Article

Federated Learning-Based Edge-Cloud Collaborative Framework for Real-Time PMSM Fault Diagnosis

Xinran Fang[1,*], Zerui Liang[1] and Qi Shao[2]

[1] Shanghai DianJi University, Shanghai, China
[2] Anhui Science and Technology University, Bengbu, China
* Correspondence: Xinran Fang, Shanghai DianJi University, Shanghai, China

Abstract: Permanent magnet synchronous motors are critical components in modern industrial automation, electric transportation, and renewable energy systems, where operational reliability directly impacts safety, efficiency, and downtime costs. Traditional fault diagnosis approaches rely on centralized data collection and cloud-based analytics, which introduce latency, bandwidth bottlenecks, and privacy vulnerabilities-especially when deployed across geographically distributed edge devices. This paper presents a federated learning-based edge-cloud collaborative framework specifically engineered for real-time fault diagnosis of permanent magnet synchronous motors. The architecture decouples model training from raw data movement: local edge nodes perform on-device feature extraction and lightweight model updates using localized motor current, vibration, and thermal signatures, while a central cloud orchestrator aggregates and refines global model parameters without accessing raw sensor streams. A time-synchronized gradient averaging protocol ensures convergence stability under heterogeneous device capabilities, intermittent connectivity, and non-IID fault data distributions. Real-world validation was conducted across twelve industrial motor testbeds operating under variable load, speed, and ambient conditions. Results demonstrate sub-second diagnostic latency, 98.7% average classification accuracy across seven fault classes-including demagnetization, inter-turn short circuits, and bearing degradation-and a 73% reduction in uplink data volume compared to conventional cloud-only pipelines. The framework maintains robustness during network partitions and adapts continuously as new fault patterns emerge at the edge. By harmonizing computational responsibility between resource-constrained endpoints and scalable cloud infrastructure, this work advances a practical, privacy-aware, and responsive paradigm for intelligent motor health monitoring in dynamic industrial ecosystems.

Keywords: Federated learning; PMSM; Fault diagnosis; Edge computing; Cloud collaboration

Received: 13 July 2026
Revised: 05 September 2026
Accepted: 16 September 2026
Published: 23 September 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Motivation and Industrial Imperative

Permanent magnet synchronous motors (PMSMs) are increasingly deployed in safety-critical and high-availability industrial systems, where operational continuity and fault resilience are non-negotiable. Undetected or delayed diagnosis of incipient faults-such as demagnetization, inter-turn short circuits, and bearing degradation-can trigger cascading system failures, unplanned production outages, and substantial maintenance overhead [1]. Traditional centralized diagnostic paradigms, reliant on aggregating raw sensor data to a remote cloud server, confront three fundamental constraints: unacceptable end-to-end latency for real-time intervention; violation of data sovereignty requirements imposed by regulatory frameworks and enterprise policies; and prohibitive bandwidth consumption when transmitting high-frequency motor current and vibration signals across wide-area networks. These limitations render legacy architectures unsuitable for time-sensitive edge environments with strict privacy, responsiveness, and resource budgets. Consequently, there exists an urgent industrial imperative to rearchitect

fault diagnosis infrastructure-not as a monolithic cloud service, but as a tightly coordinated, privacy-aware, and latency-bounded edge-cloud continuum [1, 2]. This necessitates a paradigm shift toward distributed intelligence that preserves local data fidelity while enabling collaborative model evolution across heterogeneous deployment tiers [3].

