Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/23365
DOI URL: http://dx.doi.org/10.21474/IJAR01/23365

ISSN (O): 2320-5407
ISSN (P): 3107-4928

RESEARCH ARTICLE

EDGE AI FOR INDUSTRIAL PREDICTIVE MAINTENANCE: A DUAL-ATTENTION APPROACH WITH EXPLAINABILITY AND UNCERTAINTY QUANTIFICATION

Sruti Ranjan Sahoo[1] and Dr. Kanhu Charan Bhuyan[2]

1. School of Electronics Science Odisha University of Technology & Research (OUTR) Bhubaneswar, Odisha, India.
2. Associate Professor School of Electronics Science,OUTR Bhubaneswar, Odisha, India.

Manuscript Info

Manuscript History
Received: 8 February 2026
Final Accepted: 10 March 2026
Published: April 2026

Key words:-
Edge AI, predictive maintenance, remaining useful life, dual-attention, explainable AI, uncertainty quantification, zero trust, Industrial IoT, sustainable computing.

Abstract

Industrial Internet of Things (IIoT) systems generate continuous sensor data streams from rotating machinery. Cloud based artificial intelligenc e achieves high accuracy for Remaining Useful Life (RUL) prediction but introduces 100–500 ms round-trip latency, which may be unacceptable for real-time failure prevention. Edge AI offers an alternative but faces computational, explainability, and security constrai nts that existing approaches rarely address jointly. This paper presents a unified framework integrating four dimensions: (1) a dual-attention neural architecture for RUL prediction that explicitly models channel-wise sensor importance and temporal relevance; (2) SHAP-based explainability with quantified overhead on edge hardware; (3) Monte Carlo Dropout for uncertainty calibration; and (4) a Zero Trust security layer for Operational Technology (OT) deployment. The framework targets the NVIDIA Jetson Orin Nano class of devices with an inference-latency goal of under 50 ms. This is a methodology paper; a six-phase experimental protocol benchmarks the approach against verified state-of-the-art (HMDAM: RMSE 10.82; DAST: RMSE 11.43 on NASA C-MAPSS FD001). Architecture and protocol are complete; experimental results will follow in an extended version. All quantitative claims are explicitly marked as literature-verified or target-under-validation.

Introduction:-

Industrial rotating machinery—including turbines, pumps, bearings, and turbofan engines—forms the backbone of manufacturing, energy, and transportation infrastructure. Unplanned failures incur substantial costs: data-center downtime has been estimated at an average of USD 138,000 per hour [1], with comparable or higher figures

Corresponding Author:- Sruti Ranjan Sahoo
Address**:-**School of Electronics Science Odisha University of Technology & Research (OUTR) Bhubaneswar, Odisha, India.

reported for offshore wind (O&M at 20–35% of electricity revenue) and oil & gas (maintenance at 15–70% of production cost) [1]. Predictive maintenance (PdM) aims to transition from reactive and preventive strategies to condition-based intervention triggered by degradation indicators learned directly from sensor data. The proliferation of Industrial Internet of Things (IIoT) sensors has made high-velocity multivariate time-series data abundantly available for deep-learning-based Remaining Useful Life (RUL) prediction. However, two fundamental challenges constrain practical deployment at scale.

Latency constraints. Cloud-based AI architectures require sensor data to be transmitted to remote servers, introducing 100–500 ms round-trip latency plus processing time. For rotating machinery operating at thousands of revolutions per minute, critical failure modes can develop within seconds; a 500 ms warning window may be insufficient for emergency shutdown.

Multi-dimensional deployment requirements. Industrial deployment typically requires simultaneous satisfaction of four criteria: (i) accuracy competitive with server-based models; (ii) explainability for regulatory compliance and operator trust; (iii) uncertainty quantification (UQ) for risk-based maintenance scheduling; and (iv) security appropriate for Operational Technology (OT) environments subject to cyber-physical attacks. Existing approaches typically optimize accuracy in isolation.

