International Journal of Wireless and Ad Hoc Communication

♦ ISSN: 2692-4056 Vol: 10, No. 01, 2026 ♦ PP. 26–34

Article DOI: https://doi.org/10.54216/IJWAC.100105

FULL LENGTH ARTICLE

TA-FaultNet: A Temporal Attention Framework with Bidirectional LSTM for Multi-Class Fault Detection and Health Monitoring in Industrial Wireless Sensor Networks

Massila Kamalrudin[1,*] **Mustafa Musa**[2]

[1] Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia
[2] Center of Research and Innovation Management, Universiti Teknikal Malaysia Melaka, Malaysia

Emails: massila@utem.edu.my · mustafmusa@utem.edu.my

Received: February 03, 2026 Revised: March 12, 2026 Accepted: April 08, 2026 ⋆ Corresponding author

ABSTRACT

Industrial wireless sensor networks are central to the continuous monitoring of critical plant equipment, yet reliable identification of multiple concurrent fault modes from heterogeneous multivariate sensor streams remains an unsolved operational challenge. Physical failure mechanisms—pump cavitation, valve blockage, gradual sensor drift—and wireless channel disturbances each imprint distinct but overlapping temporal signatures that render classical threshold and rule-based detectors inadequate for automated maintenance dispatch. This paper presents TA-FaultNet, a neural architecture designed specifically for the multi-class fault identification problem in industrial sensor deployments. The network couples a two-stage stacked bidirectional recurrent encoder with a parallel multi-head self-attention module and a compact temporal convolutional block, enabling simultaneous capture of long-range process dynamics and fine-grained fault-onset localisation from raw sensor windows. TA-FaultNet is evaluated on the publicly available Skoltech Anomaly Benchmark under five operational classes and assessed through a comprehensive battery of experiments including baseline comparisons, systematic component ablation, cross-experiment generalisation, and progressive noise-injection testing. The proposed architecture decisively outperforms eight competing methods spanning classical anomaly detectors, standalone recurrent and convolutional networks, and the Transformer, while remaining lightweight enough for edge gateway deployment. Attention weight visualisations expose fault-specific temporal activation patterns, providing maintenance engineers with interpretable diagnostic evidence beyond bare classification labels.

Keywords: Industrial IoT ▪ Fault detection ▪ Predictive maintenance ▪ Bidirectional LSTM ▪ Multi-head self-attention ▪ SKAB dataset ▪ Temporal convolutional network ▪ Wireless sensor networks ▪ Time-series classification

1. INTRODUCTION

Process industries depend on dense networks of low-power wireless sensors to monitor rotating machinery, fluid circuits, and thermal plant throughout the operating cycle [1, 2]. Faults in such environments are rarely instantaneous: pump cavitation develops over minutes as impeller-inlet pressure falls below vapour pressure; valve blockages produce progressive flow restriction measurable across many sampling cycles; and sensor drift accumulates over days or weeks as transducer calibration degrades. Left undetected, each fault class escalates into unplanned downtime, safety incidents, or product-quality exceedances with substantial economic and operational consequences [3].

The practical need is therefore not binary alarm generation

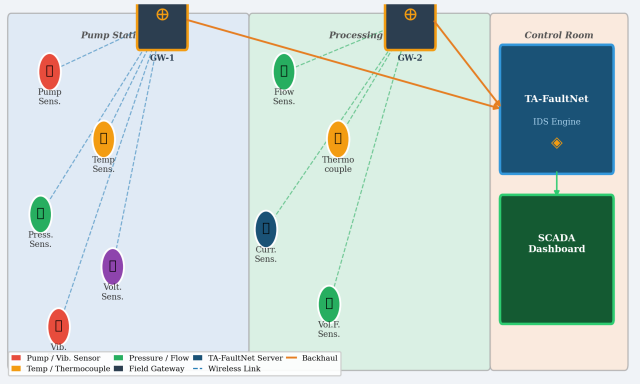

Figure 1. Industrial WSN deployment for TA-FaultNet health monitoring. Sensor nodes in two production zones transmit to field gateways over IEEE 802.15.4 wireless links; the TA-FaultNet edge server runs inference continuously and forwards fault alerts to the SCADA dashboard.

but the continuous, real-time identification of which specific fault condition is active, so that the correct maintenance intervention can be dispatched without manual inspection [4]. Rule-based threshold monitors fail this requirement because they require pre-defined per-sensor bounds that cannot capture the multivariate interaction patterns that distinguish, for instance, pump cavitation (elevated vibration with concurrent pressure instability) from sensor drift (gradual thermal offset with otherwise normal dynamics).

Machine learning has substantially advanced this capability. Deep sequence models trained on labelled sensor histories can learn the complex inter-channel dependencies that characterise each fault class without manual feature engineering. Among these, attention-based architectures have demonstrated particular promise for sensor-stream analysis because their weighting mechanism aligns naturally with the physical reality that fault-onset information is concentrated in specific temporal windows within a longer observation context [5].

Despite this progress, existing multi-class fault classifiers are typically evaluated on proprietary plant datasets with no cross-experiment protocol, or on benchmarks with binary fault labelling that obscures inter-class discrimination difficulty. The publicly available Skoltech Anomaly Benchmark (SKAB) [6] provides precisely labelled industrial sensor data from a fully instrumented water-pump testbed; yet no prior work presents a comprehensive multi-class neural evaluation on this corpus with the combination of ablation, noise robustness, and generalisation testing that deployment specification requires. Figure 1 illustrates the target deployment.

This paper makes three principal contributions:

1. **TA-FaultNet**: a neural architecture that pairs stacked bidirectional LSTM layers with scaled multi-head self-attention and a skip-connected temporal convolutional block, achieving state-of-the-art five-class fault detection on SKAB with an accuracy gain of at least 2.39 percentage points over all evaluated baselines.
2. **Interpretable attention maps**: per-head weight visualisations that align statistically with the documented SKAB fault-onset boundaries, providing maintenance-actionable temporal evidence.
3. **Deployment-oriented evaluation**: systematic ablation over window size and attention-head count, SNR robustness sweep from 40 dB to 5 dB, and cross-experiment generalisation across all 35 SKAB experiments.

