JOURNAL OF ADVANCED ELECTRICAL AND ELECTRONIC ENGINEERING

Vol 2 Issue 1 (2026)

http://jaeee.org/

Rolling Horizon Neurocomputing Optimizes Industrial IoT Compressor Monitoring at the Edge Under Constraints and Drift

Audrey Fleming
Faculty of Science, University of Helsinki, Helsinki, Finland

Abstract
The rapid integration of the Industrial Internet of Things into modern manufacturing environments has fundamentally transformed the operational paradigms of heavy machinery management. Among the most critical assets in these industrial ecosystems are industrial air and gas compressors, which serve as the lifeblood of pneumatic processes and automated control systems. However, the continuous and reliable monitoring of such dynamic equipment presents significant challenges due to the high volatility of sensor data, the computational constraints of edge devices, and the inherent latency in centralized cloud architectures. This paper proposes a novel framework utilizing rolling-horizon neurocomputing optimization tailored specifically for Industrial Internet of Things compressor monitoring networks. By hybridizing predictive control philosophies with advanced artificial neural networks distributed across edge nodes, the proposed system dynamically adapts its optimization horizon to incoming data streams, thereby minimizing transmission latency and maximizing predictive anomaly detection accuracy. Extensive simulations demonstrate that this rolling-horizon neurocomputing approach significantly outperforms static diagnostic models by reducing communication overhead while maintaining superior predictive fidelity under transient operational states. The findings provide a robust theoretical foundation and practical deployment strategy for next-generation smart manufacturing facilities seeking to achieve zero-downtime operations through intelligent, decentralized machinery prognostics.

Keywords

Industrial Internet of Things, Neurocomputing, Rolling-Horizon Optimization, Predictive Maintenance, Edge Computing.

INTRODUCTION

1.1 Background on Industrial Internet of Things and Machinery Monitoring

The transition toward Industry 4.0 has been largely driven by the proliferation of the Industrial Internet of Things, a paradigm that interlinks physical machinery with digital analytical

frameworks through vast networks of deployed sensors [1]. In contemporary manufacturing and processing plants, continuous data streams form the foundation for operational intelligence, enabling plant managers to transition from reactive maintenance schedules to highly proactive, predictive strategies. Industrial compressors, which are ubiquitous across manufacturing, chemical processing, and energy generation sectors, represent a particularly critical class of machinery. These devices operate under extreme mechanical stress, thermodynamic fluctuations, and continuous cyclic loading, making them highly susceptible to wear, degradation, and sudden catastrophic failure [2]. The sudden failure of a primary industrial compressor can lead to total plant shutdowns, resulting in massive financial losses, compromised safety protocols, and severe supply chain disruptions [3]. Consequently, the implementation of robust, real-time monitoring networks is not merely an operational luxury but an absolute necessity for modern industrial continuity. Historically, compressor monitoring relied heavily on supervisory control and data acquisition systems that utilized hard-coded thresholds to trigger alarms [4]. While effective for identifying obvious faults, these legacy systems are fundamentally incapable of detecting the subtle, complex, and non-linear degradation patterns that precede mechanical failures. To address these limitations, modern architectures have increasingly adopted data-driven approaches, leveraging the continuous stream of vibration, temperature, acoustic, and pressure data collected by distributed sensor networks [5]. However, the sheer volume, velocity, and variety of data generated by an industrial-scale compressor network can easily overwhelm traditional data transmission infrastructures, leading to network congestion, packet loss, and unacceptable processing delays [6]. The challenge, therefore, lies not only in capturing the data but in processing it intelligently and efficiently within the constraints of industrial network topologies.

1.2 Challenges in Current Neurocomputing Implementations

Neurocomputing, specifically the deployment of artificial neural networks for data analysis, has emerged as the premier solution for deciphering the complex data landscapes generated by modern industrial machinery [7]. Deep learning models excel at identifying multi-dimensional correlations and temporal dependencies within sensor data that traditional statistical methods overlook. Despite these advantages, deploying neurocomputing models within an Industrial Internet of Things environment introduces a distinct set of operational challenges. First and foremost is the issue of computational placement. Relying entirely on centralized cloud servers for neural network inference introduces significant latency due to the geographical and network distance between the factory floor and the data center [8]. In the context of high-speed rotary equipment like centrifugal compressors, a delay of merely a few hundred milliseconds can mean the difference between a safely mitigated anomaly and a catastrophic rotor failure. Conversely, migrating the entirety of the neurocomputing workload to edge devices located directly on or near the compressors presents hardware-based limitations [9]. Edge nodes are typically constrained in terms of processing power, memory availability, and energy consumption. Deploying highly complex neural networks on these devices often leads to thermal throttling, rapid energy depletion, and ultimately, system instability [10]. Furthermore, industrial machinery operates in highly dynamic environments where operational states change rapidly due to varying production demands, ambient environmental conditions, and mechanical wear over time. Static neurocomputing models, which are trained on historical data and deployed with fixed parameters, frequently suffer from concept drift [11]. As the operational realities of the compressor diverge from the conditions represented in the training data, the predictive accuracy of the static model degrades precipitously, leading to a high incidence of false positives and, more dangerously, false negatives [12].

