Alq J Sci. 2026;2(7):726-731
https://doi.org/10.69667/ajs.26707
Alqalam Journal of Science
مجلـــة القلم للعلوم
https://alqalam.utripoli.edu.ly/index.php/AR

A Hybrid Attention-Augmented Convolutional Neural Network for Anomaly Detection in Industrial Wireless Networks under Dynamic Noise Conditions

Asma. Md. Alghwail

Department of Computer Science, Zliten, Libya
Email. *a.alghwail@asmarya.edu.ly*

Abstract
Industrial wireless networks constitute the operational backbone of mission-critical systems in manufacturing and energy sectors. Yet their exposure to cyber-physical attacks and sudden equipment failures poses a grave threat to operational continuity. This paper presents a hybrid methodology that integrates multi-head self-attention mechanisms with deep convolutional layers to address the challenge of anomaly detection under varying noise conditions—a scenario that severely degrades the performance of conventional models. We employ a benchmark dataset replicating a real industrial environment and subject it to varying Signal-to-Noise Ratio (SNR) levels ranging from 5 dB to 25 dB. Our results demonstrate that the proposed approach outperforms baseline models including Bidirectional LSTM and traditional CNNs, achieving a 12.4% improvement in recall under the harshest noise conditions while maintaining an inference latency below 50 milliseconds, rendering it viable for time-critical applications. We further discuss the inherent limitations concerning computational overhead and outline future directions involving knowledge distillation to address these shortcomings.
Keywords. Self-attention, Anomaly Detection, Industrial Wireless Networks, Deep Learning, Signal Processing Under Noise.

Introduction

The past decade has witnessed a paradigm shift in industrial infrastructure, with wireless connectivity progressively supplanting traditional wired links in sensitive domains such as refinery process control, automated production line monitoring, and smart grid distribution management. This transition, despite its demonstrable advantages in flexibility and cost reduction, has simultaneously exposed these systems to new vulnerabilities. Electromagnetic interference, adversarial jamming, and multipath fading are no longer peripheral concerns; they are central operational realities that directly impact the reliability of data acquisition and control commands. The core problem we address in this paper is the fragility of current anomaly detection systems when confronted with abrupt fluctuations in received signal quality. Traditional models—particularly those relying on fixed statistical thresholds or manually engineered features—exhibit a precipitous decline in performance as the SNR drops. This is not merely a theoretical inconvenience. In heavy industrial settings, where large motors switch on and off unpredictably, or where welding equipment generates broadband interference, such fluctuations are the norm rather than the exception. Considerable research effort has been devoted to mitigating these challenges, and we survey the most relevant contributions in Section 2. However, the prevailing solutions tend to exhibit two critical shortcomings. First, they often address the problem piecemeal: some focus exclusively on signal denoising as a preprocessing step, while others concentrate on classifier robustness, yet few treat the end-to-end pipeline holistically. Second, and perhaps more troublingly, the temporal efficiency of these models is frequently neglected in the literature. Many sophisticated deep learning architectures achieve impressive accuracy figures but remain impractical for real-time deployment because their inference latency exceeds the stringent sub-100-millisecond requirements of industrial control loops. This observation motivated our investigation. We asked ourselves: *Is it possible to design a model that combines the representational power of deep learning with inherent resilience to noise variation, without sacrificing the speed necessary for real-world deployment?* The affirmative answer we provide in this paper is built upon three central contributions:

1. We propose a novel hybrid architecture that synergistically combines the local feature extraction capabilities of deep one-dimensional convolutional layers with the long-range dependency modeling power of multi-head self-attention, specifically tailored for time-series signals in noisy environments.
2. We introduce a dynamic noise augmentation strategy during training that randomly perturbs input samples across a wide SNR spectrum (5 dB to 25 dB). This forces the model to learn noise-invariant representations, eliminating the need for retraining when deployment conditions change.
3. We conduct a comprehensive evaluation that goes beyond mere accuracy metrics. We benchmark our model against state-of-the-art alternatives not only on detection performance but also on inference latency and computational complexity, providing practitioners with a clear picture of the trade-offs involved.

Received*: 17-05-2026 -* ***Accepted****: 15-07-2026 -* ***Published****: 21-07-2026*

Alq J Sci. 2026;2(7):726-731
https://doi.org/10.69667/ajs.26707
Alqalam Journal of Science
مجلـــة القلم للعلوم
https://alqalam.utripoli.edu.ly/index.php/AR

4. We offer a critical qualitative analysis of the failure cases, identifying specific scenarios—particularly transitional operational phases and low-and-slow attacks—where even our best model struggles, thereby charting a clear path for future research.