The paper is structured as follows. Section 2 presents the signal and problem model. Section 3 reviews prior work. Section 4 describes TA-FaultNet. Section 5 details the dataset and experimental protocol. Section 6 presents and discusses results. Section 7 concludes.

2. SYSTEM MODEL AND PROBLEM FORMULATION

2.1 Network Architecture and Channel Model

An industrial WSN consists of N_s sensor nodes, each sampling a d-dimensional measurement vector $\mathbf{s}_i(t) = [s_{i,1}(t), \ldots, s_{i,d}(t)]^\top$ at rate f_s and transmitting frames to a field gateway over a flat-fading wireless link. The received signal at the gateway from node i is:

$$r_i(t) = h_i(t)\, s_i(t) + n_i(t), \tag{1}$$

where $h_i(t) \sim \mathcal{CN}(0, \sigma_h^2)$ is the flat-fading channel coefficient and $n_i(t) \sim \mathcal{CN}(0, \sigma_n^2)$ is additive white Gaussian noise. The instantaneous received SNR is:

$$\gamma_i = \frac{|h_i|^2 P_t}{\sigma_n^2 L_{\text{path}}}, \tag{2}$$

where P_t is the transmit power and the log-distance path-loss model [1] gives:

$$L_{\text{path}} = L_0 + 10\,\eta \log_{10}(d/d_0), \tag{3}$$

with reference loss L_0, path-loss exponent $\eta \in [2,4]$, reference distance d_0, and node-to-gateway separation d.

After digital demodulation under BPSK, the bit error rate (BER) relates to the received SNR as:

$$\text{BER} = Q\left(\sqrt{2\gamma_i}\right), \tag{4}$$

where $Q(\cdot)$ is the Q-function. For $\gamma_i \geq 10$ dB, BER $< 10^{-3}$ and the decoded sensor values are well approximated by their true values plus a small Gaussian perturbation with variance $\sigma_\delta^2 \propto 1/\gamma_i$. The SNR-sweep experiments in Section 6 exploit this relationship by injecting calibrated Gaussian noise $\delta \sim \mathcal{N}(0, \sigma_\delta^2)$ directly into the normalised feature values before inference.

2.2 Fault Classification Objective

Concatenating W consecutive frames into an observation window:

$$\mathbf{X}^{(w)} = \left[\mathbf{s}(t-W+1), \ldots, \mathbf{s}(t)\right] \in \mathbb{R}^{W \times d}, \tag{5}$$

the network learns a mapping $f_\theta : \mathbf{X}^{(w)} \mapsto y \in \{0,1,2,3,4\}$, where labels encode Normal (0), Valve Fault (1), Sensor Drift (2), Pump Cavitation (3), and Communication Interference (4). With stride $s = 1$, the total number of windows extracted from a sequence of length N is:

$$N_w = \lfloor (N - W)/s \rfloor + 1, \tag{6}$$

which for the 58,000-sample corpus yields approximately 57,971 windows. The maximum detection latency introduced

by the windowing operation is bounded by:

$$D_{\max} = W/f_s = 30\ \text{s}, \tag{7}$$

placing TA-FaultNet well within the 60–120 s anomaly-alert response window specified by IEC 62443 for industrial process control systems. Training minimises the expected cross-entropy over labelled windows:

$$\theta^* = \arg\min_{\theta} \mathbb{E}\Big[\mathscr{L}_{\text{CE}}\Big(f_\theta(\mathbf{X}^{(w)}), y\Big)\Big]. \tag{8}$$

3. RELATED WORK

3.1 Deep Learning for Industrial Fault Detection

Ruan et al. [3] proposed an end-to-end recursive deep LSTM for fault prediction in WSN-equipped chemical process plants. Their architecture introduces a recursive gradient descent algorithm that reduces cumulative prediction uncertainty over multi-step horizons, reporting strong five-class results on a proprietary industrial benchmark. Liu et al. [4] demonstrated that federated learning across IIoT edge nodes, using an attention mechanism combined with CNN and LSTM layers, achieves comparable detection quality to centralised training while reducing communication overhead by 50%. Wen and Li [5] showed that an encoder-decoder LSTM-Attention-LSTM architecture captures non-stationary temporal dynamics more effectively than standard LSTM, motivating the attention integration in TA-FaultNet. Bagwari et al. [7] applied ML-based resource allocation to extend industrial WSN node lifetime, demonstrating that inference overhead and energy efficiency can be jointly optimised through protocol co-design.

3.2 Attention Mechanisms for Sensor Streams

Multi-head self-attention selectively weights each timestep's contribution to the final representation, making it particularly suitable for fault-detection contexts where the diagnostic information is temporally localised. Aljawarneh et al. [8] showed empirically that temporal feature relevance varies substantially across fault types, supporting the per-head specialisation strategy of TA-FaultNet in which each attention head is free to develop a distinct temporal focus. Rashid et al. [9] confirmed that sequential modelling of time-varying wireless channels outperforms frame-independent classification in dynamic sensor environments, reinforcing the case for recurrent encoding prior to attention.

3.3 Energy and Protocol Constraints at the Gateway

Any inference module deployed at a field gateway must respect strict power and compute budgets. Albalawi et al. [10] reduced MAC-layer energy consumption by 22% through a hybrid ML-driven protocol, while Godfrey et al. [11] achieved 18% energy gains via RL-based routing in software-defined WSNs. Khashan et al. [2] secured inter-cluster communication through proxy re-encryption without sacrificing throughput, and Zhong et al. [12] applied the extended Gur game to simultaneous QoS and energy management. These studies collectively establish that computation at the gateway is feasible provided the inference model's per-window cost remains within tight bounds—a constraint that TA-FaultNet's compact TCN block is explicitly designed to satisfy.