1.2. Core Technical Challenge

The core technical challenge lies in reconciling stringent real-time inference requirements with the computational and data constraints inherent to edge-cloud collaborative PMSM fault diagnosis [4]. Deep learning models demand substantial training data, high-dimensional parameter spaces, and intensive computation-resources fundamentally incompatible with resource-constrained edge devices characterized by limited memory, low-power processors, and intermittent network connectivity. Conventional transfer learning fails to address non-IID data distributions across heterogeneous motor units, while model compression techniques sacrifice diagnostic fidelity critical for early-stage fault detection [4, 5]. Moreover, regulatory and privacy mandates prohibit transmission of raw high-frequency sensor streams to centralized clouds, exacerbating data fragmentation and distribution skew. This creates a tripartite tension: latency-sensitive edge inference necessitates lightweight models, yet diagnostic robustness requires expressive architectures trained on diverse operational conditions; distributed data silos impede global model convergence; and strict data governance precludes centralized aggregation. Consequently, no existing single-method approach resolves the coupled constraints of real-time responsiveness, statistical heterogeneity, and regulatory compliance. A novel federated learning-based orchestration framework is therefore essential-one that enables collaborative model evolution without raw data exchange, preserves edge autonomy, and maintains sub-millisecond inference latency under dynamic load and fault conditions.

1.3. Proposed Framework Overview

This paper introduces a federated edge-cloud collaborative framework specifically engineered for real-time permanent magnet synchronous motor (PMSM) fault diagnosis. The architecture implements a three-tiered coordination logic: edge nodes execute local model training and low-latency inference, an orchestration layer manages asynchronous model aggregation and version control, and the cloud tier hosts global model evolution and cross-domain knowledge distillation [1]. A decentralized training loop ensures that raw sensor data never leaves the edge device, while encrypted model updates are exchanged under differential privacy constraints. A closed-loop feedback mechanism continuously refines diagnostic models using operational telemetry, enabling autonomous adaptation to evolving fault signatures and environmental perturbations. The framework is explicitly designed to achieve millisecond-level inference latency at the edge, maintain diagnostic fidelity across heterogeneous operating regimes-including transient load changes, thermal drift, and partial demagnetization-and guarantee operational resilience amid infrastructure volatility such as intermittent connectivity, node churn, or resource contention. Model convergence stability is enforced through adaptive learning rate scheduling and gradient clipping, with integrity verified via cryptographic model hashing. This structural paradigm shifts PMSM health monitoring from static, centralized analytics toward dynamic, privacy-preserving, and infrastructure-agnostic intelligence [6].

2. Literature Review

2.1. Evolution of Motor Fault Diagnosis Techniques

Motor fault diagnosis has evolved through distinct methodological paradigms. Early approaches relied on rule-based expert systems encoding domain knowledge and heuristic decision trees, constrained by limited adaptability and maintenance overhead. Subsequent advances leveraged statistical signal processing-encompassing time-domain

features, spectral analysis, and wavelet transforms-to extract discriminative patterns from vibration and current signatures. These methods improved robustness but required careful feature engineering and remained sensitive to operational variability. The deep learning era introduced end-to-end trainable architectures, notably convolutional and recurrent neural networks, which automatically learn hierarchical representations from raw sensor data [7]. However, this shift intensified dependence on large-scale, high-fidelity labeled datasets and substantial computational resources, often centralized in cloud environments [2]. Despite performance gains, persistent challenges include poor generalizability across diverse permanent magnet synchronous motor designs, variable load profiles, and heterogeneous sensor modalities or placements. Current frameworks struggle to maintain diagnostic accuracy when deployed across unseen motor configurations or under dynamic operating conditions without extensive retraining. This limitation underscores the need for adaptive, infrastructure-aware methodologies capable of preserving model fidelity while accommodating real-world deployment constraints at the edge-cloud interface.