Scope and status. This paper addresses methodology and experimental-protocol design for an edge-deployable RUL prediction framework. The following limitations are explicitly acknowledged: (1) all performance metrics for the proposed model are targets pending experimental execution; (2) validation uses NASA C-MAPSS, a simulated turbofan dataset, and generalization to other machinery types requires further study; (3) economic validation requires a multi-month pilot and is treated as a methodological proposal rather than a verified outcome; (4) security testing occurs in an isolated testbed, with production validation deferred to future work.

The specific contributions of this paper are: (1) a dual-attention neural architecture explicitly modeling channel-wise sensor importance and temporal relevance for multivariate time-series RUL prediction; (2) a six-phase experimental protocol validating accuracy, latency (<50 ms), explainability overhead, uncertainty calibration (Expected Calibration Error, ECE, <0.05), and Zero Trust security on NVIDIA Jetson Orin Nano; and (3) a claim-audit methodology that separates literature-verified values from targets-under-validation, supporting reproducibility and reviewer trust.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the proposed framework. Section IV details the experimental protocol. Section V discusses literature benchmarks and current status. Section VI concludes with future work.

Related Work:-

RUL Prediction Architectures:-

Deep learning approaches for RUL prediction have evolved from Convolutional Neural Networks (CNNs) for local feature extraction [2] to Long Short-Term Memory (LSTM) networks for temporal modeling [3]. Attention mechanisms, introduced by Vaswani et al. [4] for natural language processing, have been adapted to time series to model temporal dependencies without recurrent computation.

Dual-attention mechanisms. Recent work has explored combining multiple attention types for multivariate time series. He et al. [5] proposed HMDAM (Hybrid Model with Dual-Attention Mechanism), integrating temporal and sensor feature-extraction blocks, and reported RMSE 10.82, Score 170.07 on NASA C-MAPSS FD001 with 89.02K parameters and 821.25K FLOPs. Zhang et al. [6] introduced DAST (Dual Aspect Self-Attention Transformer), reporting RMSE 11.43 on FD001 with a 10.02% reduction on the more complex FD002 subset over prior baselines. Our work extends these foundations by targeting edge deployment under an explicit latency budget and by integrating explainability, uncertainty quantification, and security—dimensions not jointly addressed in prior dual-attention RUL work.

Explainable AI for Time Series:-

Post-hoc explanation methods—including SHAP (SHapley Additive exPlanations) [7] and LIME [8]—provide feature-importance scores for individual predictions. For time series, SHAP values indicate which sensors and time steps influenced an RUL estimate. However, SHAP computation involves multiple model evaluations, creating non-trivial overhead for resource-constrained edge devices. This overhead is rarely measured in the RUL literature and is one of the gaps our experimental protocol targets [9].

Uncertainty Quantification:-
Bayesian Neural Networks [10] and Monte Carlo (MC) Dropout [11] enable uncertainty estimation from deep models. Expected Calibration Error (ECE) measures alignment between predicted confidence and actual accuracy [12]. Abdar et al. [13] provide a comprehensive survey classifying UQ methods into Bayesian approximations and ensemble approaches. While UQ has been extensively studied for classification, its application to RUL regression under edge deployment constraints remains underexplored.

Edge AI and OT Security:-
NVIDIA Jetson platforms have enabled edge AI deployment for computer vision [14] and, increasingly, time-series applications. Jetson Orin Nano (2023) offers a 1024-core Ampere GPU and 32 Tensor Cores within a 7–15 W power envelope [15], suitable for industrial edge deployment. NIST SP 800-82 Rev. 3 [16] provides current guidance for OT security, and NIST SP 800-207 [17] codifies Zero Trust Architecture (ZTA) principles—identity verification, least privilege, micro-segmentation, and continuous monitoring—that are increasingly applied to cyber-physical systems.

