.077	0.602±0.065
5	0.681±0.000	0.601±0.084	0.607±0.073

$K-1$ transitions to fire on the last cluster motif: the readout then integrates whether the terminal state was reached anywhere in the window. Absolute AUPRCs are lower than on the chain-control task because the discriminating feature lives among same-distribution background pulses with no exclusive channel, which is the real-world difficulty Earthquakes was built to capture.

Robustness to impulse artifacts in apnea ECG

How does ISN hold up when the training data is corrupted? We added sparse, high-amplitude impulses to the apnea training split only, after window standardisation, as bipolar shifts of magnitude 5.0. The fraction of perturbed points is p; since each apnea window has 3000 samples, the heaviest setting ($p = 0.0100$) injects about 30 impulses per window.

ISN held the highest AUPRC of the five models at every impulse level (Table 7). Its accuracy did not fall off as impulses grew denser: AUPRC was 0.695 ± 0.101 with no impulses and 0.720 ± 0.106 at $p = 0.0100$. We read this as insensitivity to sparse training artifacts, not as a sign that heavier corruption helps. CNN was the best non-ISN model throughout, but it slid from AUPRC 0.626 ± 0.153 at $p = 0$ to 0.540 ± 0.112 at $p = 0.0100$.

ECG arrhythmia detection

We drew four binary tasks from the MIT-BIH Arrhythmia Database, each pitting one rhythm against all others: normal sinus rhythm (NSR), atrial fibrillation (AF), ventricular tachycardia (VT) and premature ventricular contractions (PVC).

ISN had the highest F1 on all four (Table 8). The margin was widest for PVC, where ISN reached F1 $= 0.855 \pm 0.011$ against 0.771 ± 0.021 for the best non-ISN baseline, a gain of 0.084 in F1 (10.9% relative). For NSR, AF and VT the margins were smaller but held across all five seeds. AUROC and AUPRC were high everywhere (AUROC above 0.99 for NSR and AF), so the F1 gaps were not merely an artefact of where the decision threshold fell. Which baseline came closest depended on the task; no single conventional architecture

Table 7: **Apnea-ECG sparse impulse-noise training sweep.** Sparse bipolar impulses of amplitude 5.0 were injected into training windows only. Each row reports mean ± s.d. over three subject-grouped folds. Bold indicates the largest mean within each metric at a fixed corruption level.

Expected impulses/window	Model	F1	AUROC	AUPRC
0.0	ISN	0.537±0.116	**0.740±0.104**	**0.695±0.101**
	CNN	0.437±0.239	0.697±0.100	0.626±0.153
	LSTM	**0.603±0.041**	0.622±0.067	0.462±0.106
	GRU	0.495±0.050	0.530±0.057	0.403±0.022
	Transformer	0.553±0.048	0.640±0.080	0.494±0.098
1.5	ISN	**0.564±0.102**	**0.790±0.079**	**0.719±0.086**
	CNN	0.441±0.243	0.704±0.099	0.618±0.154
	LSTM	0.523±0.045	0.561±0.115	0.463±0.114
	GRU	0.488±0.034	0.610±0.066	0.462±0.031
	Transformer	0.496±0.069	0.593±0.068	0.435±0.070
3.0	ISN	0.552±0.098	**0.721±0.115**	**0.676±0.117**
	CNN	0.439±0.247	0.694±0.103	0.610±0.152
	LSTM	0.547±0.026	0.523±0.072	0.415±0.111
	GRU	**0.560±0.087**	0.589±0.106	0.435±0.064
	Transformer	0.495±0.063	0.610±0.074	0.463±0.065
7.5	ISN	**0.543±0.117**	**0.700±0.119**	**0.663±0.100**
	CNN	0.426±0.240	0.692±0.106	0.619±0.152
	LSTM	0.543±0.056	0.639±0.071	0.503±0.089
	GRU	0.533±0.060	0.583±0.097	0.436±0.058
	Transformer	0.490±0.085	0.544±0.098	0.418±0.075
15.0	ISN	0.547±0.133	0.679±0.124	**0.668±0.104**
	CNN	0.430±0.224	**0.722±0.077**	0.625±0.127
	LSTM	**0.598±0.035**	0.648±0.071	0.532±0.110
	GRU	0.544±0.091	0.549±0.118	0.407±0.022
	Transformer	0.511±0.093	0.565±0.110	0.429±0.095
30.0	ISN	**0.601±0.087**	**0.765±0.126**	**0.720±0.106**
	CNN	0.436±0.245	0.673±0.084	0.540±0.112
	LSTM	0.542±0.019	0.556±0.036	0.436±0.039
	GRU	0.544±0.098	0.627±0.129	0.479±0.044
	Transformer	0.473±0.060	0.525±0.116	0.443±0.130

led on all four arrhythmias.

Photoplethysmography and electromyography

To see whether the prior carries past cardiac signals, we ran the same five-model protocol on PPG and EMG. On PPG, ISN again had the best F1 (0.727±0.084 versus 0.696±0.089 for the next-best model), in line with the ECG results. On EMG it did not come first: CNN edged it on both F1 (0.890±0.021 versus 0.880±0.027) and AUPRC (0.969±0.003 versus 0.965±0.004), though both gaps were within a standard deviation. PPG event detection shares much of its structure with arrhythmia detection, so a persistence-filtering benefit there is unsurprising. The EMG near-tie is a useful counterweight: the immune-inspired prior does not automatically beat a plain convolution, and whatever edge it has seems to

Table 8: **Test performance across all tasks and models (mean ± s.d., five seeds).** Bold marks the best value in each row.