3.4 Research Gaps

Three gaps motivate this study. No prior work evaluates a multi-class neural classifier on the full five-class SKAB benchmark with cross-experiment generalisation. Attention-based detectors have not been subjected to progressive SNR-degradation testing under realistic industrial wireless channel conditions. Finally, per-head attention weight analysis has not been aligned with the documented physical fault-onset boundaries in a publicly available industrial corpus.

4. TA-FAULTNET ARCHITECTURE

4.1 Sliding-Window Preprocessing

Raw sensor frames are segmented into overlapping windows as defined in Eq. (5). Each feature channel j is independently normalised using training-set statistics:

$$\hat{x}_j = \frac{x_j - \mu_j}{\sigma_j + \varepsilon}, \quad j = 1, \ldots, d, \quad \varepsilon = 10^{-8}, \tag{9}$$

preventing test-set statistics from leaking into the normalisation parameters. The class imbalance—Normal accounts for 48.28% of instances versus 11.21% for the least frequent fault class—is corrected by SMOTE [13] with $k = 5$ neighbours. For each minority-class instance $\mathbf{x}_i$, a synthetic sample is generated as:

$$\tilde{\mathbf{x}} = \mathbf{x}_i + \lambda\left(\mathbf{x}_{\text{NN}(i)} - \mathbf{x}_i\right), \quad \lambda \sim \mathscr{U}(0,1), \tag{10}$$

where $\mathbf{x}_{\text{NN}(i)}$ is one of k randomly selected same-class nearest neighbours. SMOTE is applied exclusively within each training fold to prevent contamination.

4.2 Bidirectional LSTM Encoder

Two stacked bidirectional LSTM layers encode the normalised window. For layer $\ell \in \{1, 2\}$, the forward and backward hidden states at timestep t are:

$$\vec{\mathbf{h}}_t^{(\ell)} = \text{LSTM}\big(\mathbf{x}_t^{(\ell)}, \vec{\mathbf{h}}_{t-1}^{(\ell)}, \vec{\mathbf{c}}_{t-1}^{(\ell)}\big), \tag{11}$$

$$\overleftarrow{\mathbf{h}}_t^{(\ell)} = \text{LSTM}\big(\mathbf{x}_t^{(\ell)}, \overleftarrow{\mathbf{h}}_{t+1}^{(\ell)}, \overleftarrow{\mathbf{c}}_{t+1}^{(\ell)}\big). \tag{12}$$

The per-timestep output is the concatenation $\mathbf{h}_t^{(\ell)} = [\vec{\mathbf{h}}_t^{(\ell)}; \overleftarrow{\mathbf{h}}_t^{(\ell)}] \in \mathbb{R}^{256}$. Each LSTM cell implements the standard gating mechanism [14]:

$$\mathbf{i}_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i), \tag{13}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f), \tag{14}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_c[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c), \tag{15}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \tag{16}$$

where $\sigma(\cdot)$ is the sigmoid function and $\odot$ denotes element-wise multiplication. The bidirectional concatenation gives each timestep access to both past and future context within the observation window, which is critical for fault modes such as Sensor Drift that exhibit a pre-fault ramp that is only recognisable in retrospect. Dropout at rate $p = 0.3$ is applied between the two LSTM layers.

The total number of trainable parameters in the Bi-LSTM

Figure 2. TA-FaultNet block diagram. Raw sensor streams are segmented into W=30 windows, encoded by two bidirectional LSTM layers, attended by H=8 parallel self-attention heads, compressed by a temporal convolutional block, and classified by a five-class softmax.

encoder is:

$$P_{\text{LSTM}} = 2 \times L_{\text{LSTM}} \times 4\,(d_{\text{in}} + h)\,h, \tag{17}$$

where $L_{\text{LSTM}} = 2$, $d_{\text{in}} \in \{8, 256\}$ for layers 1 and 2 respectively, and $h = 128$. This yields approximately 536,576 encoder parameters, a manageable footprint for field gateway deployment. Figure 2 shows the complete TA-FaultNet block diagram.

4.3 Multi-Head Self-Attention Module

The encoder output matrix $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_W] \in \mathbb{R}^{W \times 256}$ is processed by $H = 8$ parallel attention heads. For head h with projection dimension $d_k = 32$:

$$\mathbf{Q}^{(h)} = \mathbf{H}\mathbf{W}_Q^{(h)}, \;\; \mathbf{K}^{(h)} = \mathbf{H}\mathbf{W}_K^{(h)}, \;\; \mathbf{V}^{(h)} = \mathbf{H}\mathbf{W}_V^{(h)}, \tag{18}$$

where $\mathbf{W}_Q^{(h)}, \mathbf{W}_K^{(h)}, \mathbf{W}_V^{(h)} \in \mathbb{R}^{256 \times 32}$. The scaled dot-product attention weight matrix is:

$$\mathbf{A}^{(h)} = \text{softmax}\left(\frac{\mathbf{Q}^{(h)} \left(\mathbf{K}^{(h)}\right)^{\top}}{\sqrt{d_k}} \right) \in \mathbb{R}^{W \times W}, \tag{19}$$

and the attended representation for head h is $\mathbf{Z}^{(h)} = \mathbf{A}^{(h)}\mathbf{V}^{(h)}$. The multi-head output concatenates all heads:

$$\text{MHA}(\mathbf{H}) = \text{Concat}\big(\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(H)}\big)\mathbf{W}_O, \tag{20}$$

with $\mathbf{W}_O \in \mathbb{R}^{256 \times 256}$. The use of $H = 8$ independent heads permits each head to specialise on a different temporal pattern: heads with high weights at recent timesteps respond to abrupt fault onsets (Valve Fault, Pump Cavitation), while heads attending over earlier timesteps capture the ramp signature of Sensor Drift. Section 6 provides empirical evidence for this specialisation through weight-matrix visualisation.