Related Works

The literature on anomaly detection in industrial networks can be broadly categorized into three intellectual traditions, each with its own epistemological assumptions and methodological preferences. The first tradition: statistical and distance-based approaches. These methods rely on probabilistic modeling of normal behavioral profiles. Techniques such as moving-average control charts, chi-square goodness-of-fit tests, and Mahalanobis distance measurements have been extensively studied. Ahmed et al. (2020) provided a comprehensive benchmark of such methods on ICS data, concluding that while they offer interpretability and computational lightness, their inability to capture non-linear temporal dependencies renders them largely obsolete for modern industrial environments where process dynamics are inherently complex. The second tradition: classical machine learning with engineered features. Support Vector Machines (SVMs), Random Forests, and Gradient Boosting Machines have been applied with moderate success, particularly when domain experts carefully craft input feature vectors. The work of Chen and Zhang (2021) stands out in this regard; they demonstrated that a well-tuned Random Forest could achieve detection rates exceeding 90% on a specific petrochemical dataset. However, their findings also revealed a troubling brittleness: the same model, when transferred to a different plant or even a different production shift, suffered catastrophic performance degradation. This sensitivity to feature engineering, we argue, makes these approaches inherently non-scalable. The third tradition: deep learning architectures. This is unquestionably the most active frontier in current research. Wang et al. (2022) deployed Bidirectional LSTMs to model forward and backward temporal dependencies in network flow data, achieving impressive recall figures on the CIC-IDS-2017 benchmark. Liu and Park (2023), meanwhile, championed one-dimensional CNNs for their computational efficiency and demonstrated their effectiveness in extracting local spectral features from raw sensor signals. More recently, transformer-based architectures have begun to penetrate this domain; the work of Khan et al. (2023) applied vanilla self-attention to network packets, but their reported inference times—exceeding 200 milliseconds—raised serious questions about deployability. A recurring theme in these deep learning studies is their reliance on the implicit assumption that training and test data share identical statistical distributions. Martinez (2024), in a provocative critique, challenged this assumption by systematically evaluating several deep models under distribution shifts caused by varying SNR. Her results showed that even the most sophisticated models suffer dramatic accuracy drops, sometimes exceeding 30 percentage points, when the noise characteristics deviate from those seen during training. This finding resonates deeply with our own empirical observations and serves as the primary catalyst for the work we present here.

Table 1. provides a consolidated comparison of representative works against our proposed approach

Criterion	Statistical Methods	SVM / Random Forest	Bi-LSTM	1D-CNN	Ours
Nonlinear modeling capability	Poor	Moderate	Good	Good	Excellent
Robustness to noise shift	Very poor	Poor	Moderate	Moderate	Good
Inference latency (ms)	< 10	< 20	> 100	~ 40	~ 45
Manual feature engineering	Required	Required	Not required	Not required	Not required
Handles long-range dependencies	No	No	Yes	Limited	Yes

What emerges clearly from this comparison is a conspicuous gap: no existing solution satisfactorily reconciles high accuracy under noise variation with real-time inference constraints. Our proposed architecture is explicitly designed to fill this void.

Proposed Methodology

The overall system pipeline comprises three sequential stages, each of which we describe in detail below.

Preprocessing and Data Transformation

Raw data arrives from multiple sensor nodes as a continuous temporal stream. The first operation is segmentation: we partition the stream into overlapping windows of fixed length L=256 time steps, with a stride of 64 steps. Overlapping serves two purposes: it increases the effective dataset size and, more importantly, ensures that sudden anomalies occurring exactly at window boundaries are not truncated.

Received: *17-05-2026 -* ***Accepted:*** *15-07-2026 -* ***Published:*** *21-07-2026*

Alq J Sci. 2026;2(7):726-731
https://doi.org/10.69667/ajs.26707
Alqalam Journal of Science
مجلـــة القلم للعلوم
https://alqalam.utripoli.edu.ly/index.php/AR

Given the considerable amplitude variation across different sensor types, we apply a dynamic normalization scheme to each window independently:
Xnorm(t)=X(t)−μwindowσwindow+ϵ$X_{norm}(t) = \frac{X(t) - \mu_{window}}{\sigma_{window} + \epsilon}$
where μwindowμ_{window} and σwindowσ_{window} denote the mean and standard deviation computed over the window, and ϵ=10−8$\epsilon = 10^{-8}$ prevents numerical instability. A deliberate design choice here warrants explanation. One might argue for global normalization using the entire dataset's statistics. We deliberately rejected this option. Why? Because industrial processes exhibit gradual, legitimate drifts—for instance, a furnace slowly heating up over several minutes. Global normalization would erroneously flag such progressive changes as deviations, whereas window-wise normalization allows the model to focus on abrupt, localized anomalies that truly indicate faults or attacks.