Research Gap:-
To our knowledge, no prior work unifies dual-attention RUL prediction with explainability-overhead measurement, uncertainty calibration, and Zero Trust security validation in a single edge-deployable framework. Existing approaches optimize single dimensions in isolation; this work targets the complete deployment pipeline relevant to sustainable industrial computing.

Proposed Framework:-
System Architecture:-
The proposed framework comprises four integrated components: (1) a Dual-Attention RUL Predictor; (2) a SHAP Explainability Module; (3) a Monte Carlo Dropout UQ Module; and (4) a Zero Trust Security Agent. The input is a multivariate time-series window $X \in R^{(B\times T\times C)}$, where B is batch size, T = 50 time steps, and C = 14 sensor channels after preprocessing. The output is an RUL prediction $\hat{y} \in R^{(B\times 1)}$ together with an uncertainty estimate and a SHAP attribution.

Dual-Attention RUL Predictor:-
Channel attention computes importance weights across sensor channels using Global Average Pooling (GAP) followed by a learned projection: $\alpha_c = \text{Softmax}(W_c \cdot \text{GAP}(X) + b_c)$, with $X_{channel} = X \odot \alpha_c$. Temporal attention applies multi-head self-attention with H = 8 heads and model dimension D = 128: Q_h, K_h, V_h are learned linear projections of the channel-attended input, and $A_h = \text{Softmax}(Q_h K_h^T / \sqrt{d_k}) V_h$ with $d_k = D/H = 16$.

Dual-attention fusion combines the two streams via element-wise multiplication followed by residual connection and layer normalization: $H_{fused} = \text{LayerNorm}(X_{proj} + H_t \odot H_c)$. A position-wise feed-forward network and a GAP+linear regression head produce the RUL estimate $\hat{y}$. Training uses Huber loss with $\delta = 10$ cycles for robustness to outliers. The resulting model contains approximately 224K parameters—of the same order of magnitude as HMDAM (89.02K)—with an estimated 16M FLOPs per forward pass at T = 50. Final parameter count will be fixed after the architecture search step in EXP-01.

SHAP Explainability Module:-
We employ DeepExplainer and KernelExplainer from the SHAP library [7]. To respect edge-memory constraints, background datasets are limited to 100 samples. The SHAP decomposition $f(x) = \varphi_0 + \Sigma_i \varphi_i x_i$ expresses a prediction as a sum of feature contributions with theoretical guarantees of local accuracy, missingness, and consistency.

Monte Carlo Dropout UQ:-
Following Gal and Ghahramani [11], dropout is enabled at inference to approximate Bayesian inference. For N = 100 stochastic forward passes, we report the predictive mean $\hat{y}_{mean}$ and standard deviation $\hat{y}_{std}$. Expected Calibration Error is computed by binning predictions and comparing average confidence with observed accuracy [12]. Temperature scaling is applied post-hoc when ECE exceeds the 0.05 threshold.

Zero Trust Security Agent:-
A lightweight security agent implements ZTA principles [17] around the inference endpoint: (i) continuous device attestation via TPM 2.0; (ii) mutual TLS for all sensor-to-inference and inference-to-supervisor communications; (iii) micro-segmentation isolating the inference service from the wider OT network; and (iv) certificate-based authentication with revocation checking. The security layer is deliberately minimal to keep added inference-path latency within a target budget of 5 ms.

Experimental Protocol:-
A six-phase protocol ensures systematic validation across the four deployment dimensions. All phases use standardized hardware (Jetson Orin Nano Developer Kit, JetPack 5.1.2, TensorRT 8.6, CUDA 11.8) and fixed random seeds to support reproducibility.

EXP-01: RUL Benchmarking:-
The dual-attention model is compared against HMDAM [5] and DAST [6] on NASA C-MAPSS FD001–FD004 [18]. Metrics: RMSE, asymmetric Score (penalizing late predictions more heavily), and MAE. Protocol: 20 runs (5 seeds × 4 subsets) per model; report mean ± standard deviation. Success target: RMSE within 10% of HMDAM on FD001 (i.e., <11.90).