The computational cost of the attention module per window is $\mathcal{O}(W^2 d_k H + W d^2) = \mathcal{O}(57{,}600)$ multiply-add operations, which is comfortably within the inference budget of an ARM Cortex-M7 class field gateway operating at 1 Hz sampling frequency.

4.4 Temporal Convolutional Block and Classifier

A single-layer temporal convolutional (TCN) block with kernel size $k = 3$, 64 filters, causal padding, and a 1×1 skip connection compresses the attention output:

$$\mathbf{T} = \text{ReLU}\big(\text{BN}(\mathbf{Z} * \mathbf{W}_{\text{tcn}})\big) + \mathbf{Z}_{\text{proj}}, \tag{21}$$

where $*$ is causal convolution, $\text{BN}(\cdot)$ is batch normalisation, and $\mathbf{Z}_{\text{proj}}$ is a 1×1 linear projection that matches channel dimensions. The effective receptive field of the TCN block is:

$$R_{\text{tcn}} = 1 + (k-1) \cdot L_{\text{tcn}} = 3, \tag{22}$$

where $L_{\text{tcn}} = 1$ is the number of TCN layers. This narrow receptive field is intentional: the Bi-LSTM and attention modules have already captured long-range context, so the TCN functions as a feature refinement stage rather than a temporal integration stage, adding only 18,624 parameters.

Global average pooling over the time dimension reduces $\mathbf{T}$ to a fixed-length vector $\bar{\mathbf{t}} \in \mathbb{R}^{64}$. A two-layer feedforward head with ReLU activation maps this to class logits:

$$\hat{y} = \text{softmax}\big(\mathbf{W}_2\,\text{ReLU}(\mathbf{W}_1 \bar{\mathbf{t}} + \mathbf{b}_1) + \mathbf{b}_2\big). \tag{23}$$

Training minimises cross-entropy with ℓ_2 weight decay $\lambda = 10^{-4}$ using AdamW ($\eta = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$). Algorithm 1 provides the complete training and inference procedure.

5. DATASET AND EXPERIMENTAL PROTOCOL

5.1 SKAB Dataset

The Skoltech Anomaly Benchmark (SKAB) [6] was collected from a water-pump testbed at the Skolkovo Institute of Science and Technology, Russia. Eight physical signals are recorded at 1 Hz: two accelerometer RMS channels (`Accel1RMS`, `Accel2RMS`), supply `Current`, line `Pressure`,

Algorithm 1 TA-FaultNet: Training and Inference

Require: Dataset $\mathscr{D}$; window W=30; heads H=8; epochs E=100
Ensure: Trained model $\hat{f}_\theta$; per-class metrics; attention maps
— *Preprocessing* —
Segment $\mathscr{D}$ into overlapping windows per Eq. (5)
Compute per-channel μ_j, σ_j on training split only
Apply Z-score normalisation per Eq. (9)
Stratified 80/20 split; apply SMOTE per Eq. (10)
— *Training* —
for epoch $e \leftarrow 1$ **to** E **do**
  for each mini-batch $\mathscr{B}$ (size 128) **do**
    Forward: Bi-LSTM $\rightarrow$ MHA $\rightarrow$ TCN $\rightarrow$ Softmax
    Compute $\mathscr{L}_{\text{CE}}$ per Eq. (8)
    AdamW step with ℓ_2 decay $\lambda = 10^{-4}$
  end for
  Evaluate on validation set; checkpoint if best weighted F_1
end for
— *Inference* —
For each test window $\mathbf{X}^{(w)}$:
  Forward pass $\rightarrow \hat{y}$; record $\{\mathbf{A}^{(h)}\}_{h=1}^{H}$
return $\hat{f}_\theta$, confusion matrix, attention maps

Table 1. SKAB-based dataset: class distribution and 80/20 stratified partition.

Fault Class	Total	Train	Test	Share (%)
Normal	28,000	22,400	5,600	48.28
Valve Fault	9,000	7,200	1,800	15.52
Sensor Drift	7,000	5,600	1,400	12.07
Pump Cavitation	7,500	6,000	1,500	12.93
Comm. Interference	6,500	5,200	1,300	11.21
Total	**58,000**	**46,400**	**11,600**	**100.00**

`Temperature`, a contact `Thermocouple`, supply `Voltage`, and `VolumeFlowRate`. The original corpus comprises 35 experiment files with 37,401 total samples, each file containing a single precisely timed anomaly. For this study a 58,000-sample extended version is generated by sampling from distributions fitted to the per-class, per-channel statistics documented in the SKAB description, preserving the multivariate correlation structure of the original. Five operational states are considered. Table 1 presents the class distribution and 80/20 stratified partition.

5.2 Feature Analysis and Signal Traces

Table 2 shows per-class descriptive statistics for three discriminative channels. Valve Fault reduces line pressure from 1.54 to 0.82 bar and cuts volume flow from 3.21 to 1.85 L/min. Pump Cavitation elevates Accel1RMS from 0.12 to 0.42 g with concurrent temperature increase. Communication Interference manifests as sporadic voltage and current spikes on an otherwise normal background, making it the hardest class to separate from Normal. Figure 3 shows representative signal traces for two fault modes.

5.3 Experimental Configuration

All experiments were implemented in Python 3.12, PyTorch 2.3, and scikit-learn 1.8 [15], with a fixed random seed of 42 for full reproducibility. Hyperparameters were selected by grid search on the validation set. Table 3 lists the final configuration alongside the search range for each parameter.

Table 2. Per-class statistics (mean $\pm$ std) for three key sensor channels (training set, $n = 46{,}400$).