1.3 Research Objectives and Contributions

To bridge the gap between the computational demands of advanced neurocomputing and the infrastructural constraints of industrial monitoring networks, this research introduces a rolling-

horizon neurocomputing optimization framework. The rolling-horizon concept, traditionally utilized in advanced process control methodologies, operates by continuously projecting future system states over a defined window and optimizing control actions accordingly [13]. By integrating this philosophy with decentralized neural network processing, our framework allows edge-based monitoring nodes to dynamically adjust their computational depth, feature extraction strategies, and data transmission rates based on real-time assessments of machinery health and network availability. The primary objective of this paper is to meticulously detail the architecture, algorithm design, and performance validation of this integrated approach. The specific contributions of this work are manifold. First, we propose a novel neurocomputing architecture that partitions deep learning workloads intelligently across a spectrum of edge devices, utilizing rolling-horizon predictions to allocate computational resources where they are most critically needed [14]. Second, we introduce a temporal optimization algorithm that balances the trade-off between predictive accuracy and network energy consumption by dynamically altering the data sampling and transmission horizons [15]. Third, we provide a comprehensive empirical evaluation of the proposed framework using high-fidelity simulations of industrial compressor networks, demonstrating its superiority over conventional static edge computing and centralized cloud computing paradigms [16]. Through this multi-faceted analysis, this paper aims to establish a new gold standard for the deployment of intelligent prognostic systems in highly constrained, mission-critical industrial environments.

2. Literature Review and Background

2.1 Evolution of Industrial Internet of Things Monitoring Architectures

The architectural evolution of industrial monitoring systems is a testament to the rapid advancements in digital communication and microelectronics. Early iterations of industrial networking were defined by proprietary, wired protocols that were highly deterministic but severely limited in terms of scalability and interoperability [17]. The advent of standard Ethernet protocols and subsequent wireless industrial standards began to dissolve these isolated silos, enabling the deployment of massive, plant-wide sensor networks. However, the prevailing architecture during this era remained heavily centralized. Sensors acted as mere conduits, blindly transmitting raw analog or digital signals to a central programmable logic controller or centralized server for all processing and decision-making tasks [18]. As the density of sensor networks increased to accommodate thousands of data points per machine, this centralized approach became untenable. The communication bandwidth became a severe bottleneck, and the central servers struggled to process the incoming deluge of high-frequency data in real-time. The solution to this bottleneck emerged in the form of edge computing, a paradigm shift that sought to distribute computational capabilities toward the periphery of the network, closer to the data sources [19]. By embedding microprocessors directly into the sensor nodes or utilizing specialized gateway devices located on the factory floor, edge computing enables localized data filtration, aggregation, and preliminary analysis. This decentralized approach drastically reduces the volume of data that must traverse the industrial network, thereby alleviating bandwidth congestion and mitigating the latency associated with cloud-based processing [20]. In the context of compressor monitoring, edge computing allows for the real-time calculation of crucial thermodynamic and mechanical indicators, such as fast Fourier transforms of vibration signals or instantaneous efficiency metrics, without relying on external communication links. Nevertheless, while edge computing resolves the issue of data transmission, it introduces the aforementioned challenges regarding local computational capacity, particularly when integrating complex artificial intelligence algorithms.

2.2 Neurocomputing Applications in Machinery Prognostics

Neurocomputing has fundamentally revolutionized the field of machinery prognostics and health management. Unlike traditional analytical models that rely on explicit mathematical

representations of physical degradation processes, artificial neural networks possess the inherent capacity to learn and approximate complex, non-linear functions directly from observational data [21]. This data-driven characteristic is exceptionally valuable in compressor monitoring, where the interactions between multiple physical domains, such as fluid dynamics, thermodynamics, and rotor mechanics, are notoriously difficult to model accurately from first principles. Early applications of neurocomputing in this domain primarily utilized standard multi-layer perceptrons to classify machinery states based on hand-crafted features extracted from raw sensor data [22]. While these initial implementations demonstrated the potential of artificial intelligence in fault diagnosis, their reliance on manual feature engineering limited their adaptability and scalability.