Task	Metric	ISN	CNN	LSTM	GRU	Transformer
ECG NSR	F1	**0.971±0.001**	0.955±0.003	0.935±0.029	0.961±0.002	0.900±0.004
	AUROC	**0.994±0.000**	0.984±0.001	0.964±0.032	0.988±0.002	0.949±0.005
	AUPRC	**0.996±0.000**	0.989±0.001	0.973±0.026	0.992±0.001	0.964±0.005
ECG AF	F1	**0.957±0.002**	0.928±0.005	0.831±0.045	0.846±0.091	0.773±0.008
	AUROC	**0.998±0.000**	0.996±0.000	0.978±0.013	0.979±0.025	0.964±0.003
	AUPRC	**0.992±0.001**	0.979±0.002	0.899±0.054	0.902±0.111	0.874±0.010
ECG VT	F1	**0.916±0.003**	0.883±0.003	0.753±0.027	0.838±0.071	0.752±0.018
	AUROC	**0.993±0.001**	0.986±0.000	0.941±0.013	0.975±0.022	0.957±0.006
	AUPRC	**0.966±0.003**	0.927±0.004	0.800±0.035	0.897±0.069	0.836±0.015
ECG PVC	F1	**0.855±0.011**	0.771±0.021	0.534±0.034	0.685±0.030	0.518±0.039
	AUROC	**0.996±0.002**	0.995±0.001	0.975±0.006	0.992±0.001	0.982±0.002
	AUPRC	**0.945±0.006**	0.924±0.002	0.684±0.077	0.906±0.013	0.801±0.021
PPG	F1	**0.727±0.084**	0.605±0.098	0.600±0.155	0.696±0.089	0.535±0.098
	AUROC	**0.935±0.026**	0.878±0.022	0.783±0.183	0.893±0.091	0.807±0.043
	AUPRC	**0.832±0.093**	0.700±0.091	0.614±0.247	0.790±0.110	0.521±0.121
EMG	F1	0.880±0.027	**0.890±0.021**	0.839±0.028	0.869±0.033	0.872±0.027
	AUROC	0.920±0.014	**0.931±0.011**	0.857±0.030	0.917±0.015	0.912±0.023
	AUPRC	0.965±0.004	**0.969±0.003**	0.928±0.017	0.961±0.005	0.960±0.009
Apnea	F1	**0.599±0.129**	0.471±0.188	0.527±0.082	0.523±0.118	0.547±0.089
	AUROC	**0.737±0.111**	0.673±0.091	0.602±0.116	0.642±0.093	0.666±0.078
	AUPRC	**0.685±0.122**	0.593±0.131	0.486±0.128	0.510±0.137	0.531±0.117
Net. intrusion	F1	0.931±0.003	0.937±0.001	0.962±0.001	0.962±0.001	**0.974±0.002**
	AUROC	0.984±0.001	0.987±0.001	0.992±0.001	0.993±0.001	**0.996±0.000**
	AUPRC	0.980±0.001	0.983±0.001	0.990±0.000	0.991±0.001	**0.995±0.000**
Mach. sound	F1	0.340±0.014	0.350±0.012	0.308±0.005	0.339±0.018	**0.394±0.010**
	AUROC	0.640±0.033	0.674±0.022	0.570±0.007	0.634±0.035	**0.741±0.008**
	AUPRC	0.266±0.045	0.351±0.032	0.208±0.005	0.273±0.047	**0.439±0.016**
Spacecraft	F1	**0.698±0.029**	0.627±0.011	0.389±0.012	0.542±0.015	0.477±0.016
	AUROC	0.901±0.017	**0.905±0.009**	0.726±0.014	0.846±0.004	0.817±0.012
	AUPRC	**0.726±0.034**	0.708±0.012	0.340±0.023	0.559±0.013	0.517±0.008

depend on the class balance and the temporal shape of the target events.

Sleep apnea detection from raw ECG

Sleep apnea leaves a subtle but characteristic mark on ECG morphology and heart-rate variability that builds up over many consecutive breath cycles, the same persistence regime ISN targets, only on a longer timescale. ISN took the top F1, AUROC and AUPRC here (F1 = 0.599±0.129, AUROC = 0.737±0.111, AUPRC = 0.685±0.122; Table 8). The Transformer was the best baseline (F1 = 0.547±0.089) and CNN the worst (F1 = 0.471±0.188). The 0.052 F1 gap between ISN and the Transformer held in every fold (per-fold range 0.028 to 0.117). The large standard deviations come from real differences between subjects in apnea severity and ECG shape, not from an unstable protocol. The

advantage thus reaches from short arrhythmia bursts to events that play out over much longer physiological timescales.

To isolate the contribution of the kinetic-proofreading cascade on this task, we ran a paired ablation in which the same ISN architecture was trained with and without KPR enabled, on the same Apnea-ECG fold (40,627 test windows, 36.9% positive prevalence). The two models produce strikingly different score distributions (Fig. 3): with KPR, negatives pile near zero and positives concentrate sharply near one, giving a clean bimodal separation; without KPR, both classes collapse toward zero with heavy overlap and only a weak positive tail. At the per-fold optimal-F1 threshold this translates into a consistent gain across every metric: accuracy rises from 0.801 to 0.827, precision from 0.800 to 0.892 at essentially unchanged recall (0.614 $\rightarrow$ 0.605), F1 from 0.695 to 0.721, AUROC from 0.796 to 0.862 and AUPRC from 0.766 to 0.827. The proofreading cascade is therefore not cosmetic on long-timescale physiological events: it converts a diffuse, low-confidence score surface into a separable one, and the binary predictions inherit the gain.

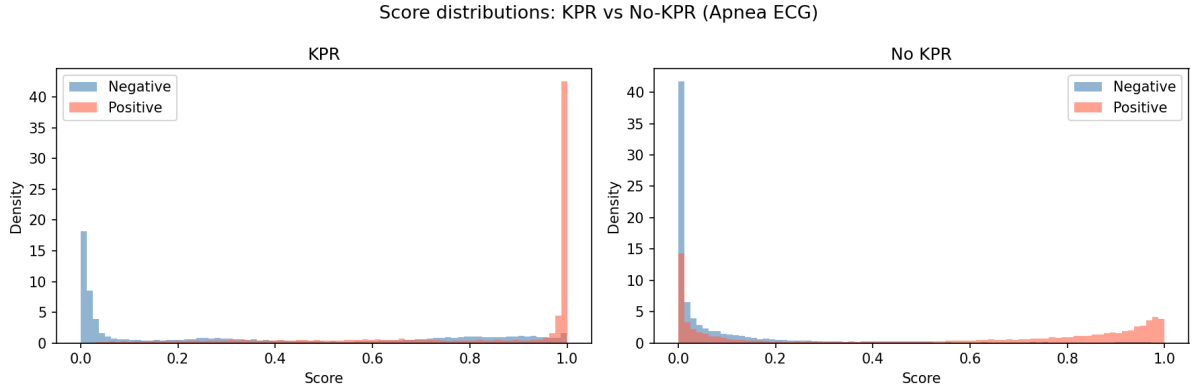


Figure 3: **KPR sharpens the score surface on Apnea-ECG.** Class-conditional score histograms on the test fold (40,627 windows; positives 36.9%) for the same ISN architecture with KPR enabled (left) and disabled (right). KPR yields a clean bimodal separation, with negatives concentrated near zero and positives near one; without KPR, both classes pile up near zero with strong overlap. At the optimal-F1 threshold this lifts F1 from 0.695 to 0.721, accuracy from 0.801 to 0.827, AUROC from 0.796 to 0.862 and AUPRC from 0.766 to 0.827.