EXP-02: Latency Validation:-
Inference latency is measured on Jetson Orin Nano using CUDA events across 1,000 runs per precision mode (FP32 baseline, FP16, INT8 with calibration), with 100 warm-up runs discarded. Reported statistics: mean, standard deviation, p50, p95, p99. Success targets: mean < 50 ms (FP16), p95 < 75 ms, p99 < 100 ms.

EXP-03: SHAP Overhead:-
Latency and memory overhead of DeepExplainer (100-sample background) and KernelExplainer (50 and 100 samples) are measured by A/B comparison against inference-only baseline. Explanation stability is assessed via correlation across five runs on identical input. Targets: latency overhead < 20%, memory overhead < 50%.

EXP-04: UQ Calibration:-
MC Dropout with N = 100 forward passes produces per-prediction uncertainty. Calibration is evaluated via ECE and Maximum Calibration Error (MCE), with reliability diagrams generated on the validation split. Targets: ECE < 0.05 and MCE < 0.15 after optional temperature scaling.

EXP-05: Zero Trust Validation:-
An isolated testbed with PLC simulators (Modbus TCP on ports 5020–5024) and a Kali Linux attacker VM is used. Authentication scenarios (valid, expired, missing, revoked certificates) are each run 100 times to measure rejection accuracy. Penetration testing includes nmap scanning, Hydra brute force, and Ettercap man-in-the-middle. Target: <5 ms average authentication overhead.

EXP-06: Economic Validation Methodology:-
A reproducible cost-benefit methodology is defined using two years of historical maintenance records (MTBF, MTTR, downtime cost, labor, spares) as baseline, combined with a pilot deployment on 5–10 representative machines over six months. Monte Carlo simulation propagates uncertainty in cost inputs. No ROI figure is claimed in the present paper; the methodology itself is the deliverable.

Literature Benchmarks and Current Status:-
Verified State-of-the-Art:-
Table I summarizes peer-reviewed benchmarks from the dual-attention RUL literature. These are the targets our experimental protocol must match to establish competitive accuracy. All listed values are reproduced from the cited papers and are literature-verified; our proposed-model row reports targets that are currently under validation.

Table:- **I**
Literature Benchmarks and Proposed Target On NASA C-MAPSS FD001.

Model	Year	RMSE	Score	Params
HMDAM [5]	2025	10.82	170.07	89.0K

DAST [6]	2022	11.43	203.15	n/r
Proposed (target)	2026	<11.90	<187	~224K

Notes: HMDAM values from He et al. [5] (RMSE 10.82, Score 170.07, 89.02K params, 821.25K FLOPs). DAST values from Zhang et al. [6] (RMSE 11.43, 10.02% improvement over baseline LSTM on FD002). Proposed target is within 10% of HMDAM; 'n/r' = not reported in source paper.

Current Status:-
The dual-attention architecture and experimental protocol are complete. HMDAM-baseline reproduction and dataset preprocessing pipeline are in progress. EXP-01 through EXP-05 are scheduled across the next six months, with EXP-06 running concurrently for a further six months. The present paper reports design and methodology; comprehensive results will be reported in an extended version.

Claim Audit:-
Table II applies an evidence-discipline audit to every numeric claim in the paper, separating literature-verified values, preliminary design estimates, and targets under validation. This is intended to support both reviewer scrutiny and reproducibility.

Table:- II
Claim audit: evidence status for numeric claims.