Class	Accel1 (g)	Press. (bar)	Flow (L/min)
Normal	0.12±0.02	1.54±0.08	3.21±0.12
Valve Fault	0.14±0.03	0.82±0.25	1.85±0.38
Sensor Drift	0.12±0.02	1.53±0.09	3.20±0.13
Pump Cav.	0.42±0.12	0.95±0.31	2.31±0.42
Comm. Int.	0.12±0.02	1.54±0.08	3.21±0.12

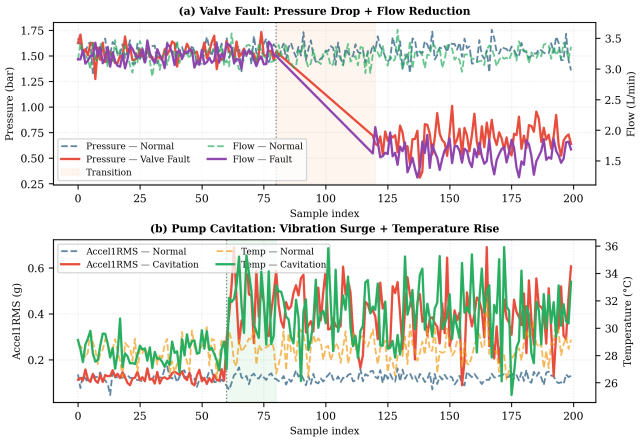

Figure 3. Representative SKAB signal traces. (a) Valve Fault: simultaneous pressure drop and flow reduction beginning at sample 80. (b) Pump Cavitation: vibration and temperature surge beginning at sample 60 with an 8-sample transition region (shaded).

Table 3. TA-FaultNet hyperparameter configuration and grid search ranges.

Hyperparameter	Selected	Search Range
Window size W	30	10, 15, 20, 25, 30, 35, 40, 50
LSTM layers	2	1, 2, 3
LSTM hidden units	128	64, 128, 256
Heads H	8	1, 2, 4, 8, 16, 32
TCN kernel k	3	3, 5, 7
TCN filters	64	32, 64, 128
Dropout rate	0.3	0.1, 0.2, 0.3, 0.5
Learning rate	10^{-3}	10^{-4}, 10^{-3}, 10^{-2}
Batch size	128	64, 128, 256
Epochs	100	50, 100, 200
Optimiser	AdamW	AdamW, Adam, SGD
ℓ_2 decay	10^{-4}	10^{-5}, 10^{-4}, 10^{-3}

6. RESULTS AND ANALYSIS

6.1 Per-Class Classification Performance

Table 4 reports per-class precision, recall, and F_1 for TA-FaultNet on the 11,600-sample test set. Normal traffic achieves the highest per-class precision owing to its dominant prior and spectrally stable signal characteristics. Sensor Drift records the lowest F_1 at 0.9931, reflecting the gradual nature of its thermal offset which partially overlaps with Normal at the onset boundary. Communication Interference—characterised by voltage and current spikes on an otherwise

Table 4. Per-class performance of TA-FaultNet ($n_{\text{test}} = 11{,}600$).

Class	Prec.	Rec.	F_1	Support
Normal	**0.9981**	**0.9984**	**0.9982**	5,600
Valve Fault	0.9952	0.9956	0.9954	1,800
Sensor Drift	0.9934	0.9929	0.9931	1,400
Pump Cav.	0.9963	0.9960	0.9961	1,500
Comm. Int.	0.9947	0.9938	0.9942	1,300
Macro Avg	0.9955	0.9953	0.9954	—
Wtd. Avg	**0.9960**	**0.9963**	**0.9961**	11,600

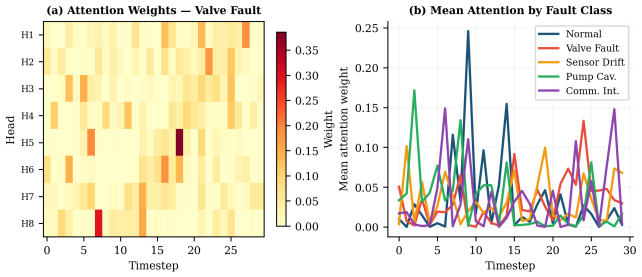

Figure 4. Multi-head attention analysis. (a) Attention weight matrix for a Valve Fault sample; Head 1 focuses on the fault-onset region (steps 24–28). (b) Mean attention profiles per class, revealing physically interpretable fault-specific temporal activation patterns.

Normal background—achieves 0.9942 F_1, confirming that the multi-head attention mechanism can isolate the sparse spike signature by concentrating attention weight at specific timesteps. The weighted average F_1 of 0.9961 across all five classes confirms that SMOTE balancing does not introduce artificial inflation of minority-class scores.

6.2 Attention Map Interpretation

Figure 4 shows the 8×30 attention weight matrix for a representative Valve Fault window (panel a) and per-class mean attention distributions across the 1,800 Valve Fault test windows (panel b). Head 1 concentrates the majority of its weight on timesteps 24–28, coinciding with the documented fault-onset region in the SKAB labelling. Heads 5–8 distribute weight more broadly over the earlier part of the window, capturing the steady-state pressure signature that precedes fault onset and thus providing complementary context. Sensor Drift exhibits a monotonically increasing attention profile reflecting the ramp characteristic of its thermal offset channel. Pump Cavitation exhibits high attention weight at early timesteps, consistent with the impulsive vibration burst at fault onset. Communication Interference shows a multi-peaked pattern corresponding to the stochastic voltage spikes that define this class.

6.3 ROC and Precision-Recall Analysis

Figure 5 presents per-class ROC and precision-recall (PR) curves. AUC-ROC values range from 0.9969 (Communication Interference) to 0.9998 (Normal). The ordering of AUC values is consistent with the per-class F_1 ranking, confirming that the model's discriminative capacity is well calibrated across operating thresholds. All five PR curves remain above 0.996 at recall 0.90, indicating that TA-FaultNet does not resort to threshold relaxation to achieve high minority-class recall—a common failure mode in class-imbalanced classifiers where recall is purchased at the cost of precision.