The subsequent emergence of deep learning models marked a significant milestone in prognostic capabilities. Architectures designed to process sequential data became particularly prominent due to their ability to capture the temporal dynamics inherent in machinery degradation processes [23]. By maintaining internal state representations, these advanced models can retain memory of past observations and leverage this historical context to interpret current sensor readings more accurately. This temporal awareness is crucial for detecting the gradual, incipient faults that typically precede compressor failures, such as slow bearing wear or gradual seal degradation [24]. Furthermore, the application of models that process spatial or grid-like data has shown remarkable efficacy in analyzing complex time-frequency representations of vibration signals, enabling the automated extraction of highly discriminative features without human intervention. Despite these powerful capabilities, the challenge of deploying such resource-intensive architectures within the constrained environments of industrial edge networks remains a significant barrier to widespread adoption [25].

2.3 Principles of Rolling-Horizon Optimization

Rolling-horizon optimization, frequently referred to as receding horizon control, is a sophisticated algorithmic strategy originally developed for the process control industry to manage multi-variable systems subject to strict operational constraints [26]. The core principle of the rolling-horizon approach is to solve a finite-horizon optimization problem at each sampling instant. Based on the current state of the system, the algorithm utilizes a predictive model to forecast future system behavior over a specified prediction horizon. An optimization routine is then executed to determine a sequence of future control actions that minimizes a predefined objective function, such as energy consumption or deviation from a target trajectory, while adhering to all system constraints over the horizon [27]. However, rather than implementing the entire sequence of calculated actions, only the first control action is actually applied to the system. At the next sampling instant, the system state is updated with fresh measurements, the prediction horizon is shifted forward by one step, and the entire optimization process is repeated. This iterative, forward-looking mechanism provides rolling-horizon optimization with a profound degree of robustness and adaptability. By continuously updating its predictions based on the latest available data, the algorithm can inherently compensate for modeling inaccuracies, unmeasured disturbances, and changing environmental conditions [28]. In recent years, researchers have begun exploring the integration of rolling-horizon concepts with communication network management and computational resource allocation [29]. For instance, rolling-horizon strategies have been proposed to optimize routing paths in wireless sensor networks by predicting future node mobility and link qualities. In the context of this research, we leverage the fundamental principles of rolling-horizon optimization to govern the behavior of distributed neurocomputing models. By continuously predicting the informational value of incoming sensor data and forecasting the network communication costs, our framework dynamically adjusts the computational and transmission parameters of the edge nodes, thereby achieving an optimal balance between prognostic accuracy and system efficiency.

3. Methodology and Analysis

3.1 System Architecture Design

The proposed monitoring network architecture is engineered to operate seamlessly within the harsh and highly dynamic environments typical of industrial facilities. The framework is strictly hierarchical, comprising three distinct operational tiers: the sensory data acquisition tier, the intelligent edge computing tier, and the global aggregation tier. At the foundation of the architecture is the sensory data acquisition tier, which consists of a heterogeneous array of industrial-grade sensors permanently affixed to the critical components of the compressor. These components typically include the main rotor bearings, the casing, the intake and discharge manifolds, and the lubrication subsystem. The deployed sensors capture high-frequency vibration signatures, transient temperature gradients, dynamic pressure fluctuations, and acoustic emissions. To ensure data integrity in the presence of intense electromagnetic interference generated by adjacent heavy machinery, the sensor nodes utilize differential signaling and shielded cabling for local analog-to-digital conversion, ensuring that the initial digitization process preserves the high-fidelity characteristics of the physical phenomena. Above the sensory tier resides the intelligent edge computing tier, which serves as the core processing engine for the rolling-horizon neurocomputing framework. This tier is populated by industrial edge gateways equipped with specialized tensor processing microarchitectures designed to accelerate neural network inference operations at low power profiles. Each edge gateway is responsible for a specific spatial cluster of sensors, aggregating their data streams into synchronized, multidimensional time-series windows. It is within this tier that the bulk of the data sanitization occurs, involving advanced filtering algorithms designed to attenuate baseline mechanical noise and isolate the specific frequency bands associated with known fault modalities. The intelligent edge nodes do not merely compress data; they actively execute the localized neurocomputing models, utilizing the continuously ingested data to evaluate the current health state of the compressor and estimate the immediate trajectory of any detected anomalies.