Claim	Value	Status
HMDAM RMSE (FD001)	10.82	Literature-verified [5]
HMDAM Score (FD001)	170.07	Literature-verified [5]
HMDAM parameters	89.02K	Literature-verified [5]
HMDAM FLOPs	821.25K	Literature-verified [5]
DAST RMSE (FD001)	11.43	Literature-verified [6]
DAST FD002 improvement	10.02%	Literature-verified [6]
Orin Nano GPU	1024 Ampere cores	Literature-verified [15]
Orin Nano power	7–15 W	Literature-verified [15]
Datacenter downtime cost	$138K/hour	Literature-cited [1]
Proposed parameters	~224K	Preliminary design
Proposed FLOPs	~16M	Preliminary design
Target latency (FP16)	<50 ms	Target — to be measured
Target RMSE (FD001)	<11.90	Target — to be measured
SHAP overhead target	<20% latency	Target — to be measured
UQ ECE target	<0.05	Target — to be measured
ZTA auth overhead target	<5 ms	Target — to be measured

Limitations:-
(1) C-MAPSS represents simulated turbofan degradation; generalization to pumps, bearings, and gearboxes requires additional validation. (2) An isolated testbed cannot replicate long-horizon Advanced Persistent Threat scenarios. (3) Results are hardware-specific; other edge platforms (Intel NUC, Coral TPU, Raspberry Pi) may exhibit different latency–accuracy profiles. (4) Economic figures will remain methodological proposals until the six-month pilot concludes.

Conclusion and Future Work:-

This paper presented a unified framework for edge-deployable predictive maintenance that integrates dual-attention RUL prediction, SHAP explainability, Monte Carlo Dropout uncertainty quantification, and Zero Trust security. A six-phase protocol benchmarks the approach against verified state-of-the-art (HMDAM RMSE 10.82, DAST RMSE 11.43) and validates latency (<50 ms), explainability overhead, UQ calibration (ECE <0.05), and security authentication overhead on Jetson Orin Nano. Immediate future work (0–8 months) executes EXP-01 through EXP-06 and reports statistically significant results. Near-term work (6–12 months) targets industrial pilot deployment; long-term work (12–24 months) extends to additional machinery types and investigates federated learning for multi-site deployment. The evidence-discipline framing—explicit separation of literature-verified values from targets-under-validation—is proposed as a replicable practice for methodology papers in sustainable industrial computing.

References:-

[1] Z. Zhu et al., "A survey of predictive maintenance: Systems, purposes and approaches," arXiv preprint, 2024.
[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.
[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[4] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
[5] K. He, Z. Li, C. Zheng, Z. Zhang, and L. Zhang, "A hybrid model based on a dual-attention mechanism for the prediction of remaining useful life of aircraft engines," Sensors, vol. 25, no. 18, art. 5682, 2025.
[6] Y. Zhang, W. Song, and Q. Li, "Dual aspect self-attention based on Transformer for remaining useful life prediction," IEEE Trans. Instrumentation and Measurement, 2022, arXiv:2106.15842.
[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[9] S. Moss et al., "Explainable AI in IoT: A survey of challenges, advancements, and pathways to trustworthy automation," Applied Sciences, MDPI, 2025.
[10] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," in Proc. ICML, 2015, pp. 1613–1622.
[11] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. ICML, 2016, pp. 1050–1059.
[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. ICML, 2017, pp. 1321–1330.
[13] M. Abdar et al., "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," Information Fusion, vol. 76, pp. 243–297, 2021.
[14] A. Chakraborty et al., "Profiling concurrent vision inference workloads on NVIDIA Jetson—Extended," arXiv:2508.08430, 2025.
[15] NVIDIA Corporation, "Jetson Orin Nano Developer Kit: Technical specifications," NVIDIA Developer Documentation, 2023.
[16] National Institute of Standards and Technology, "Guide to Operational Technology (OT) security," NIST SP 800-82 Rev. 3, Sept. 2023.
[17] National Institute of Standards and Technology, "Zero Trust Architecture," NIST SP 800-207, 2020.
[18] A. Saxena and K. Goebel, "Turbofan engine degradation simulation data set," NASA Ames Prognostics Data Repository, 2008.