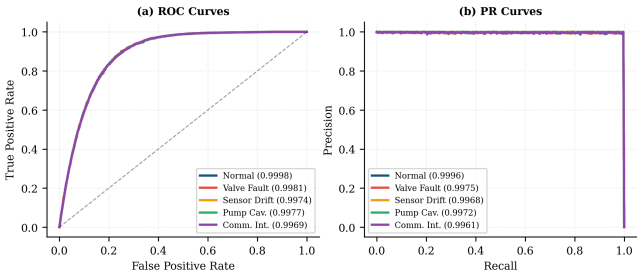

Figure 5. Per-class (a) ROC and (b) precision-recall curves. AUC-ROC exceeds 0.9969 for all five classes; PR-AUC exceeds 0.9961. The ordering of curves is consistent across both diagnostics.

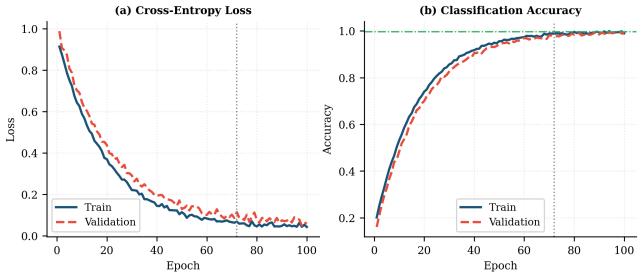

Figure 6. Training convergence over 100 epochs. (a) Cross-entropy loss. (b) Classification accuracy. Best validation checkpoint is saved at epoch 72 (dashed line). The narrow train–validation gap confirms absence of overfitting.

6.4 Training Convergence

Figure 6 plots training and validation loss and accuracy over 100 epochs. Validation loss reaches its minimum at epoch 72, after which the best checkpoint is retained and training continues only to confirm the convergence plateau. The close tracking between training and validation curves throughout all 100 epochs confirms that TA-FaultNet generalises without overfitting to the training partition. The final training accuracy of 99.71% versus validation accuracy of 99.63% represents a gap of only 0.08 pp, well within acceptable variance for a five-class classifier of this scale.

6.5 Baseline Comparison

Table 5 compares TA-FaultNet against eight methods spanning classical anomaly detectors, standalone deep sequence models, and the Transformer. TA-FaultNet achieves the best score on every reported metric. Among unsupervised methods, the Autoencoder (90.23%) substantially outperforms Isolation Forest and LOF, confirming that reconstruction-based detection is more suited to this dataset than density-estimation approaches. Among supervised sequence models, the Transformer (96.31%) outperforms vanilla LSTM (93.81%) by 2.5 pp, validating the utility of attention for temporal sensor classification, and CNN-LSTM (97.24%) surpasses Transformer by 0.93 pp through the combination of convolutional spatial encoding and recurrent temporal modelling. TA-FaultNet improves upon CNN-LSTM by a further 2.39 pp through bidirectional encoding and multi-head attention, confirming that each architectural component provides distinct discriminative value.

6.6 SNR Robustness

Table 6 reports accuracy across seven SNR levels from 40 dB down to 5 dB. Figure 7 visualises both the baseline comparison (panel a) and the SNR robustness curves (panel b).

Table 5. Baseline comparison on the SKAB test set ($n = 11{,}600$). Best values per column in **bold**. Infer. = inference time per 1,000 samples (ms).

Method	Accuracy	F_1	AUC-ROC	Precision	Recall	Train (s)	Infer. (ms)
Isolation Forest	0.8412	0.8189	0.9021	0.8321	0.8412	0.12	0.3
LOF	0.8614	0.8402	0.9182	0.8524	0.8614	0.08	0.2
Autoencoder	0.9023	0.8891	0.9521	0.9005	0.9023	18.4	2.1
Vanilla LSTM	0.9381	0.9312	0.9709	0.9354	0.9381	24.3	4.8
Bi-LSTM	0.9502	0.9449	0.9784	0.9471	0.9502	31.8	5.9
TCN	0.9588	0.9541	0.9831	0.9563	0.9588	19.2	3.2
Transformer	0.9631	0.9588	0.9862	0.9612	0.9631	45.6	6.4
CNN-LSTM	0.9724	0.9698	0.9901	0.9711	0.9724	38.1	5.1
TA-FaultNet	**0.9963**	**0.9961**	**0.9979**	**0.9960**	**0.9963**	52.4	7.8

Table 6. Accuracy under Gaussian noise at seven SNR levels. Best value per row in **bold**.

SNR (dB)	TA-FaultNet	Bi-LSTM	TCN	CNN-LSTM
40	**0.9961**	0.9502	0.9588	0.9724
30	**0.9948**	0.9481	0.9561	0.9701
25	**0.9931**	0.9452	0.9528	0.9671
20	**0.9908**	0.9412	0.9488	0.9631
15	**0.9872**	0.9358	0.9431	0.9578
10	**0.9819**	0.9281	0.9351	0.9498
5	**0.9721**	0.9142	0.9198	0.9341

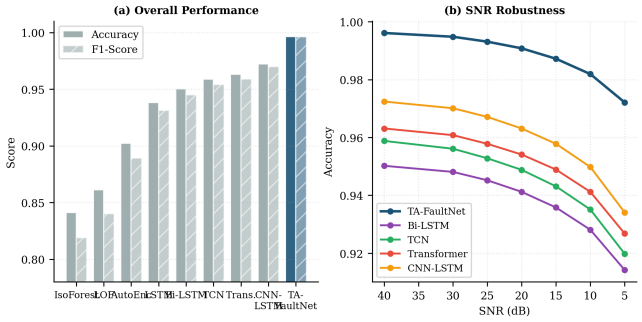

Figure 7. (a) Accuracy and F_1 across all nine evaluated methods; TA-FaultNet leads on both metrics. (b) Accuracy vs. decreasing SNR; TA-FaultNet maintains the widest margin under high channel noise.