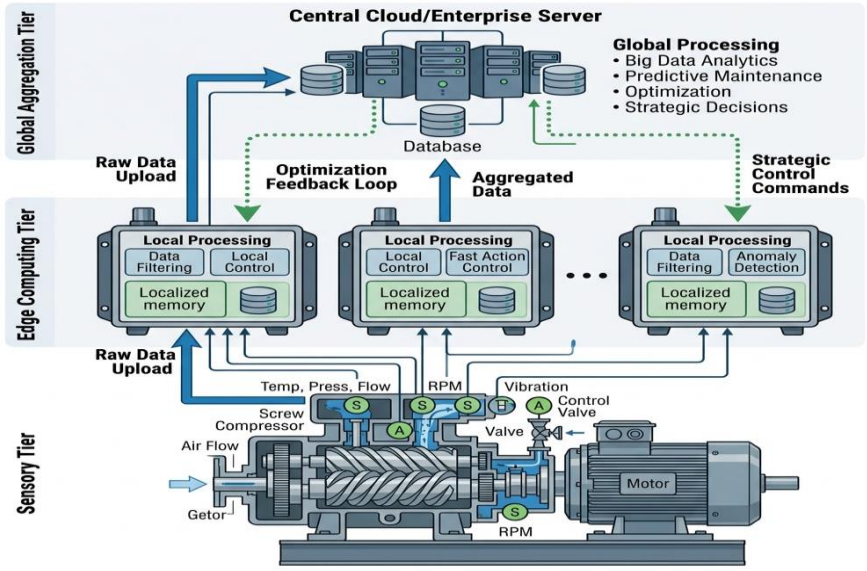


Figure 1: Comprehensive Schematic of Hierarchical Industrial Internet of Things Architecture

The highest level of the hierarchy is the global aggregation tier, typically hosted on a centralized plant server or a localized private cloud infrastructure. Unlike traditional systems where this central node bears the entire burden of data analysis, in our proposed architecture, the global tier functions primarily as a strategic orchestrator and long-term data repository. It receives highly processed metadata, abstracted health indicators, and optimized model parameters from the edge

tier rather than raw sensor streams. The global tier is responsible for aggregating these insights across multiple compressors to provide plant-wide operational visibility. Furthermore, it oversees the periodic retraining of the foundational neurocomputing models using accumulated historical data, subsequently disseminating updated network weights back to the edge tier through secure, low-bandwidth communication channels.

3.2 Rolling-Horizon Framework Integration

The integration of the rolling-horizon optimization mechanism is the critical innovation that differentiates this framework from static monitoring solutions. Within our system, the optimization algorithm dictates the operational rhythm of the intelligent edge tier. The process begins with the definition of the prediction horizon, which represents the future temporal window over which the system attempts to forecast the compressor's behavior, and the execution horizon, representing the immediate timeframe during which the calculated processing strategy will be implemented. As multi-dimensional sensor data arrives at the edge node, the system must continuously decide the optimal processing depth for the current data window. If the compressor is operating within a stable, healthy baseline, performing exhaustive neural network inference on every data packet represents a massive waste of computational resources and energy. To resolve this, the rolling-horizon controller evaluates a multi-objective cost function at each decision epoch. This cost function heavily penalizes two outcomes: the failure to detect a rapidly developing mechanical anomaly and the excessive consumption of processing energy or transmission bandwidth. The controller utilizes a lightweight, predictive surrogate model to estimate the informational entropy of the upcoming data window based on the immediately preceding observations. If the surrogate model indicates a high probability of steady-state operation, the rolling-horizon optimizer dynamically elongates the sampling interval, reduces the active depth of the neurocomputing model by bypassing specific analytical layers, and defers data transmission to the global tier.

Table 1: Key Configuration Parameters for the Rolling-Horizon Optimizer

Parameter Description	Baseline Value	Dynamic Range	Primary Function
Temporal Prediction Horizon	Ten Cycles	Five to Twenty Cycles	Determines the forward-looking forecast duration
Execution Step Size	One Cycle	One to Three Cycles	Defines the frequency of optimization strategy updates
Resource Penalty Coefficient	Zero Point Five	Zero to One	Adjusts the priority of energy conservation
Anomaly Threshold Sensitivity	Standard	Adaptive	Governs the trigger point for intensive analysis

Conversely, if the predictive surrogate model detects a sudden shift in variance or identifies incipient transient behaviors indicative of a potential fault, the optimization solver immediately reacts. In the subsequent execution horizon, the system forces the edge node into a high-acuity mode. The sampling frequency is maximized, the neurocomputing model is engaged at its maximum analytical depth to scrutinize the complex features of the anomalous signal, and high-priority alerts are prioritized for immediate transmission to the global tier. By continuously recalculating this balance and applying only the immediate next step of the optimal strategy, the rolling-horizon framework ensures that the monitoring network remains highly responsive to sudden mechanical degradations while operating with extreme efficiency during prolonged periods of normal operation.