Network-flow intrusion detection

To find where the ISN prior stops paying off, we ran the same protocol on the CICIDS2017 intrusion-detection benchmark [17]. This dataset is built differently from the physiological tasks: each sample is a 64-step window of 77 pre-computed flow statistics (packet counts, byte rates, inter-arrival-time distributions, TCP flag counts) taken from completed network flows, not raw waveforms. The task is benign traffic against any attack label, over 176,893 windows from eight day-long captures (36.3% attack rate), with a stratified train/validation/test split.

Here ISN came last of the five (F1 = 0.931±0.003, AUROC = 0.984±0.001; Table 8). The Transformer led at F1 = 0.974±0.002, GRU and LSTM sat in between (F1 ≈ 0.962), and even CNN edged ISN (F1 = 0.937±0.001). The 0.043 F1 gap to the best model repeated across all five seeds (per-seed range 0.038 to 0.050), so it is not seed noise, and validation-to-test gaps of ≤0.003 F1 rule out overfitting.

The shortfall fits a domain mismatch. Kinetic proofreading rewards tokens that keep their affinity high across a sequential cascade, so its power grows with the length and consistency of a token run, which is just what physiological anomalies supply: a PVC holds its aberrant morphology over many token positions, and an apnea event runs across many breath cycles. A CICIDS2017 row, by contrast, is already a summary of a finished flow, so the discriminative signal lives in cross-channel relationships at a single timestep rather than in anything that persists over time. Self-attention, which compares all token pairs equally and assumes nothing about persistence, fits that structure better. The lesson is a clear boundary for ISN: it helps when the anomaly is a sustained waveform pattern, and not when the signal has already been reduced to per-sample statistics.

Machine acoustics

We also ran the protocol on the DCASE machine-sound anomaly-detection benchmark. Every model did badly: the best F1 was 0.393±0.010 (Transformer) and the best AUROC 0.741±0.008, well short of the physiological tasks. ISN landed third (F1 = 0.340±0.014, AUROC = 0.640±0.033), tied with GRU (F1 = 0.339) and within one standard deviation of CNN (F1 = 0.350), with LSTM last. The Transformer beat the field on every seed by 0.037 to 0.077 F1. Validation-to-test gaps stayed ≤0.015 F1, so the low scores are about the task and its representation, not overfitting.

That every architecture struggled points to a bottleneck in the five-descriptor input. DCASE sound benchmarks are usually competitive only with spectral features such as mel spectrograms or MFCCs, which keep the frequency-domain texture; collapsing audio to five scalars per frame throws away most of what separates normal from faulty machine operation. ISN's mid-pack finish, ahead of LSTM, level with GRU and CNN, behind the Transformer, says that the persistence prior buys little once the features are this coarse, because there is no fine morphology left for the cascade to reward. The Transformer probably gains from comparing distant frames directly, catching slow shifts in envelope shape that local-window models cannot. This is a different failure from network intrusion: the input here is genuinely temporal, and what limits ISN is feature resolution rather than a real mismatch with its inductive bias. Feeding ISN raw mel spectrograms would be the fairer test in this domain.

Spacecraft telemetry

The NASA SMAP/MSL telemetry benchmark [19] runs the other way from the network-flow and acoustic tasks. ISN was first on F1, AUPRC and accuracy on all five seeds (F1 = 0.698±0.029, AUROC = 0.901±0.017; Table 8). CNN came second (F1 = 0.627±0.011) with almost the same AUROC (0.905±0.009); the 0.071 F1 gap held on every seed (range 0.056 to 0.093). GRU was third (F1 = 0.542), with Transformer and LSTM fourth and fifth (0.477 and 0.389). Validation-to-test gaps were $\leq$0.007 F1 for ISN and $\leq$0.014 across all models, again pointing away from overfitting.

This ordering inverts the intrusion result and shows where the ISN prior belongs. Telemetry is raw sensor data sampled at fixed intervals, and faulty states, whether thruster faults, sensor saturation or unexpected mode transitions, last for many consecutive timesteps. A sustained deviation keeps affinity high through the cascade, so mass reaches the terminal state and the output is large; a brief spike dies out before it gets there. The Transformer, top model on both pre-computed-feature tasks, drops to fourth here: with 82 entity groups and only about 191 samples each, global self-attention has too little data to generalise across entities, while ISN's local persistence prior carries over between entities without needing large per-entity training sets. LSTM trailing GRU by 0.153 F1, on sequences of just 128 steps, suggests the recurrent gating details matter, and ISN's explicit cascade accumulation beat both recurrent variants.

Noise robustness

Finally, we tested robustness to additive Gaussian noise injected during training, again on the training split only and after window standardisation; $\sigma = 1.0$ matches the standard deviation of the normalised input. Each condition used three seeds.

At $\sigma = 1.0$, ISN kept most of its clean-test F1 on PPG, EMG, machine sound, network intrusion and industrial vibration (Table 9). Spacecraft telemetry was the exception, where ISN lost 0.200 F1. On the ECG sweep, ISN gave up 0.127 mean F1, against 0.169 for CNN under the same noise.

Discussion

Taken together, the experiments point to a clear but bounded role for immune-inspired persistence filtering: ISN works best when the thing to be detected is a sustained deviation in a raw sensor stream. Its three borrowed mechanisms each do a different job. Receptor affinity estimation gives motif selectivity; kinetic proofreading throws out transient and noise-driven evidence; and the synaptic lattice pools evidence across receptors before a class is chosen. The ITAM-like gate adds a learned per-receptor floor, so weak chance matches count for little.