2.2. Edge Intelligence in Industrial Monitoring

Edge intelligence has emerged as a pivotal enabler for real-time industrial monitoring, particularly in resource-constrained environments where latency, thermal dissipation, and power budgets impose strict operational boundaries. Embedded inference strategies increasingly favor lightweight convolutional neural networks optimized for microcontroller units and programmable logic controllers, balancing representational fidelity against computational throughput [1]. Quantized recurrent architectures further reduce memory footprint and inference latency while preserving temporal modeling capacity for time-series diagnostics [5]. On-chip anomaly detection mechanisms-often leveraging statistical divergence metrics or unsupervised reconstruction residuals-enable autonomous fault flagging without cloud round-trips [5]. However, these approaches confront inherent trade-offs: aggressive model compression risks degrading sensitivity to incipient faults, while hardware-aware optimizations must reconcile clock frequency throttling, voltage scaling, and thermal derating across heterogeneous edge nodes. Empirical evaluations indicate that inference latency below $\tau_{\max} = 50\ \text{ms}$ remains critical for closed-loop PMSM protection, yet achieving such bounds under sustained thermal load necessitates co-design of algorithmic sparsity, mixed-precision arithmetic, and adaptive sampling policies. Consequently, the frontier lies not in isolated model reduction, but in holistic edge-cloud orchestration wherein local models perform rapid triage while delegating ambiguous cases to higher-fidelity cloud analytics. Beyond designing for low latency, the ability to predict inference delay before deployment is equally important for reliable edge operation. Prior work employed machine learning algorithms to predict inference delay through feature engineering and regression modeling, enabling pre-deployment performance estimation. Although developed for a different inference scenario, this "delay prediction for deployment planning" principle is methodologically relevant to the deployment optimization of edge-based PMSM fault diagnosis models, where runtime latency must be bounded under varying computational loads and thermal conditions [6].

2.3. Federated Paradigms Beyond Healthcare and Finance

Early federated learning applications extended beyond healthcare and finance into distributed cyber-physical systems, where heterogeneous edge devices generate high-frequency time-series sensor data under stringent latency constraints [3]. These deployments necessitated structural adaptations to accommodate non-IID temporal distributions, asynchronous device availability, and hardware-induced sampling drift. Critical modifications included the introduction of sliding-window gradient accumulation to preserve temporal coherence during local model updates, dynamic alignment of gradient timestamps across devices to mitigate phase misalignment in motor current signatures, and fault-aware aggregation weights that scale inversely with per-class support in local datasets. Such weighting schemes explicitly address severe class

imbalance-particularly for incipient fault modes-by modulating contribution magnitudes based on $n_{\text{fault}}/n_{\text{total}}$ ratios observed at each node. Furthermore, update windows were decoupled from global round boundaries to enable real-time responsiveness, while temporal interpolation protocols ensured consistent feature vectorization despite variable sampling intervals [7]. These innovations collectively reconfigure federated optimization to prioritize temporal fidelity and diagnostic sensitivity over statistical convergence guarantees, establishing foundational principles for edge-cloud collaborative fault diagnosis in permanent magnet synchronous motor systems.

3. Materials and Methods

3.1. System Architecture and Data Flow

The proposed framework implements a three-tier edge-cloud collaborative architecture for real-time permanent magnet synchronous motor (PMSM) fault diagnosis. At the sensor layer, current transducers, triaxial accelerometers, and infrared thermal arrays continuously acquire multimodal operational data with synchronized timestamps. These sensors interface directly with industrial gateways at the edge layer, which feature ARM Cortex-A72 processors and run a real-time Linux kernel to ensure deterministic latency bounds for local preprocessing, feature extraction, and on-device inference. The cloud layer hosts a containerized orchestration platform that manages federated learning workflows, global model aggregation, and long-term analytics. A secure handshake protocol establishes authenticated channels between edge nodes and the cloud prior to any data exchange. All model parameter updates are encrypted using symmetric-key cryptography before transmission, and synchronization occurs at fixed heartbeat-driven intervals governed by Δt_{sync} , where Δt_{sync} is dynamically adjusted based on network stability metrics and computational load variance across participating nodes [1]. Data flow follows a strict pull-based paradigm: edge nodes request updated global models only upon successful local training completion and integrity verification, while raw sensor measurements remain strictly confined to the edge. This design enforces data sovereignty, minimizes upstream bandwidth consumption, and guarantees temporal coherence across distributed diagnostic agents without compromising real-time responsiveness or cryptographic assurance [4].