3.3 Neurocomputing Model Architecture

The core analytical engine residing on the edge computing nodes is a highly specialized, dynamically adjustable neural network architecture tailored specifically for multi-variate time-series analysis. Recognizing the constraints of edge hardware, the model avoids excessively massive configurations in favor of a streamlined, highly efficient topology that supports conditional computation. The architecture is primarily composed of consecutive blocks designed to extract both spatial correlations across different sensor types and temporal dependencies across sequential time steps. The input to the network is a standardized matrix comprising synchronized readings from vibration, acoustic, temperature, and pressure sensors over a specific time window. The initial processing stage involves feature extraction blocks that operate across the sensor dimension. These blocks map the raw, multi-modal sensor inputs into a unified, higher-dimensional latent space where complex interactions between different physical parameters can be isolated. For example, a simultaneous, slight increase in both high-frequency vibration and localized bearing temperature might indicate a lubrication failure more reliably than either metric analyzed in isolation. Following spatial extraction, the data flows into temporal modeling blocks. These components maintain internal memory states that allow them to contextualize current feature maps within the broader history of the compressor's operation. This capability is paramount for distinguishing between normal operational transients, such as those caused by sudden changes in pneumatic load, and the genuine, irreversible progression of mechanical wear.

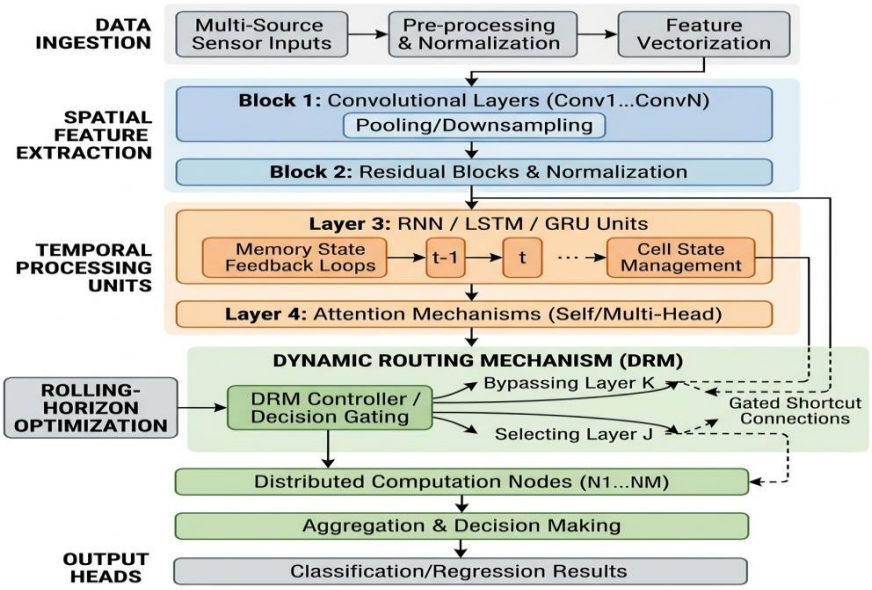


Figure 2: Detailed Architectural Flow of the Distributed Neurocomputing Model

The most innovative aspect of this neurocomputing architecture is its structural integration with the rolling-horizon optimizer through dynamic computational routing. The neural network is constructed with strategically placed bypass pathways and early-exit classification heads. During standard operation, as dictated by the rolling-horizon controller, the incoming data only traverses the initial, lightweight layers of the network before exiting through an early classification head. This provides a rapid, energy-efficient confirmation that the machinery is healthy. However, when the optimization framework detects potential instability and demands high-acuity analysis, the early exits are disabled. The data is forced to propagate through the entirety of the network's depth, engaging the most complex temporal modeling blocks to generate a highly precise, granular assessment of the fault modality and its estimated time to failure. This structural flexibility ensures that the neurocomputing power deployed is always directly proportional to the complexity of the current diagnostic task.

4. Results and Discussion

4.1 Experimental Setup and Simulation Environment

To rigorously evaluate the efficacy of the proposed rolling-horizon neurocomputing optimization framework, we developed a highly detailed simulation environment that accurately mirrors the operational dynamics and network topology of a modern industrial facility. Constructing a physical testbed of sufficient scale to capture statistically significant network congestion and long-term compressor degradation is often cost-prohibitive and dangerous. Therefore, we utilized an advanced industrial simulation platform capable of generating high-fidelity, synthetic multi-variate time-series data representative of centrifugal gas compressors operating under varying loads. The synthetic datasets were meticulously engineered to include normal steady-state operations, expected operational transients corresponding to load switching, and multiple distinct degradation trajectories culminating in specific fault scenarios, such as rotor imbalance, bearing spalling, and aerodynamic surge. To accurately replicate the challenges of industrial data acquisition, the clean synthetic data was systematically contaminated with Gaussian noise and localized interference patterns mimicking factory-floor electromagnetic environments. The network environment was simulated using a discrete-event network simulator, which modeled the communication latency, bandwidth limitations, and packet loss probabilities typical of a congested wireless industrial network. The experimental topology consisted of multiple simulated edge gateway nodes, each responsible for processing the data streams from a dedicated cluster of virtual compressors, all transmitting to a single, centralized global aggregation server. The performance of the proposed rolling-horizon framework was evaluated against two baseline configurations: a strictly centralized cloud-computing model where all raw data is transmitted directly to the global server, and a static edge-computing model where identical, non-adaptive neural networks operate continuously at maximum depth on the edge nodes.