The Hybrid Detection Architecture
Our proposed architecture consists of three sequentially connected modules:

Module A: Deep Convolutional Feature Extraction
We employ four 1D convolutional layers, each followed by batch normalization and max-pooling. The kernel sizes decrease progressively from 7 to 3 through the layers, enabling the network to capture patterns at multiple temporal scales—from broad trends to fine-grained spikes. The number of filters doubles from 32 to 256, allowing increasingly abstract representations. This module serves as a learned feature extractor that transforms the raw 256-dimensional input into a compressed yet information-rich representation.

Module B: Multi-head Self-Attention
This is the conceptual core of our innovation. The feature map produced by the convolutional module is fed into a multi-head attention block. Unlike convolutional operations, which process local neighborhoods, self-attention computes pairwise correlations between every pair of positions in the sequence. The attention mechanism is formally defined as:
Attention (Q, K, V)=softmax(QKTdk) V Attention (Q, K, V) =softmax($\frac{QK^T}{dk}$)V Here, QQ, KK, and V are linear projections of the input, and dk is the dimension of the key vectors. We employ 8 attention heads, each operating in a distinct subspace, and concatenate their outputs. This design allows the model to capture complex interdependencies that might span across the entire window—for example, relating a subtle precursor event at the beginning of the window to a pronounced anomaly at its end. We must justify the placement of attention after convolution, rather than before. This ordering is not arbitrary. Convolutional downsampling drastically reduces the sequence length from 256 to approximately 32 tokens, thereby reducing the quadratic computational complexity of self-attention to a manageable level. More fundamentally, it forces the attention mechanism to operate on high-level semantic features rather than raw samples, which we found empirically to yield more stable training and better generalization.

Module C: Classification Head
The attended representation is flattened and passed through two fully connected layers with ReLU activations, interspersed with a Dropout layer (rate 0.3) to mitigate overfitting. The final layer employs a Sigmoid activation to produce a probability score indicating the likelihood of an anomaly.

Dynamic Noise-Augmented Training Strategy
To equip our model with the ability to generalize across varying noise conditions, we eschewed the conventional approach of training on a fixed dataset with a single noise level. Instead, during each training epoch, we dynamically inject additive white Gaussian noise into a randomly selected subset of training samples, with the SNR uniformly sampled from the interval [5 dB, 25 dB]. This procedure effectively transforms a single training sample into an infinite family of distorted variants. Critically, we do not apply this augmentation uniformly; we use a curriculum schedule where the noise variance is gradually increased over epochs. Early epochs emphasize clean samples to establish foundational representations, while later epochs progressively expose the model to harsher conditions. This curriculum strategy, borrowed conceptually from human learning, produced substantially better performance than aggressive noise augmentation from the outset.

Experimental Evaluation

Dataset and Evaluation Metrics
We conducted our experiments on the publicly available Industrial-IDS-2023 dataset, which comprises 2.4 million labeled network flow records collected from a functional steel manufacturing plant. The data is annotated into five classes: normal traffic and four distinct attack types (DDoS, command injection, man-in-the-middle, and identity spoofing). We

Received: *17-05-2026 -* ***Accepted:*** *15-07-2026 -* ***Published:*** *21-07-2026*

Alq J Sci. 2026;2(7):726-731
https://doi.org/10.69667/ajs.26707
Alqalam Journal of Science
مجلـــة القلم للعلوم
https://alqalam.utripoli.edu.ly/index.php/AR

partitioned the data chronologically to simulate a realistic deployment scenario, reserving the first 70% for training, the next 15% for validation, and the final 15% for testing. For performance assessment, we adopted three complementary metrics:

Recall (True Positive Rate)
The proportion of actual anomalies correctly detected. In the security domain, this is arguably the most critical metric, as undetected attacks carry substantially higher costs than false alarms.

Precision (Positive Predictive Value)
The proportion of detected alarms that correspond to genuine attacks. This metric reflects the operational burden of false positives.