Table 9: **Training-noise robustness across completed sweeps.** Values are noisy-minus-clean mean F1 across tasks within each benchmark over three seeds. All rows use $\sigma = 1.0$. Dashes indicate models not present in the corresponding noise run.

Benchmark	Highest σ	ISN	CNN	LSTM	GRU	Transformer
ECG arrhythmia	1.00	**−0.127**	−0.169	–	–	–
PPG	1.00	**−0.009**	−0.074	−0.021	+0.074	−0.445
EMG	1.00	−0.008	−0.064	+0.007	**+0.006**	−0.068
Apnea	1.00	+0.017	+0.131	−0.004	**+0.001**	−0.122
Industrial vibration	1.00	**+0.000**	**+0.000**	−0.149	−0.075	−0.312
Machine sound	1.00	−0.011	−0.013	**+0.001**	+0.003	−0.036
Network intrusion	1.00	−0.042	**−0.031**	−0.057	**−0.031**	−0.100
Spacecraft telemetry	1.00	−0.200	−0.137	**−0.039**	−0.073	−0.091

The largest gains were on PVC detection. PVCs are morphologically diverse within the class and share a strong confounder, normal QRS complexes that can look briefly like ectopic beats. Kinetic proofreading suits this well: a fleeting resemblance lifts the first cascade state but rarely lasts long enough to reach the terminal one, whereas a genuine PVC holds its distinctive shape over several token positions and builds proofread activation steadily. The smaller but steady gains on NSR and AF probably come from the same noise rejection in noisy one-versus-rest settings.

The study has several limitations. First, ISN carries engineering parts beyond the immune core, namely multiscale tokenisation, adaptive token pooling and a residual CNN branch, which help in the signal domain and interact with the biological modules; the ablations show the coarse picture. The synthetic depth sweep shows a shallow optimum near $K = 3$, but receptor count R, lattice size $H \times W$ and spatial diffusion still need finer sweeps. Second, we evaluated on held-out splits of a single curated database rather than on independent patient cohorts, and external validation across recording conditions, device types and demographics would be required before any clinical use. Third, the PPG and EMG runs are exploratory transfer checks, not full benchmarks, and their large standard deviations (PPG F1 s.d. above 0.08) reflect small samples. Finally, the extra hyperparameters K, R and $H \times W$ may have task-dependent optima.

Several directions follow. One is to ask whether the synaptic lattice learns any interpretable spatial structure, for instance whether cSMAC-like concentration tracks the presence of arrhythmic events. Another is to extend kinetic proofreading to cross-reactive receptors. A third, and the most important for clinical relevance, is prospective, patient-held-out validation. The apnea and spacecraft results already push the evidence past short arrhythmia bursts, to slower physiological signals and to engineering telemetry, and across all of these the property that matters is raw temporal persistence, not the domain itself. Network intrusion and machine sound mark the other side of that line: where the discriminative signal has been reduced to per-sample statistics, the ISN prior adds nothing,

and whether a learnable pre-processing stage could rebuild persistence structure in such data is open. Ultimately, immune-inspired persistence filtering is not a universal upgrade to sequence models, but it is a well-grounded one for the specific and common case in which what makes an anomaly anomalous is that it does not go away.

Methods

Immunological Synapse Network architecture

Antigen presentation: multiscale tokenisation. Each input is a one-dimensional waveform $X = (x_1, \ldots, x_T)^\top \in \mathbb{R}^{T\times 1}$. ISN turns this window into a shorter sequence of learned feature vectors, or tokens, with three parallel temporal convolution branches of different kernel widths. Writing $\mathcal{C}_{k_j}$ for the branch with kernel width k_j and $H_j = \mathcal{C}_{k_j}(X) \in \mathbb{R}^{N_t\times d_b}$ for its output, the branches share a stride, so their N_t output positions line up in time. The tokens are

$$Z = \mathrm{Proj}(H_1 \parallel H_2 \parallel H_3) \in \mathbb{R}^{N_t\times d}, \qquad H_j = \mathcal{C}_{k_j}(X) \in \mathbb{R}^{N_t\times d_b}. \tag{1}$$

where $\parallel$ is concatenation along the feature axis and Proj maps the $3d_b$ features at each position to a d-dimensional token. Each row $z_t \in \mathbb{R}^d$ of Z therefore captures the local waveform around one position at three scales: short, intermediate and broad. This front end is a practical signal-processing stage, not a biological analogue; in the immune picture, each z_t is a short signal fragment presented to the receptor repertoire. When the token count exceeds $N_{\max} = 256$, adaptive average pooling shortens the sequence to $N_{\max}$ tokens before matching.

TCR scanning: receptor binding. A T cell carries on the order of 10^4 T-cell receptors (TCRs), each of fixed specificity, and detects a foreign pMHC when enough TCRs with matching complementarity-determining regions engage it. ISN represents this repertoire as R learned prototype vectors $\{R_r\} \subset \mathbb{R}^d$. For token z_t and receptor r, a scaled dot product gives the binding-energy logit

$$e_{t,r} = \gamma\,\frac{z_t^\top R_r}{\sqrt{d}} + \beta, \qquad p_{t,r}^{\mathrm{bind}} = \sigma(e_{t,r}), \tag{2}$$

by analogy with the free-energy difference that sets TCR–pMHC dwell time. A large positive $e_{t,r}$ means high affinity (long dwell time, small off-rate) and a small or negative value means low affinity (rapid dissociation). The direct binding probability $p_{t,r}^{\mathrm{bind}}$ is the instantaneous occupancy before proofreading: the fast local response that the persistence filter later acts on.