3.2. Local Feature Engineering Pipeline

Each edge node executes a deterministic, hardware-optimized feature engineering pipeline to ensure reproducibility and fixed-point arithmetic compatibility. Raw current and voltage signals undergo adaptive noise filtering using a recursive least-squares algorithm with time-varying forgetting factor λ_t , calibrated per motor load condition. Subsequent spectral decomposition employs a sliding window of length $N_w = 1024$ samples with 50% overlap, preserving temporal locality while accommodating non-stationary fault transients. Time-frequency representation is generated via synchrosqueezed continuous wavelet transform (CWT), where the mother wavelet is the complex Morlet function $\psi_{\text{morlet}}(\eta) = \pi^{-1/4} e^{i\omega_0 \eta} e^{-\eta^2/2}$, followed by instantaneous frequency reassignment to sharpen energy concentration along fault-related ridges. Resulting time-frequency scalograms are normalized to unit L_2 norm and downsampled to 64×64 resolution to constrain memory footprint. Dimensionality reduction applies supervised locality-preserving projections (SLPP), optimizing the projection matrix $\boldsymbol{W}_{\text{SLPP}} \in \mathbb{R}^{4096 \times d}$ to maximize inter-class separability while preserving local manifold structure-where $d = 128$ balances discriminative power and edge inference latency [3]. All operations are implemented in Q15 fixed-point arithmetic with saturation arithmetic and range-aware quantization thresholds derived from empirical signal statistics across operational regimes. This pipeline ensures consistent feature fidelity across heterogeneous edge hardware while maintaining strict determinism for federated model convergence.

3.3. Federated Training Protocol

The federated training protocol implements a robust, iterative cycle to ensure convergence and resilience in heterogeneous edge environments [3]. Each round begins with global model initialization, followed by local training on recent fault segments collected at each edge device. Local updates employ mini-batch stochastic gradient descent with adaptive learning rates tuned per device based on signal-to-noise ratio and fault class imbalance. Gradient clipping enforces ℓ_2 -norm bounds to mitigate explosion, while top- k sparsification reduces communication overhead without compromising diagnostic fidelity. Aggregation weights are computed as the product of local data volume and label diversity entropy, ensuring underrepresented fault modes receive proportional influence [5]. A versioned global model repository maintains immutable snapshots with cryptographic hashes; rollback is triggered automatically upon detection of persistent divergence exceeding τ_{div} or anomalous gradient norms inconsistent with historical baselines. Client divergence is monitored via cosine similarity between local gradients and the current global gradient direction, augmented by statistical outlier detection on per-class confusion matrix shifts. Compromised clients-identified through Byzantine behavior signatures including gradient inversion, repeated zero-updates, or temporal correlation anomalies-are quarantined and excluded from aggregation for subsequent rounds. All metadata, including client health scores, update timestamps, and version lineage, is stored in a tamper-evident ledger. This design ensures real-time adaptability, fault tolerance, and auditability across distributed PMSM monitoring infrastructure. The coordination logic among federated nodes also resonates with structured multi-agent collaboration, where role separation and ordered interaction protocols improve task reliability. Zhang et al. proposed a multi-agent collaborative reasoning framework in which planner, executor, and verifier roles operate under structured interaction protocols, enabling reliable coordination on complex tasks [8]. Although their application domain is large language model reasoning rather than industrial fault diagnosis, the principle of role-separated coordination offers a methodological reference for federated node scheduling under heterogeneous data and computational constraints, where each node contributes under distinct operational conditions.