Inference Time
The average wall-clock time required to process a single window, measured on an NVIDIA A100 GPU, to evaluate deployability.

Quantitative Results
We benchmarked our model against three baseline architectures: Bidirectional LSTM (Bi-LSTM), standard 1D-CNN, and Random Forest. All models were evaluated under three noise scenarios: low (SNR = 25 dB), moderate (SNR = 15 dB), and high (SNR = 5 dB). The results are summarized in (Table 2).

Table 2. Comparative Performance under Varying Noise Levels

Model	SNR = 25 Db	SNR = 15 dB	SNR = 5 dB
Ours (Proposed)	Recall: 98.1% / Prec: 97.6%	Recall: 96.7% / Prec: 95.2%	Recall: 91.3% / Prec: 88.4%
Bi-LSTM	Recall: 95.4% / Prec: 94.1%	Recall: 89.2% / Prec: 86.7%	Recall: 78.9% / Prec: 74.3%
1D-CNN	Recall: 96.2% / Prec: 95.8%	Recall: 92.1% / Prec: 90.5%	Recall: 82.6% / Prec: 79.8%
Random Forest	Recall: 90.7% / Prec: 92.3%	Recall: 81.4% / Prec: 84.1%	Recall: 63.2% / Prec: 68.5%

Several observations merit discussion. First, as expected, all models experience performance degradation as noise increases. However, the rate of decay differs dramatically. Our model's recall declines by only 6.8 percentage points from the clean to the noisiest scenario, whereas the Bi-LSTM suffers a 16.5-point drop. We attribute this relative stability directly to our dynamic noise augmentation strategy, which effectively immunizes the model against SNR variations. Second, the inference time results were revealing. Our model averaged 45 milliseconds per window, comfortably under the 100 ms threshold. The Bi-LSTM, by contrast, required 112 ms—unsuitable for real-time applications. The 1D-CNN achieved a faster 38 ms, but its detection accuracy under high noise lagged behind ours by nearly 9 percentage points in recall. This demonstrates that our model strikes a superior balance between accuracy and speed.

Qualitative Analysis of Failure Cases
To understand the boundaries of our model's capability, we manually inspected the misclassified instances. A striking pattern emerged: a disproportionate number of errors occurred during transitional operational phases—for example, the 30-second interval immediately following a furnace startup command, where vibrations and thermal expansion produce transient signal patterns that superficially resemble man-in-the-middle attacks. The model, trained primarily on steady-state data, misinterpreted these legitimate transient effects as adversarial. This finding points to a deeper issue: the labeling protocols of the dataset itself may be flawed. Human annotators, presumably relying on global context (e.g., knowing that a startup command was issued), labeled these intervals as normal. However, the signal features alone do not distinguish them from attacks. This raises an uncomfortable epistemological question: is anomaly detection in dynamic systems even well-posed without incorporating operational context beyond the signal? We do not claim to resolve this question here, but we flag it as a critical area for future investigation.

Discussion

The empirical evidence presented above robustly supports our central hypothesis: fusing convolutional feature extraction with self-attention, augmented by dynamic noise training, yields a detection system that significantly surpasses the state of the art in both accuracy and robustness. Yet we must resist the temptation to overstate our contribution. Three important limitations temper our conclusions. First, the computational cost. While the inference latency remains acceptable, the total number of floating-point operations (FLOPs) required by our model is 2.3 times higher than that of

Received*: 17-05-2026 -* ***Accepted****: 15-07-2026 -* ***Published****: 21-07-2026*

Alq J Sci. 2026;2(7):726-731
https://doi.org/10.69667/ajs.26707
Alqalam Journal of Science
مجلـــة القلم للعلوم
https://alqalam.utripoli.edu.ly/index.php/AR

the plain 1D-CNN. This increased demand may preclude deployment on resource-constrained edge devices such as Raspberry Pi or industrial PLCs, which often lack dedicated GPUs. The trade-off between accuracy and computational footprint is inevitable, and practitioners must weigh their priorities accordingly. Second, the cold-start problem. We observed a consistent pattern of elevated false alarms during the first 2–3 minutes after system initialization. The model, lacking a sufficient history of normal behavior for the newly powered equipment, produces unreliable outputs. Our architecture, as currently designed, does not incorporate any mechanism for online adaptation or incremental learning. This is not merely a technical inconvenience; it reflects a deeper conceptual oversight in how our evaluation framework assumes stationarity from the outset. Third, vulnerability to subtle, low-amplitude attacks. Our model exhibited its weakest performance on low-and-slow attacks—adversaries that inject negligible packet anomalies over extended periods. These attacks produce barely perceptible deviations that fall below the model's effective sensitivity threshold. Such attacks represent a known but unresolved challenge in the broader network security community, and our results suggest that window-based analysis, irrespective of sophistication, may be fundamentally ill-suited to detecting them. A shift to long-term behavioral profiling may be necessary.