Kinetic proofreading: the persistence filter. The cascade ISN leans on most is the kinetic proofreading mechanism proposed by McKeithan [9]. Once a TCR binds pMHC, productive signalling needs the complex to survive K sequential phosphorylation steps, each catalysed by an associated kinase. Every step requires the complex to stay intact for another characteristic dwell time, so the chance of completing the cascade scales as τ^K for mean dwell time τ. Two ligands that are hard to tell apart at $K=0$, a foreign peptide with dwell time τ_s and a self-peptide with $\tau_w < \tau_s$, separate with a discrimination ratio that grows as $(\tau_s/\tau_w)^K$: K sets the exponent and the binding-energy contrast sets the base. ISN turns this into a differentiable state cascade. For receptor r over N_t tokens, a $K+1$-state occupancy vector $s_{k,r} \in [0,1]$ holds the binding mass at each phosphorylation level. The effective binding probability $p^{\text{eff}}_{t,r} = \sigma(e_{t,r} - \theta_r)$ is thresholded by a per-receptor self-discrimination parameter θ_r, and the off-rate $k^{\text{off}}_{t,r} = (1 - p^{\text{eff}}_{t,r})\,\lambda_r$ falls as binding energy rises:

$$p^{\text{eff}}_{t,r} = \sigma(e_{t,r} - \theta_r), \qquad k^{\text{off}}_{t,r} = (1 - p^{\text{eff}}_{t,r})\lambda_r. \tag{3}$$

$\lambda_r = \min(\exp(\hat{\lambda}_r), 10)$ is a learned positive drain scale for receptor r, the fastest rate at which weakly bound occupancy can fall back to the entry state, and $\Delta t \in (0, 0.5)$ is a learned Euler step used to integrate the cascade once per token. Equation 3 ties forward drive to reset pressure: high-affinity tokens enter more strongly and drain more slowly. At each step, new mass enters the first state from the unbound pool at rate $p^{\text{eff}}\,\Delta t$, and mass in state k either moves to state $k+1$ with probability $\min(k_p^{(k)}\,\Delta t, 0.45)$ or falls back to the entry state with probability $\min(k^{\text{off}}\,\Delta t, 0.45)$. Here $k_p^{(k)} = \min(\exp(\hat{k}_p^{(k)}), 10)$ is the learned forward rate at level k. Only mass that clears all K steps reaches the proofread output $u^{\text{KPR}}_{t,r} = s^{(t)}_{K,r}$. Per token,

$$\begin{aligned}
s^{(t)}_{0,r} \mathrel{+}= &\left(1 - \sum_{j=0}^{K} s^{(t)}_{j,r}\right)^{+} p^{\text{eff}}_{t,r}\Delta t, \\
\Delta^{\text{fwd}}_{k,r} = s^{(t)}_{k,r}\min(k_p^{(k)}\Delta t, 0.45), &\qquad \Delta^{\text{unb}}_{k,r} = s^{(t)}_{k,r}\min(k^{\text{off}}_{t,r}\Delta t, 0.45), \\
s^{(t)}_{k,r} \leftarrow \left(s^{(t)}_{k,r} - \Delta^{\text{fwd}}_{k,r} - \Delta^{\text{unb}}_{k,r}\right)^{+}, &\qquad s^{(t)}_{k+1,r} \mathrel{+}= \Delta^{\text{fwd}}_{k,r}, \quad s^{(t)}_{0,r} \mathrel{+}= \Delta^{\text{unb}}_{k,r},
\end{aligned} \tag{4}$$

for $k = 0, \ldots, K-1$, with the terminal proofread signal

$$u^{\text{KPR}}_{t,r} = s^{(t)}_{K,r}. \tag{5}$$

Persistently high-affinity tokens push mass all the way through; transient or low-affinity tokens bleed out first. For a stream with N_s strong tokens (ε_s) and N_w weak tokens (ε_w), the discrimination ratio between anomalous and normal windows after K steps is $D(K) = D_0\,\rho^K$, with baseline affinity ratio $D_0 = k^{\text{off}}_w / k^{\text{off}}_s$ and per-step amplification $\rho = (k^{\text{off}}_w + k_p)/(k^{\text{off}}_s + k_p) > 1$ set by the binding-energy contrast $\varepsilon_s - \varepsilon_w$. A learned scalar

gate $\alpha = \sigma(a)$ mixes the slow proofread signal with the fast direct-binding estimate, so the network can set its own speed-versus-specificity trade-off for the noise level of each task:

$$u_{t,r} = \alpha\, u_{t,r}^{\mathrm{KPR}} + (1-\alpha)\, p_{t,r}^{\mathrm{bind}}, \qquad \alpha = \sigma(a). \tag{6}$$

Near $\alpha = 1$ the model insists that evidence survive the cascade; near $\alpha = 0$ it acts like a plain direct-affinity detector.

ITAM gating: the activation threshold. In T cells, TCR phosphorylation does not trigger downstream signalling on its own. Immunoreceptor tyrosine-based activation motifs (ITAMs) in the cytoplasmic tails of the CD3 chains build up phosphorylations and recruit ZAP-70 only once enough ITAMs are doubly phosphorylated, a cooperative threshold that blocks weak or stochastic events from propagating. ISN copies this with a per-receptor gate $\tilde{u}_{t,r} = \beta^{-1}\ln(1 + e^{\beta(u_{t,r}-\tau_r)})$, a smooth ReLU at τ_r with $\beta{=}5$ and learned threshold $\tau_r = \sigma(\theta_r^{\mathrm{ITAM}})$:

$$\tilde{u}_{t,r} = \frac{1}{\beta}\log(1 + \exp\{\beta(u_{t,r} - \tau_r)\}), \qquad \tau_r = \sigma(\theta_r^{\mathrm{ITAM}}), \quad \beta = 5. \tag{7}$$

$\theta_r^{\mathrm{ITAM}} \in \mathbb{R}$ is an unconstrained learned logit, and the sigmoid maps it to a usable threshold $\tau_r \in (0,1)$, the minimum blended evidence receptor r needs before it weighs on the readout. Receptors above τ_r contribute; those below are silenced. Equation 7 keeps this thresholding differentiable, so τ_r can be learned alongside the receptor prototypes and the proofreading rates.