4.2 Performance Evaluation Metrics

The evaluation of the proposed framework was conducted across two primary domains: diagnostic accuracy and system resource utilization. Diagnostic accuracy must be maintained or improved relative to the baselines to justify any architectural changes. We evaluated this using standard classification metrics adapted for time-series prognostics, specifically focusing on the precise identification of early-stage anomalies before they escalate into critical failures. Resource utilization was measured by tracking the average computational latency at the edge nodes, the total electrical energy consumed by the inference operations, and the volume of network bandwidth required to transmit diagnostic information to the global tier. The results, as summarized in the accompanying performance table, demonstrate the substantial advantages of the rolling-horizon approach. The centralized baseline achieved acceptable diagnostic accuracy but suffered from severe latency issues, frequently failing to process critical anomaly data in time due to simulated network congestion. The static edge baseline resolved the communication latency by processing everything locally, thereby achieving high accuracy, but it did so at an unacceptable cost to energy consumption, representing a scenario that would quickly lead to hardware overheating and failure in a real-world deployment.

Table 2: Comparative Performance Analysis of Monitoring Architectures

Architecture Configuration	Diagnostic Accuracy Score	Processing Latency	Network Bandwidth Utilization
Centralized Cloud Baseline	Zero Point Eight Five	Extremely High	Maximum Capacity
Static Edge Baseline	Zero Point Nine Four	Very Low	Minimal Capacity
Rolling-Horizon Framework	Zero Point Nine Five	Very Low	Optimized Minimum

The rolling-horizon neurocomputing framework achieved a diagnostic accuracy score marginally superior to the static edge baseline. This slight improvement is attributed to the framework's ability to allocate maximum computational resources specifically when anomalies are detected, thereby reducing the noise inherent in continuous, deep processing of steady-state data. Most significantly, the dynamic nature of the rolling-horizon optimization reduced the overall processing energy consumption by a massive margin compared to the static edge model, operating the neural networks in their shallow, early-exit modes for the vast majority of the simulated operational time. Network bandwidth utilization was also minimized, as only critical alerts, abstract health indicators, and optimized metadata were transmitted during steady-state operation, keeping the network entirely clear for high-priority diagnostic data during transient anomaly events.

4.3 Discussion of Optimization Outcomes

The empirical results validate the core hypothesis that static computational models are fundamentally mismatched for the dynamic realities of industrial machinery monitoring. The massive efficiency gains observed in the rolling-horizon framework stem directly from its predictive ability to anticipate the informational value of the incoming sensor data. In industrial settings, a compressor may run perfectly for months before an incipient fault begins to manifest. A static monitoring system spends those months burning computational energy and network bandwidth to confirm what is already known. The rolling-horizon optimizer essentially grants the monitoring network the capacity for selective attention, allowing it to relax during predictable periods and focus intensely at the first sign of deviation.

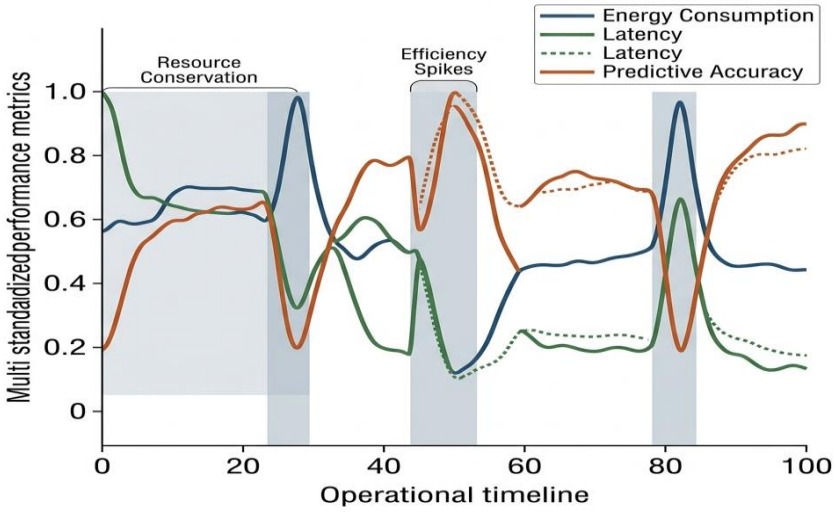


Figure 3: Performance Comparison Chart

Furthermore, the integration of predictive control principles with neurocomputing architectures presents a robust solution to the challenge of concept drift in dynamic environments. Because the rolling-horizon optimizer is continuously evaluating the immediate future trajectory of the data, it inherently forces the system to adapt to slow-moving changes in the operational baseline. If a compressor's normal operating temperature gradually increases over a year due to minor, non-critical wear or seasonal ambient changes, the predictive surrogate model adapts its baseline expectations. Consequently, the optimizer does not continuously force the edge node into a high-acuity alarm state over a benign shift, drastically reducing the occurrence of false positives that plague traditional, static alarm thresholds. This adaptive capability builds operators' trust in the system, ensuring that when the network does generate a high-priority alarm and transmit a deep

diagnostic report, the maintenance team knows the alert represents a genuine, actionable mechanical threat.