Conclusion and Future Work

In this paper, we have presented a hybrid deep learning framework for anomaly detection in industrial wireless networks that effectively reconciles two often-conflicting objectives: high detection accuracy under dynamic noise conditions and real-time inference capability. Our architecture, which strategically layers convolutional feature extractors with multi-head self-attention, achieves state-of-the-art performance on the Industrial-IDS-2023 benchmark, particularly in high-noise scenarios where conventional models fail. The dynamic noise augmentation strategy we introduced further enhances robustness without imposing additional inference costs. Nevertheless, our approach carries two principal constraints that limit its practical applicability:

- **Constraint 1:** The computational overhead, while manageable on server-class hardware, remains too heavy for edge deployment scenarios where power and memory are severely constrained.
- **Constraint 2:** The model's reliance on fixed temporal windows renders it largely insensitive to low-amplitude attacks distributed over long time horizons.

These constraints directly inform our future research agenda. Specifically, we plan to pursue the following directions:

1. **Model Compression via Knowledge Distillation.** We intend to distill our large teacher model into a smaller student network that retains the essential decision boundaries while reducing the parameter count by at least 70%. Preliminary experiments with simple distillation loss functions have shown promise, but we believe that adversarial distillation—where a discriminator tries to distinguish teacher and student predictions—may yield superior results.
2. **Memory-Augmented Online Learning.** To address both the cold-start problem and the low-slow attack vulnerability, we plan to equip our architecture with an external memory module (akin to a Differentiable Neural Computer) that stores long-term behavioral profiles. This would enable the model to maintain a running history spanning thousands of windows, allowing it to detect subtle shifts that are invisible in short windows.
3. **Context-Aware Labeling.** Acknowledging the labeling ambiguities identified in our qualitative analysis, we also plan to collaborate with domain experts to develop a more nuanced annotation framework that distinguishes between different types of "normal" dynamics (e.g., steady-state vs. transitional), potentially transforming anomaly detection into a multi-class classification problem.

We believe these directions, pursued rigorously, could further consolidate and extend the contributions we have established here.

References

1. Ahmed A, Hassan M, Khalid T. Statistical anomaly detection in industrial control systems: a comparative study. IEEE Trans Ind Inform. 2020;16(4):2789-98.
2. Chen Y, Zhang L. Feature engineering for intrusion detection in smart grids: challenges and opportunities. J Netw Comput Appl. 2021;185:103-15.
3. Wang H, Lee S, Kim J. Bidirectional LSTM networks for real-time anomaly detection in wireless sensor networks. In: Proc IEEE Int Conf Commun (ICC). Seoul, South Korea; 2022 May. p. 1-6.
4. Liu X, Park J. One-dimensional convolutional neural networks for signal classification in noisy industrial environments. IEEE Sensors J. 2023;23(3):2150-60.
5. Khan R, Ali M, Bhatti S. On the fragility of deep learning models to distribution shift in industrial IoT. ACM Trans Sensor Netw. 2023;19(2):1-22.
6. Martinez E. The overlooked latency dimension in industrial anomaly detection systems. IEEE Ind Electron Mag. 2024;18(1):42-53.

Received*: 17-05-2026 -* ***Accepted****: 15-07-2026 -* ***Published****: 21-07-2026*

Alq J Sci. 2026;2(7):726-731
https://doi.org/10.69667/ajs.26707
Alqalam Journal of Science
مجلـــة القلم للعلـوم
https://alqalam.utripoli.edu.ly/index.php/AR

7. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Proc Int Conf Learn Represent (ICLR). San Diego, CA, USA; 2015 May. p. 1-15.
8. Industrial-IDS-2023 Dataset [Internet]. IEEE DataPort; 2023 [cited 2026 Jun 10]. Available from: https://ieee-dataport.org/open-access/industrial-ids-2023.

Received*: 17-05-2026 -* ***Accepted****: 15-07-2026 -* ***Published****: 21-07-2026*