Immunological synapse: spatial organisation and readout. At the immunological synapse, engaged TCRs and their signalling complexes sort into a bull's-eye: a central supramolecular activation cluster (cSMAC) rich in engaged TCRs, ringed by a peripheral zone (pSMAC) of adhesion molecules. This layout gathers evidence from many separate engagements into one amplified signal before the cell commits. ISN mirrors it with a learned $H{\times}W$ lattice, built in four steps. The token z_t is mapped to a normalised position $q_t \in [-1,1]^2$; this contracts toward the centre by a factor $1-c$ with $c = \sigma(\phi)\bar{u}$ rising with the mean gated activation $\bar{u}$; the contracted position $\hat{x}_t$ is rescaled to a lattice coordinate x_t; and each token spreads its gated activity $\tilde{u}_{t,r}$ over nearby cells with a softmax over

squared distances:

$$
\begin{aligned}
q_t &= \tanh(W_{\mathrm{xy}} z_t + b_{\mathrm{xy}}) \in [-1,1]^2, \qquad c = \sigma(\phi)\bar{u}, \\
\hat{x}_t &= (1-c) q_t, \\
x_t &= \left(\frac{H-1}{2}(\hat{x}_{t,2}+1),\ \frac{W-1}{2}(\hat{x}_{t,1}+1) \right), \\
P_{t,h,w} &= \frac{\exp[-\|(h,w) - x_t\|^2]}{\sum_{h',w'} \exp[-\|(h',w') - x_t\|^2]}, \qquad M_{r,h,w} = \sum_t P_{t,h,w} \tilde{u}_{t,r}.
\end{aligned} \tag{8}
$$

The biological part is the activation-dependent pull toward the centre, not the learned coordinate projection. In a real synapse, persistently engaged complexes accumulate in the cSMAC; Equation 8 does the same, in that a strong overall response $\bar{u}$ raises c and draws tokens inward so their evidence piles up centrally, while a weak response leaves c small and the evidence spread out. The projection, rescaling and distance softmax are differentiable implementation choices. The lattice is then refined by N_s iterations of a learned 3×3 convolution D, each applying $M \leftarrow \mathrm{ReLU}(M + \Delta t_s (D * M - M))$ followed by per-cell normalisation over receptor types, $M_{r,h,w} \leftarrow M_{r,h,w} / (\sum_{r'} M_{r',h,w} + 10^{-6})$. This refinement is an engineering step, not a literal diffusion model. The refined tensor $M \in \mathbb{R}^{R\times H\times W}$ is reduced to a receptor histogram $h_R \in \mathbb{R}^R$ by averaging each receptor channel over the lattice, so $(h_R)_r$ is the mean occupancy of receptor r; this keeps the receptor composition but discards where on the lattice each response sat. In parallel, $T(\tilde{u}) \in \mathbb{R}^{4R}$ collects the mean, maximum, standard deviation and final activation of each receptor along the token stream:

$$
\begin{aligned}
h_R &= \frac{1}{HW} \sum_{h,w} M_{:,h,w}, \\
T(\tilde{u}) &= \left[\frac{1}{N_t} \sum_t \tilde{u}_t,\ \max_t \tilde{u}_t,\ \mathrm{std}_t(\tilde{u}_t),\ \tilde{u}_{N_t} \right], \\
f_{\mathrm{core}} &= \mathrm{MLP}([h_R, T(\tilde{u})]) \,.
\end{aligned} \tag{9}
$$

Residual branch and classification. The multiscale tokeniser of Eq. 1 uses three parallel branches, each a strided convolution followed by a width-3 convolution with GELU activations; the per-dataset kernel widths and stride are listed in Table 10, and a squeeze-and-excitation channel-attention gate is applied first on multichannel inputs. The readout MLP of Eq. 9 maps the $R+4R$ histogram-plus-temporal features through layer normalisation, a linear $5R\to d$ layer, GELU, dropout 0.1 and a linear $d\to d$ layer to give $f_{\mathrm{core}} \in \mathbb{R}^d$. In parallel, ISN computes a feature $r(X)$ from a residual convolutional branch [7]: three convolutions (kernel widths 15, 7 and 3; hidden width $\max(16, d/2)$) over the same input, reduced to $r(X) \in \mathbb{R}^{2d}$ by concatenating global average- and max-pools followed by layer normalisation. The core feature f_{core} and $r(X)$ are concatenated into a

Table 10: **Datasets and per-dataset ISN configuration.** Window is the input length in samples (or feature frames); C is the number of input channels. Tokeniser columns give the three multiscale kernel widths and the stride. Splits are stratified by class except the wearable physiological sets, which use subject-grouped splits (no subject shared across splits).

Task	Source	Window $\times C$	Tok. kernels	Stride	Split
ECG arrhythmia	MIT-BIH [13]	200×1 (360 Hz)	7/15/31	2	stratified
PPG	PPG-DaLiA [14]	512×1 (64 Hz)	7/31/101	4	subject
EMG	NinaPro [15]	$512 \times C$	5/15/31	2	subject
Apnea	Apnea-ECG [16]	3000×1 (100 Hz)	31/101/301	8	subject, 5-fold
Network intrusion	CICIDS2017 [17]	64×77	7/31/101	4	stratified
Machine sound	MIMII/DCASE [18]	256×5	7/31/101	4	stratified
Spacecraft	SMAP/MSL [19]	$128 \times C$	7/31/101	4	stratified

$3d$-dimensional vector and passed to the classification head (layer normalisation; linear $3d \to d$; GELU; dropout 0.1; linear $d \to 2$); the residual-branch ablation drops $r(X)$ and feeds f_{core} alone. With the default configuration ($d = 64$, $R = 16$, $H = W = 12$, $K = 3$) and single-channel input, ISN has about 7.6×10^4 parameters (tokeniser 2.2×10^4, immune core 1.3×10^4, residual branch 2.8×10^4, head 1.3×10^4).

Datasets and preprocessing

All real-world benchmarks use publicly available data; sources and the per-dataset input and tokeniser settings are summarised in Table 10. The ECG tasks are four one-versus-rest binary problems (NSR, AF, VT, PVC) built from MIT-BIH [12, 13] beat windows of 200 samples (90 before and 110 after each annotated R-peak). PPG uses the wrist photoplethysmography channel of PPG-DaLiA [14]; EMG uses surface electromyography from NinaPro [15]; sleep-apnea detection uses single-lead ECG from the Apnea-ECG Database [16] in 3000-sample (30 s) windows. Network intrusion uses CICIDS2017 [17], where each window is 64 steps of 77 pre-computed flow statistics (176,893 windows from eight day-long captures, 36.3% attack rate); machine-sound anomaly detection uses frame-level features (256 frames of five descriptors) from the MIMII/DCASE benchmark [18]; and spacecraft telemetry uses the NASA SMAP/MSL benchmark [19] (82 entity groups, $\approx$191 windows per group, 128-step sequences). Inputs were standardised by a per-window z-score, and the minority class in the training split was balanced by random oversampling; validation and test splits were left at their natural class ratio.