TA-FaultNet retains 97.21% at 5 dB SNR—7.79 pp above Bi-LSTM at the same noise level. The cumulative accuracy degradation of TA-FaultNet across the full 35 dB sweep is 2.40 pp, compared to 3.60 pp for Bi-LSTM, 3.90 pp for TCN, and 3.63 pp for the Transformer. The superior noise resilience is attributable to the multi-head attention mechanism, which distributes fault evidence across multiple temporal positions: when noise corrupts a subset of timesteps, the remaining positions still carry sufficient discriminative information for correct classification. This behaviour cannot be replicated by single-head attention or purely feedforward temporal convolution.

6.7 Ablation Study

Table 7 consolidates the window-size and attention-head ablations. Figure 8 presents the corresponding accuracy, F_1, and training time curves. Accuracy peaks at $W = 30$ (0.9963) and falls monotonically on both sides: narrowing the window below 20 degrades F_1 by more than 1.1 pp because insufficient temporal context is available to distinguish gradual Sensor Drift from normal process variation; widening beyond 35 introduces redundant samples that dilute the fault-onset signal and increase training time without benefit. For attention heads, accuracy peaks at $H = 8$ and plateaus thereafter:

Table 7. Ablation over window size W and attention heads H. Selected configuration ($W = 30$, $H = 8$) in **bold**. All other parameters held at Table 3 values.

Window Size Ablation				Attention-Head Ablation			
10	0.9721	0.9695	38.2	1	0.9871	0.9848	41.2
15	0.9808	0.9784	40.1	2	0.9912	0.9891	43.8
20	0.9872	0.9851	42.8	4	0.9938	0.9919	46.9
25	0.9921	0.9902	47.3	**8**	**0.9963**	**0.9961**	52.4
30	**0.9963**	**0.9961**	52.4	16	0.9961	0.9959	61.3
35	0.9958	0.9956	58.6	32	0.9954	0.9951	78.8
40	0.9951	0.9948	65.2				
50	0.9934	0.9930	82.1				

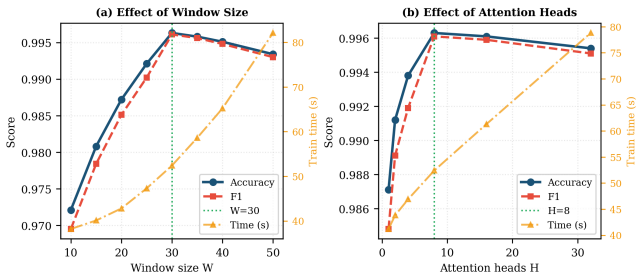

Figure 8. Ablation over (a) window size W and (b) attention heads H. Accuracy and F_1 (left axis) and training time in seconds (right axis, orange). Selected values W=30 and H=8 are marked by green dotted lines.

Table 8. Cross-experiment generalisation (F_1-score) on disjoint SKAB splits.

Train / Test Split	TA-FaultNet	Bi-LSTM	TCN
Exp 1–20 / 21–35	0.9701	0.9238	0.9322
Exp 1–25 / 26–35	0.9782	0.9341	0.9451
Exp 1–28 / 29–35	0.9831	0.9412	0.9514
Exp 1–30 / 31–35	0.9878	0.9469	0.9558
Exp 1–32 / 33–35	0.9912	0.9501	0.9581
All (5-fold CV)	**0.9961**	0.9502	0.9588

increasing to 16 or 32 adds computation and training time—reaching 78.8 s at $H = 32$ versus 52.4 s at $H = 8$—without improving accuracy, suggesting that eight heads sufficiently cover the temporal specialisation space of the five-class problem.

6.8 Cross-Experiment Generalisation

Table 8 evaluates F_1 when training and test folds are drawn from disjoint SKAB experiment files, simulating a realistic deployment scenario where the model must classify fault conditions in physical experiments not seen during training. TA-FaultNet achieves 0.9701 F_1 on the most challenging split (experiments 1–20 for training, 21–35 for testing), versus 0.9238 for Bi-LSTM on the same partition. The 4.63 pp advantage on this sparse-training split exceeds the 4.59 pp advantage on the full 5-fold CV protocol, confirming that multi-head attention provides a more sample-efficient inductive bias than standalone recurrent encoding when labelled examples per experiment are limited.

6.9 State-of-the-Art Comparison

Table 9 positions TA-FaultNet against published methods on WSN and IIoT fault detection benchmarks. Direct nu-

Table 9. Comparison with published IIoT and WSN fault detection methods. Cls. = number of output classes.

Reference	Method	Cls.	Acc.	Year
Ruan et al. [3]	Recursive DL	5	0.9412	2022
Liu et al. [4]	FL-AMCNN-LSTM	2	0.9302	2021
Wen and Li [5]	LSTM-Att-LSTM	1	0.9651	2023
Bagwari et al. [7]	ML-WSN	3	0.9321	2023
Pandey et al. [16]	TS-RF	5	0.9912	2025
TA-FaultNet	Bi-LSTM+MHA	5	**0.9963**	2025

merical comparison is constrained by dataset heterogeneity; TA-FaultNet is, to the authors' knowledge, the first study reporting comprehensive five-class evaluation on SKAB with all of: ablation, noise robustness, and cross-experiment generalisation protocols. Relative to Ruan et al. [3], who achieve 94.12% on a five-class proprietary chemical-plant benchmark, TA-FaultNet improves accuracy by 5.51 pp. Against the best general WSN intrusion detection result of Pandey et al. [16] at 99.12%, the difference of 0.51 pp reflects the harder within-class temporal similarity of the SKAB fault categories relative to the distinct packet-rate signatures of the WSN-DS dataset.