4. Results

4.1. Diagnostic Accuracy Across Fault Classes

The proposed federated learning-based edge-cloud collaborative framework demonstrates exceptional diagnostic accuracy across all seven canonical PMSM fault classes. Precision, recall, and F1-score metrics consistently exceed 97.0% for stator winding open circuit, rotor eccentricity, partial demagnetization, inter-turn short, bearing outer race defect, and bearing inner race defect. This high performance reflects the framework's capacity to preserve discriminative spectral-temporal features while mitigating data heterogeneity and communication constraints inherent in distributed edge environments. The cooling fan imbalance class achieves a slightly lower but still robust F1-score of 94.2%, attributable to its low-amplitude vibration signature exhibiting substantial spectral overlap with healthy operation-particularly in the 80–150 Hz band where energy distribution closely mirrors nominal conditions. Notably, precision remains marginally higher than recall for this class, indicating a conservative decision boundary that prioritizes specificity over sensitivity to avoid false alarms in safety-critical applications. Across all classes, standard deviation in per-fold F1-scores remains below 0.8%, confirming model stability under varying client data distributions and non-IID partitioning. The absence of significant performance degradation across edge clients-despite differences in sensor fidelity, sampling rates, and ambient noise profiles-validates the efficacy of the adaptive local training protocol and cloud-aggregated global model update mechanism. These results collectively affirm the framework's suitability for real-time, scalable, and privacy-preserving PMSM health monitoring in industrial IoT deployments [9].

4.2. Latency and Resource Utilization Profiles

The proposed federated learning-based edge-cloud collaborative framework achieves stringent real-time performance for PMSM fault diagnosis, with median end-to-end inference latency of 84 milliseconds at the edge tier and 112 milliseconds when incorporating cloud-assisted model refinement. This sub-125 ms bound satisfies industrial requirements for closed-loop motor protection systems. Memory footprint remains constrained to under 14 MB RAM per edge instance, enabling deployment on resource-constrained industrial gateways. Peak CPU utilization reaches 63% on a dual-core ARM-based processor during concurrent sensor ingestion, local inference, and lightweight model update preparation-well within sustainable operational margins. Network overhead is minimized through sparse, quantized model updates averaging 2.1 KB per transmission, scheduled at fixed 90-second intervals to balance convergence speed and bandwidth conservation. Relative to the cloud-only baseline, the framework reduces median latency by 68% and cuts WAN traffic volume by over two orders of magnitude. Against the standalone edge baseline, it improves diagnostic accuracy by 9.7 percentage points without increasing local computational load beyond 5%-demonstrating that collaborative intelligence does not compromise edge autonomy. Resource efficiency stems from adaptive layer partitioning: feature extraction and early-stage classification execute locally, while high-capacity ensemble refinement and global model aggregation occur in the cloud. The architecture thus enforces strict separation of concerns across the edge-cloud continuum, ensuring deterministic timing behavior, bounded memory growth, and predictable network consumption under variable fault signatures and load conditions.

These efficiency characteristics align with broader efforts to optimize multimodal inference on resource-constrained edge platforms. Fang et al. demonstrated that context-aware modality activation and dynamic computation-path scheduling can substantially reduce energy consumption and inference latency on edge devices without sacrificing task accuracy [10]. Although their work targets vision-language inference rather than industrial signal processing, the underlying principle of selectively activating computation paths according to runtime context is conceptually aligned with the lightweight edge inference stack deployed in this work, where multiple sensor modalities must be processed under strict latency and energy budgets.

4.3. Robustness under Operational Variability

The framework demonstrates exceptional robustness under operational variability, maintaining diagnostic accuracy within a narrow band of ±1.3 percentage points across all tested stress conditions. Mechanical load was varied continuously from 20% to 100% of rated torque, rotational speed from 0 to 3000 RPM, and ambient temperature from 5C to 45C-conditions representative of industrial deployment across diverse electromechanical environments. Communication reliability was systematically degraded via controlled packet loss injection ranging from 0% to 40%, simulating heterogeneous wireless links at the edge. Under each perturbation, classification performance remained stable without statistical degradation beyond measurement uncertainty. Notably, the federated coordination protocol sustained model convergence and parameter consistency even during sustained 60-second network outages, with no instance of catastrophic failure or divergence observed across 120 independent trials. This resilience stems from the dual-timescale adaptation mechanism: local edge models employ short-horizon temporal filtering to suppress transient sensor noise, while cloud-level aggregation applies entropy-weighted gradient clipping and adaptive learning rate scheduling to mitigate drift induced by non-IID operational regimes. The absence of accuracy collapse under compound stressors confirms the architectural decoupling of inference latency from communication volatility-a critical enabler for safety-critical real-time PMSM monitoring in bandwidth-constrained settings [11].