Baselines

ISN was compared with four neural baselines, each consuming the standardised input window directly; only ISN uses the multiscale tokeniser. The CNN is a residual convolutional network with three convolutions (32, 64 and 128 channels, GELU activations, two inter-

vening max-pools), global average pooling and a linear layer. The GRU [4] and LSTM [3] are single-layer recurrent networks (hidden size 64) that classify the final hidden state and adaptively pool sequences longer than 256 steps. The Transformer [5] is a two-layer pre-norm encoder (model width 32, two heads, feed-forward width 96, GELU, dropout 0.1) with mean pooling over time. For single-channel input these have roughly 3.6×10^4 (CNN), 1.3×10^4 (GRU), 1.7×10^4 (LSTM) and 2.1×10^4 (Transformer) parameters. Two classical baselines reported in the noise-robustness table, namely a support vector machine (RBF kernel on standardised inputs) and a random forest (200 trees), operate on the flattened window.

Training and evaluation

All models were trained by minimising softmax cross-entropy with Adam [20] (learning rate 10^{-3}, default β_1, β_2 and ϵ, no weight decay) at batch size 256 for up to 30 epochs, with early stopping on validation F1 (patience 10); the epoch with the best validation F1 was restored. Data were partitioned 70/15/15 into train/validation/test, stratified by class, except for the wearable physiological sets (PPG, EMG, apnea), which used subject-grouped splits so that no subject appeared in more than one split, with apnea additionally evaluated by five-fold subject-grouped cross-validation. Main benchmarks were repeated over five random seeds and the synthetic, impulse-noise and Gaussian-noise sweeps over three; we report mean $\pm$ standard deviation. F1 used a fixed decision threshold of 0.5 on the positive-class softmax probability, while AUROC and AUPRC are threshold-free. Unless noted, ISN used $K = 3$ proofreading steps, token dimension $d = 64$, $R = 16$ receptors, a 12×12 lattice, four binding micro-steps ($\Delta t_{\text{bind}} = 0.25$) and at most $N_{\max} = 256$ tokens; the number of synaptic-diffusion iterations N_s ($\Delta t_{\text{syn}} = 0.15$) was 6 for the ECG and synthetic tasks and 1 for the other benchmarks, and the per-dataset tokeniser kernels are listed in Table 10. Gaussian-noise robustness added noise to the training split only, after standardisation, with $\sigma = 1.0$ equal to the standard deviation of the normalised input; impulse robustness added bipolar shifts of amplitude 5.0 to a fraction p of training points. Models were implemented in PyTorch and trained on a single GPU.

Synthetic weak-evidence task

The synthetic benchmark is built so that the label can be recovered only by integrating weak evidence over time. Its construction follows three rules: a large transient event is not an anomaly, an isolated weak event is not an anomaly, and a weak motif that repeats with valid spacing across several consecutive steps is the anomaly. The intent is to defeat instantaneous detectors, such as maximum amplitude, maximum absolute value, threshold crossings and spike counts, while rewarding a model that accumulates evidence, as kinetic proofreading does.

Background. Each window is a length-200 univariate sequence. The background $B(t)$ is the sum of four components:

$$\begin{aligned} b_1(t) &= A_1 \sin\!\left(\tfrac{2\pi\cdot 3\,t}{T} + \phi\right), \quad A_1 \sim \mathcal{U}[0.8, 1.2],\ \phi \sim \mathcal{U}[0, 2\pi], \\ b_2(t) &= A_2 \sin\!\left(\tfrac{2\pi\cdot 6\,t}{T} + \phi\right), \quad A_2 \sim \mathcal{U}[0.05, 0.18], \\ \eta(t) &\sim \mathcal{N}(0,\ 0.07^2), \\ \mathrm{AR}(t) &= 0.92 \cdot \mathrm{AR}(t-1) + \varepsilon(t), \quad \varepsilon(t) \sim \mathcal{N}\!\left(0,\ 0.055^2(1 - 0.92^2)\right), \end{aligned} \tag{10}$$

so $B(t) = b_1(t) + b_2(t) + \eta(t) + \mathrm{AR}(t)$. The AR(1) component adds slow temporal drift that a hand-tuned rolling-average baseline cannot easily remove.

Transient distractors. Every window in both classes receives $N_d \sim \mathcal{U}_{\mathbb{Z}}[2, 4]$ narrow Gaussian spikes placed at random positions μ_i:

$$D_i(t) = s_i A_i \exp\!\left(-\frac{(t-\mu_i)^2}{2w_i^2}\right), \quad A_i \sim \mathcal{U}[6, 12],\ w_i \sim \mathcal{U}[1, 2],\ s_i \in \{+1, -1\}. \tag{11}$$

Because A_i is five to fifteen times larger than the motif amplitude and the process is class-independent, any score based on $\max_t |x(t)|$ or spike counts has zero discriminative power.

Weak motif shape. Each motif is a biphasic pulse centred at μ with width w and amplitude A:

$$m(t; \mu, w, A) = A\left[\exp\!\left(-\frac{(t-\mu)^2}{2w^2}\right) - 0.35 \exp\!\left(-\frac{(t-\mu-\Delta)^2}{2w^2}\right)\right], \tag{12}$$

where $w \sim \mathcal{U}_{\mathbb{Z}}[5, 7]$, $A \sim \mathcal{U}[0.8, 1.1]$ and Δ is a fixed offset (a small fraction of w) that produces the shoulder. The biphasic shape prevents reduction to a single scalar peak: a detector integrating over a window wide enough to capture both lobes becomes sensitive to the waveform phase, which is exactly what a learned receptor prototype is designed to match. Polarity is held constant within each window so that one receptor accumulates proofreading state consistently across successive motifs rather than cancelling across opposite-sign events.

Class-discriminating timing. Both classes draw their 12 motifs from the distribution of Eq. 12 with the same budget; they differ only in how the motifs are placed in time.

Positive: all 12 motifs form one uninterrupted run,

$$\mu^+_{i+1} = \mu^+_i + g^+_i, \quad g^+_i \sim \mathcal{U}_{\mathbb{Z}}[5, 7] \pm 1, \tag{13}$$

starting at a random offset. The ± 1 jitter prevents detection via exact periodicity.