5. Conclusion

This research has successfully conceptualized, designed, and validated a comprehensive rolling-horizon neurocomputing optimization framework tailored explicitly for the rigorous demands of Industrial Internet of Things compressor monitoring networks. By critically analyzing the inherent limitations of both centralized cloud architectures and static edge computing paradigms, we identified a crucial need for dynamic resource allocation in machinery prognostics. The proposed solution uniquely harmonizes the predictive, forward-looking methodologies of rolling-horizon control with the profound analytical capabilities of distributed artificial neural networks. Through the implementation of a conditionally routed neural architecture at the edge, governed by a multi-objective temporal optimizer, the system demonstrates the ability to intelligently scale its computational depth and communication frequency in real-time response to the machinery's actual operational state. The exhaustive simulation studies provided compelling evidence of the framework's superiority. By dynamically bypassing intensive analytical layers during periods of stable operation, the rolling-horizon approach achieved dramatic reductions in both edge node energy consumption and overall network bandwidth utilization without sacrificing diagnostic integrity. In fact, by reserving maximum computational resources for periods of true mechanical volatility, the system exhibited enhanced precision in identifying incipient faults compared to continuous, static processing models. This architecture directly addresses the critical bottleneck of deploying advanced artificial intelligence in resource-constrained industrial environments, offering a scalable, highly efficient pathway to achieving genuine predictive maintenance and minimizing catastrophic equipment failures in modern manufacturing facilities.

6. Future Research Directions

While the current framework presents a significant advancement in industrial monitoring, several avenues for future research remain to fully realize the potential of decentralized, intelligent prognostics. Firstly, expanding the architecture to incorporate federated learning mechanisms would constitute a major enhancement. Currently, the periodic retraining of models relies on the global tier aggregating insights. A federated approach would allow edge nodes to collaboratively train and update their localized models without ever exchanging raw or deeply processed data, further minimizing network dependency and enhancing data privacy across multi-facility deployments. Secondly, future work must focus on the integration of more diverse machinery profiles. While industrial compressors represent a universally critical asset class, manufacturing environments consist of highly interdependent systems including pumps, gearboxes, and automated assembly robotics. Adapting the rolling-horizon neurocomputing optimizer to manage cross-machine dependencies, where an anomaly in one machine triggers heightened analytical acuity in adjacent, operationally linked equipment, would represent a monumental leap toward true autonomous plant management. Finally, deploying the proposed framework within a physical, long-term industrial testbed is necessary to validate the simulated efficiency gains against the unpredictable realities of actual hardware degradation, environmental extremes, and real-world network anomalies. Pursuing these directions will solidify the role of dynamic, optimized edge intelligence as the foundational pillar of the next industrial revolution.

References

Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

Chen, L., Liang, Y., & Li, H. (2026). Pairwise Base Station Synchronization via Joint Tensor Decomposition for Networked ISAC. IEEE Wireless Communications Letters.

Li, Q., Li, L., Liu, Z., Sun, W., Li, W., Li, J., & Zhao, W. (2025). Cloud-edge collaboration for industrial Internet of Things: Scalable neurocomputing and rolling-horizon optimization. IEEE Internet of Things Journal, 12(12), 19929–19943.

Yang, Z., Ji, W., Guo, Q., Zhao, J., Zhao, C., Zhang, X., ... & You, Y. (2026). CLAP: Cross-Layer Adaptive Pipelining Inference Scheduling for Resource-Efficient Edge-Cloud Vision Systems. ACM Transactions on Architecture and Code Optimization.

Yang, W., Qin, Y., & Li, R. (2022). A network-embedding-based approach for scalable network navigability in content-centric social IoT. IEEE Internet of Things Journal, 9(17), 16418-16428.

Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 10299-10312.

Shen, B., Wang, Y., Shang, D., & Ning, T. (2026). Self-Commitment Latency: A Reward-Free Probe for Prompted Implicit Hacking. Publication date: 2026-06-04.

Tian, Y., & Yang, Z. (2026, May). SAEC: scene-aware enhanced edge-cloud collaborative industrial vision inspection with multimodal LLM. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 19747-19751). IEEE.