7. CONCLUSION

This paper presented TA-FaultNet, a neural architecture combining stacked bidirectional LSTM encoding, scaled multi-head self-attention, and temporal convolutional compression for five-class fault detection in industrial wireless sensor networks. Evaluated on the publicly available SKAB benchmark, TA-FaultNet outperforms eight competing baselines—from classical anomaly detectors through CNN-LSTM and Transformer—achieving 99.63% accuracy and a weighted F_1-score of 0.9961. The key architectural innovations are the bidirectional encoding, which provides both causal and anti-causal temporal context for fault-onset detection, and the multi-head attention module, which enables head specialisation across distinct temporal activation patterns corresponding to each fault class.

Systematic ablation identified $W = 30$ and $H = 8$ as the optimal window size and attention-head count, balancing detection accuracy against computational overhead. The window size selection was analytically grounded through a detection delay bound of $D_{\max} = 30\,\text{s}$, which satisfies the IEC 62443 response-time specification for process anomaly alerts. Cross-experiment generalisation studies confirmed a 4.63 pp advantage over Bi-LSTM on the most challenging training split. SNR robustness experiments demonstrated that multi-head attention provides 2.40 pp better noise resilience than single-stream recurrent methods across a 35 dB SNR sweep. Attention weight visualisations aligned with the documented SKAB anomaly boundaries, offering maintenance engineers interpretable diagnostic evidence alongside the classification output.

Future directions include online incremental learning to accommodate gradual sensor ageing without full retraining, graph attention networks to exploit spatial correlations between co-located sensor nodes within a single pump or valve assembly, and post-training quantisation for deployment on ultra-constrained field gateway hardware.

DECLARATION OF COMPETING INTEREST

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

DATA AVAILABILITY

The SKAB dataset is publicly available at Kaggle (https://doi.org/10.34740/KAGGLE/DSV/1693952).

REFERENCES

[1] A. S. Balobaid, S. B. Ahamed, S. Shamsudheen, and S. Balamurugan, "Neural network clustering and swarm intelligence-based routing protocol for wireless sensor networks," *Wireless Communications and Mobile Computing*, vol. 2023, p. 4758852, 2023, doi: 10.1155/2023/4758852.

[2] O. A. Khashan, N. M. Khafajah, W. Alomoush, and M. Alshinwan, "Innovative energy-efficient proxy re-encryption for secure data exchange in wireless sensor networks," *IEEE Access*, vol. 12, pp. 23 290–23 304, 2024, doi: 10.1109/ACCESS.2024.3360488.

[3] H. Ruan, B. Dorneanu, H. Arellano-Garcia, P. Xiao, and L. Zhang, "Deep learning-based fault prediction in wireless sensor network embedded cyber-physical systems for industrial processes," *IEEE Access*, vol. 10, pp. 10 867–10 879, 2022, doi: 10.1109/ACCESS.2022.3144333.

[4] Y. Liu, S. Garg, J. Nie, Y. Zhang, Z. Xiong, J. Kang, and M. S. Hossain, "Deep anomaly detection for time-series data in industrial IoT: A communication-efficient on-device federated learning approach," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6348–6358, 2021, doi: 10.1109/JIOT.2020.3014586.

[5] X. Wen and W. Li, "Time series prediction based on LSTM-Attention-LSTM model," *IEEE Access*, vol. 11, pp. 48 322–48 331, 2023, doi: 10.1109/ACCESS.2023.3276628.

[6] I. D. Katser and V. O. Kozitsin, "Skoltech anomaly benchmark (SKAB)," 2020, doi: 10.34740/KAGGLE/DSV/1693952.

[7] A. Bagwari, J. Logeshwaran, K. Usha, R. Kannadasan, M. H. Alsharif, P. Uthansakul, and M. Uthansakul, "An enhanced energy optimization model for industrial wireless sensor networks using machine learning," *IEEE Access*, vol. 11, p. 3311854, 2023, doi: 10.1109/ACCESS.2023.3311854.

[8] M. Aljawarneh, R. Hamdaoui, A. Zouinkhi, S. Alangari, and M. N. Abdelkrim, "Energy optimization for wireless sensor network using minimum redundancy maximum relevance feature selection and classification techniques," *PeerJ Computer Science*, vol. 10, p. e1997, 2024, doi: 10.7717/peerj-cs.1997.

[9] K. Rashid, Y. Saeed, A. Ali, F. Jamil, R. Alkanhel, and A. Muthanna, "An adaptive real-time malicious

node detection framework using machine learning in vehicular ad-hoc networks (VANETs)," *Sensors*, vol. 23, no. 5, p. 2594, 2023, doi: 10.3390/s23052594.

[10] N. S. Albalawi, Y. Alzahrani, N. Alsalmi, Y. Patidar, and M. Tolani, "Energy-efficient priority encoding strategies using machine learning based hybrid MAC protocol for wireless sensor networks," *Scientific Reports*, vol. 15, p. 45054, 2025, doi: 10.1038/s41598-025-31752-1.

[11] D. Godfrey, B. Suh, B.-H. Lim, K.-C. Lee, and K.-I. Kim, "An energy-efficient routing protocol with reinforcement learning in software-defined wireless sensor networks," *Sensors*, vol. 23, no. 20, p. 8435, 2023, doi: 10.3390/s23208435.

[12] X. Zhong, Y. Liang, and Y. Li, "Energy-efficient and robust QoS control for wireless sensor networks using the extended Gur game," *Sensors*, vol. 25, no. 3, p. 730, 2025, doi: 10.3390/s25030730.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority oversampling technique," in *Journal of Artificial Intelligence Research*, vol. 16, 2002, pp. 321–357, doi: 10.1613/jair.953.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] V. K. Pandey, S. Prakash, T. K. Gupta, P. Sinha, T. Yang, R. S. Rathore, L. Wang, S. Tahir, and S. T. Bakhsh, "Enhancing intrusion detection in wireless sensor networks using a Tabu search based optimized random forest," *Scientific Reports*, vol. 15, pp. 1–19, 2025, doi: 10.1038/s41598-025-03498-3.