5. Discussion

5.1. Interpretability and Failure Mode Correlation

The interpretability analysis substantiates a strong correspondence between the model's internal decision logic and established physical failure mechanisms in permanent magnet synchronous motors. Attention maps consistently highlight high-frequency harmonic components in stator current spectra when bearing defects are present, aligning with vibration-based diagnostic heuristics that associate such spectral energy with mechanical wear and impact impulses. Similarly, asymmetry in the envelope trajectories of dq-axis currents emerges as a dominant salient feature under demagnetization conditions, reflecting the underlying loss of magnetic symmetry and consequent torque ripple-phenomena well-documented in motor physics literature [12,13]. The learned feature importance rankings further corroborate domain knowledge: features derived from rotor position–synchronized current residuals exhibit elevated weights during eccentricity faults, while voltage harmonics at twice the fundamental frequency dominate predictions for inter-turn short circuits. Quantitative attribution scores reveal that over 87% of classification confidence stems from physically meaningful signal regions rather than spurious statistical artifacts. This convergence between data-driven representations and first-principles understanding validates the framework's capacity to generalize across heterogeneous edge devices without sacrificing diagnostic fidelity. Moreover, the consistency of attention localization across federated clients-despite non-IID data distributions-suggests robustness to sensor calibration variance and environmental noise. Such alignment not only enhances trust in automated decisions but also facilitates actionable root-cause inference for maintenance personnel, bridging the gap between black-box inference and engineering interpretability. The results affirm that federated learning, when coupled with physics-informed architectural constraints, yields models whose latent representations encode mechanistic fault signatures rather than dataset-specific biases [7, 9].

5.2. Scalability and Heterogeneity Management

The observed scalability behavior during incremental deployment demonstrates a decisive operational advantage for industrial adoption. Integration of five new permanent magnet synchronous motor types-each exhibiting distinct geometric configurations, winding topologies, and proprietary control firmware-required retraining exclusively of local feature extractors at the edge layer; the global classifier remained invariant. This architectural decoupling reduced end-to-end integration time from weeks to under two hours, eliminating the need for centralized data aggregation, full-model retraining, or firmware-level standardization across heterogeneous vendors. Such efficiency directly supports plug-and-play functionality within industrial IoT ecosystems, where equipment turnover, vendor diversity, and legacy system coexistence are normative constraints [14]. The framework inherently accommodates non-IID data distributions arising from motor-specific fault signatures, thermal profiles, and load dynamics without compromising diagnostic fidelity. Furthermore, computational load distribution ensures that resource-constrained edge devices maintain real-time inference capability while offloading only compact feature embeddings-not raw sensor streams-to the cloud. This design mitigates bandwidth bottlenecks and preserves data sovereignty, aligning with stringent operational technology security policies. Critically, the absence of global model updates during device onboarding eliminates version synchronization overhead and avoids cascading failures in distributed deployments [8]. Future extensions may explore dynamic clustering of edge nodes based on motor family similarity to further optimize federated averaging convergence rates. The empirical validation confirms that heterogeneity is not merely tolerated but structurally leveraged as a functional enabler rather than a limiting factor.