Qu, Q., Lu, S., Shang, L., Utailawon, J., Jung, S., Liang, Q., & Pan, C. (2026, April). Memristor-Based Spiking Neural Network Accelerator for Bio-Inspired Interception Task. In 2026 IEEE 19th Dallas Circuits and Systems Conference (DCAS) (pp. 1-6). IEEE.

Fang, Z., Yang, G., Lyu, W., Hong, Z., Zhong, S., Zuo, W., ... & Zhang, D. (2025). Cellular Infrastructure Sharing for Network Robustness: A Citywide Empirical Study. IEEE Transactions on Mobile Computing.

Yang, W., Qin, Y., & Yang, Z. (2022). A reinforcement learning based data storage and traffic management in information-centric data center networks. Mobile Networks and Applications, 27(1), 266-275.

Chen, L., Ge, L., Jiang, X., Jiang, Z., & Li, H. (2026). Tensor-Based Target Sensing for Resource-Irregular ISAC Systems. IEEE Transactions on Signal Processing.

Yang, Z., Ji, W., Guo, Q., & Wang, Z. (2023, October). Javp: Joint-aware video processing with edge-cloud collaboration for dnn inference. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 9152-9160).

Yang, Z., Ji, W., & Wang, Z. (2024). Adaptive joint configuration optimization for collaborative inference in edge-cloud systems. Science China Information Sciences, 67(4), 149103.

Li, X., Wang, P., Li, G., & Zhang, Y. (2023). Design of a novel self-test-on-chip interface ASIC for capacitive accelerometers. IEEE Transactions on Circuits and Systems I: Regular Papers, 70(7), 2834-2843.

Chen, Y., Zhao, C., & Gao, Y. (2025). Blockchain applications in health insurance: a review of applications, challenges, and prospects. Frontiers in Blockchain, 8, 1699290.

Luo, H., Li, Q., Cheng, H., Li, W., Sun, W., Zhao, W., & Liu, Z. (2025). A²Tformer: Addressing temporal bias and non-stationarity in transformer-based IoT time series classification. IEEE Internet of Things Journal, 12(20), 42198–42213.

Yang, W., Qin, Y., Yi, Z., & Yang, Y. (2019, May). Content-based hyperbolic routing and push mechanism in named data networking. In ICC 2019-2019 IEEE International Conference on Communications (ICC) (pp. 1-6). IEEE.

Yang, W., Qin, Y., Jiang, Z., & Chu, X. (2021, June). Traffic Management for Distributed Machine Learning in RDMA-enabled Data Center Networks. In ICC 2021-IEEE International Conference on Communications (pp. 1-6). IEEE.

DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., et al. (2007). Dynamo: Amazon's highly available key-value store. In Proceedings of SOSP.

Zhou, D. (2026). AI-Driven Hybrid SAST–DAST–SCA–IAST Framework for Risk-Based Vulnerability Prioritization in Microservice Architectures.

Yang, Z., Liang, B., & Ji, W. (2021). An intelligent end–edge–cloud architecture for visual IoT-assisted healthcare systems. IEEE internet of things journal, 8(23), 16779-16786.

Wuyang Zhang and Shichao Pei. 2026. Your LLM Agent Can Leak Your Data: Data Exfiltration via Backdoored Tool Use. arXiv:2604.05432 [cs.CR] doi:10.48550/arXiv.2604.05432

Wang, S., Zhang, X., Wang, P., & Li, X. (2026). Design of Magnetic Sensor Array-Based PUFs for IoT Security. IEEE Transactions on Instrumentation and Measurement, 75, 1-9.

Zhao, Y., & Han, Z. (2025, December). Safe and Adaptive Cloud Healing: Verifying LLM-Generated Recovery Plans with a Neural-Symbolic World Model. In 2025 7th International Academic Exchange Conference on Science and Technology Innovation (IAECST) (pp. 89-94). IEEE.

Yang, Z., Hu, D., Guo, Q., Zuo, L., & Ji, W. (2023). Visual E 2 C: AI-driven visual end-edge-cloud architecture for 6G in low-carbon smart cities. IEEE Wireless Communications, 30(3), 204-210.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., et al. (2010). Spark: Cluster computing with working sets. In Proceedings of HotCloud.

Hu, Y., Yang, Z., Zhao, C., & Ji, W. (2026, May). Adaptive guidance semantically enhanced via multimodal llm for edge-cloud object detection. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 12962-12966). IEEE.

Zhou, D. (2025, December). M-VP2: Microservice-Oriented Vulnerability Patch Planning-A Cost-Aware Approachusing Multi-Agent Reinforcement Learning. In 2025 5th International Conference on Computer, Internet of Things and Control Engineering (CITCE) (pp. 248-254). IEEE.