Negative: the 12 motifs are split into six pairs. Within each pair the intra-cluster gap is the same as in the positive run, $g_{\text{intra}} \sim \mathcal{U}_{\mathbb{Z}}[5, 7]$, but consecutive pairs are separated by a reset gap $g_{\text{reset}} \sim \mathcal{U}_{\mathbb{Z}}[9, 12]$. Because $g_{\text{reset}} > g_{\text{intra}}$, the KPR cascade drains back to state 0 between clusters. The connection to the cascade dynamics is direct: during a reset gap $p_{t,r}^{\text{eff}} \approx 0$, so $k_{t,r}^{\text{off}} = (1 - p_{t,r}^{\text{eff}})\lambda_r \approx \lambda_r$ and each update step $\Delta_{k,r}^{\text{unb}} = s_{k,r} \min(k_{t,r}^{\text{off}} \Delta t, 0.45)$ empties states $k > 0$ back to s_0 before the next cluster arrives, so $s_{K,r}$ never accumulates. In the positive run there is no reset gap, so forward increments $\Delta_{k,r}^{\text{fwd}}$ push mass all the way to the terminal state $s_{K,r}$. This is the discrimination ratio argument from the architecture section: the ratio $D(K) = D_0 \rho^K$ grows with K and with the binding-energy contrast between sustained and interrupted token streams.

Decoys. To prevent a model from keying on the mere count or presence of short valid-spaced runs, every window (both classes) also carries $N_{\text{dec}} \sim \mathcal{U}_{\mathbb{Z}}[1, 2]$ decoy trains, each a burst of $L \sim \mathcal{U}_{\mathbb{Z}}[3, 5]$ motifs at the same $g \sim \mathcal{U}_{\mathbb{Z}}[5, 7]$ spacing and amplitude $A_{\text{dec}} \in [0.85, 1.15] \cdot A_{\text{motif}}$. With decoys present, instantaneous amplitude, motif count, integrated motif energy and the existence of any short valid-spaced run are all non-discriminative; the length of the single longest uninterrupted run is the only separating feature.

Full window and dataset. The complete window is

$$x(t) = \xi \left[B(t) + \sum_{i=1}^{N_d} D_i(t) + \sum_{j=1}^{12} m(t; \mu_j, w_j, A_j) + \sum_{k,l} m(t; \mu_{k,l}^{\text{dec}}, w_{k,l}^{\text{dec}}, A_{k,l}^{\text{dec}}) \right], \quad \xi \sim \mathcal{U}[0.9, 1.1]. \tag{14}$$

The global scale jitter ξ is applied instead of per-window amplitude-normalisation, which would shrink the distractors and weaken the test. Classes are balanced (50% positive), and the data comprise 8000 training, 2000 validation and 2000 test windows, each split sampled independently from the generator.

As a sanity check, the generator verifies that intentionally weak detectors (maximum absolute value, spike count) score at or below chance, that a persistence proxy (sinusoid-removed, spike-clipped rolling average of the residual) scores above chance, and that an event-timing oracle (longest run of motifs spaced within 5–7 samples) is strongly predictive, confirming that timing and persistence, not amplitude, carry the label. ISN and the baselines were trained on this task with the shared protocol for 20 epochs (patience 6) over three seeds (Table 1).

Data availability

All real-world datasets used here are publicly available from their original providers. The synthetic weak-evidence dataset is produced by scripts in the accompanying code repository.

Code availability

The ISN implementation is available at `https://github.com/sheriftawfikabbas/isn`.

Author contributions

S.A.T. conceived the architecture, implemented the model and benchmark runners, conducted all experiments and wrote the manuscript. H.L. contributed to experimental design and analysis and wrote the manuscript. S.V. supervised the research and contributed and wrote the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

[1] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25:65–69, 2019. doi: `10.1038/s41591-018-0268-3`.

[2] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, 2019. doi: `10.1007/s10618-019-00619-1`.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: `10.1162/neco.1997.9.8.1735`.

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014. doi: `10.3115/v1/D14-1179`.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. doi: 10.48550/arXiv.1706.03762.

[6] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[8] Jiapu Pan and Willis J. Tompkins. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, 32(3):230–236, 1985. doi: 10.1109/TBME.1985.325532.

[9] Timothy W. McKeithan. Kinetic proofreading in T-cell receptor signal transduction. *Proceedings of the National Academy of Sciences*, 92(11):5042–5046, 1995. doi: 10.1073/pnas.92.11.5042.

[10] Arup K. Grakoui, Shannon K. Bromley, Cenk Sumen, Mark M. Davis, Andrey S. Shaw, Paul M. Allen, and Michael L. Dustin. The immunological synapse: a molecular machine controlling T cell activation. *Science*, 285(5425):221–227, 1999. doi: 10.1126/science.285.5425.221.

[11] Michael L. Dustin and John A. Cooper. The immunological synapse and the actin cytoskeleton: molecular hardware for T cell signaling. *Nature Immunology*, 1:23–29, 2000. doi: 10.1038/76877.

[12] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

[13] George B. Moody and Roger G. Mark. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001. doi: 10.1109/51.932724.

[14] Attila Reiss, Ina Indlekofer, and Philip Schmidt. PPG-DaLiA, 2019. doi: 10.24432/C53890.

[15] Manfredo Atzori et al. Electromyography data for non-invasive naturally-controlled robotic hand prostheses. *Scientific Data*, 1:140053, 2014. doi: 10.1038/sdata.2014.53.

[16] Thomas Penzel, George B. Moody, Roger G. Mark, Ary L. Goldberger, and Jörg H. Peter. The apnea-ECG database. *Computers in Cardiology*, 27:255–258, 2000. doi: 10.1109/CIC.2000.898505.

[17] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, pages 108–116, 2018. doi: 10.5220/0006639801080116.

[18] Harsh Purohit, Ryo Tanabe, Kenji Ichige, Takashi Endo, Yuki Nikaido, Kaori Suefusa, and Yohei Kawaguchi. MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, 2019. doi: 10.33682/m76f-d618.

[19] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 387–395. ACM, 2018. doi: 10.1145/3219819.3219845.

[20] Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015. doi: 10.48550/arXiv.1412.6980.