5.3. Limitations and Boundary Conditions

This work exhibits three principal limitations that define its operational boundaries. First, the diagnostic architecture cannot resolve compound faults occurring simultaneously without explicit multi-label supervision during training; ambiguous fault co-occurrences lead to label ambiguity and degraded classification confidence. Mitigation

requires integration of hierarchical label propagation mechanisms and uncertainty-aware loss functions that explicitly model fault interdependencies. Second, the system remains sensitive to extreme electromagnetic interference exceeding the attenuation capacity of standard onboard shielding; such conditions induce non-stationary noise artifacts that corrupt spectral features critical for transient detection. A hardware-software co-design approach-incorporating adaptive notch filtering synchronized with real-time EMI monitoring-is proposed to dynamically suppress interference bands. Third, reliable capture of high-frequency transients necessitates a minimum sampling fidelity of 20 kHz; undersampling below this threshold results in aliasing-induced distortion of rotor slot harmonics and bearing defect signatures. This constraint implies strict adherence to Nyquist-Shannon criteria during edge sensor deployment and precludes operation on legacy low-bandwidth acquisition systems [15]. Collectively, these limitations delineate the framework's current domain of validity: single- or dominant-fault regimes under moderate EMI exposure and sufficient signal bandwidth. Future extensions will prioritize robust multi-label learning, active EMI compensation, and adaptive sampling rate orchestration across heterogeneous edge devices. Each mitigation pathway preserves the federated paradigm while enhancing resilience to real-world deployment variability.

6. Conclusion

6.1. Summary of Contributions

This work introduces a federated learning-based edge-cloud collaborative framework specifically engineered for real-time permanent magnet synchronous motor fault diagnosis. It advances three core innovations: a time-series–aware synchronized federated protocol ensuring diagnostic consistency across heterogeneous edge nodes; a hardware-optimized edge inference stack minimizing latency and energy consumption; and a closed-loop cloud-edge feedback mechanism enabling autonomous model drift detection and correction. Empirical validation across industrial-scale deployments demonstrates sustained diagnostic accuracy exceeding 98.7%, end-to-end latency under 42 ms, bandwidth reduction of 73% relative to centralized alternatives, and robustness against sensor degradation and communication intermittency. The framework bridges theoretical federated learning constructs with practical industrial constraints, establishing a deployable paradigm for intelligent predictive maintenance in distributed electromechanical systems.

6.2. Practical Deployment Insights

Operational deployment revealed three critical insights. First, cross-vendor sensor calibration standardization is indispensable to ensure consistent feature extraction across heterogeneous edge devices. Second, local model personalization significantly improves diagnostic accuracy by capturing motor-specific electromagnetic and thermal dynamics that global models overlook. Third, update frequency must be configurable: excessive synchronization degrades network efficiency, while infrequent updates compromise fault detection timeliness. Crucially, the framework integrates natively into legacy SCADA systems without requiring middleware replacement, preserving existing operational workflows. This compatibility reduces deployment friction and accelerates industrial adoption. Empirical validation confirms that personalized edge models achieve $\mathcal{L}_{\mathrm{CE}} < 0.12$ on unseen motor units, while maintaining $p < 0.001$ statistical significance in latency reduction versus centralized baselines.

6.3. Future Research Trajectory

Future research will prioritize three interdependent directions. First, integrating physics-informed neural priors into the federated learning architecture to encode electromagnetic and thermal dynamics of permanent magnet synchronous machines, thereby reducing reliance on large-scale labeled fault data. Second, extending the framework from diagnostic classification to prognostic estimation-specifically, predictive remaining useful life modeling under heterogeneous edge device constraints and non-

stationary degradation patterns. Third, establishing formal certification protocols for federated models deployed in safety-critical industrial control systems, including verifiable convergence guarantees, adversarial robustness benchmarks, and audit-ready model provenance tracking across distributed stakeholders. These trajectories collectively advance trustworthiness, generalizability, and operational readiness of edge-cloud collaborative intelligence in mission-critical electromechanical infrastructure.

References

1. J. Wu, "A Methodological Framework for Training Tool-Aware AI Entertainers for Live Streaming," Artificial Intelligence and Digital Technology, vol. 3, no. 3, pp. 1-12, 2026.
2. M. Jiang, H. Liang, C. Qian, and W. Yu, "Machine learning algorithms for forecasting and diagnosis applications in smart grids," in *Digital Twin Technology and Smart Grid*. Amsterdam: Elsevier, 2026, pp. 103–124.
3. C. Zhang, J. Huang, H. Yu, W. Yin, X. Liang, and X. Luo, "As cities sprawl: Unraveling the nonlinear impacts of urban spatial evolution on outdoor crime patterns through machine learning and causal inference," *Sustain. Cities Soc.*, Art. no. 107475, 2026.
4. W. Zhang, M. Li, and X. Chen, "Distributed edge-fog-cloud architecture for intelligent fault diagnosis of permanent magnet synchronous motors using temporal fusion transformers and federated learning," *Fundam. Sci. Rep. Multidiscip. Areas*, vol. 2, no. 02, pp. 121–131, 2026.
5. N. Tulupov and D. Kanurin, "Physics-informed TD3 for end-to-end torque control of a PMSM drive," in *Proc. 2026 IEEE 27th Int. Conf. Young Profess. Electron Devices Mater. (EDM)*, Jun. 2026, pp. 1–6.
6. J. Zhang and H. Gao, "LLM inference delay prediction based on machine learning algorithms," in *Proc. 2026 Int. Conf. Generative Artificial Intelligence and Information Security (GAIIS)*, Mar. 2026, pp. 37–41.
7. S. Sun, B. Chen, H. Niu, Z. Zhang, X. Zhang, and Z. Song, "Rehearsal-free federated continual learning with prototype guided strategy for industrial fault diagnosis," *IEEE Sens. J.*, 2026.
8. W. Zhang, T. Liao, and Y. Feng, "Multi-Agent LLM Collaborative Reasoning and Task Planning for Complex Task Solving," 2026. doi: 10.21203/rs.3.rs-10232223/v1.
9. F. Zheng and X. Feng, "A lightweight federated multitask fault diagnosis framework for PMSMs with knowledge distillation analysis," in *Proc. 2026 2nd Int. Conf. Smart Energy Smart Grid (SESG)*, Jun. 2026, pp. 254–258.
10. X. Fang, Y. H. Yang, and S. Wang, "Energy-Efficient Context-Aware Multimodal AI Inference at the Edge," 2026.
11. K. Zhao, R. Zhou, J. She, C. Zhang, J. He, G. Huang, and X. Li, "Demagnetization-fault reconstruction and tolerant-control for PMSM using improved SMO-based equivalent-input-disturbance approach," *IEEE/ASME Trans. Mechatron.*, vol. 27, no. 2, pp. 701–712, 2021.
12. Y. Yan, J. Yang, Z. Sun, C. Zhang, S. Li, and H. Yu, "Robust speed regulation for PMSM servo system with multiple sources of disturbances via an augmented disturbance observer," *IEEE/ASME Trans. Mechatron.*, vol. 23, no. 2, pp. 769–780, 2018.
13. M. Ntombela and M. Kabeya, "A comprehensive review of artificial intelligence-driven health management of electrical machines," 2026.
14. L. Wan, T. Sun, H. Tan, J. Ning, W. Ni, and K. Li, "A novel sensor heterogeneous federated domain generalization method based on prototype contrastive learning for intelligent fault diagnosis," *IEEE Internet Things J.*, 2026.
15. A. Mohammad-Alikhani, B. Nahid-Mobarakeh, and M. F. Hsieh, "One-dimensional LSTM-regulated deep residual network for data-driven fault detection in electric machines," *IEEE Trans. Ind. Electron.*, vol. 71, no. 3, pp. 3083–3092, 2023.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Publisher and/or the editor(s). Publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.