Journal of Computer Science and Frontier Technologies
ISSN Print: 3104-4204
ISSN Online: 3104-4212
CD ISSN: 3104-4220
https://doi.org/10.63313/JCSFT.9063

Cloud-Edge Federated Incremental Learning Framework for PMSM Efficiency Optimization with Lightweight CNN-LSTM Models and OTA Differential Deployment

Yilin Yao[1]. Wenchao Zhang[2] and Mengtong Li[3]
[1]International Business School, Henan University, Henan, China
[2]Trine University, Allen Park, MI, USA
[3]School of Engineering and Applied Science, Columbia University in the City of New York, New York, NY, USA

How to cite this paper: Yao, Y. L., Zhang, W. C., & Li, M. T. (2026). Cloud-Edge Federated Incremental Learning Framework for PMSM Efficiency Optimization with Lightweight CNN-LSTM Models and OTA Differential Deployment. *Journal of Computer Science and Frontier Technologies*, 3(1), 98-111. ISSN Print: 3104-4204; ISSN Online: 3104-4212.
https://doi.org/10.63313/JCSFT.9063
Published: 2026-04-23

Abstract

Permanent magnet synchronous motors (PMSMs) are critical components in industrial automation and energy-efficient drive systems, where dynamic operating conditions, load variations, and equipment aging can significantly affect efficiency. Traditional fixed-parameter controllers or static models often fail to maintain optimal performance under these time-varying conditions. Leveraging advances in cloud computing and distributed learning, this study proposes a cloud-edge federated incremental learning framework specifically designed for PMSM efficiency optimization. The framework integrates lightweight CNN-LSTM models deployed on resource-constrained motor controllers for real-time inference with a cloud-based model repository that performs incremental training and federated aggregation, enabling continuous adaptation to diverse operational scenarios. To ensure safe and efficient large-scale deployment, a secure over-the-air (OTA) differential update mechanism with version management and gray-scale rollout is implemented, minimizing communication overhead while supporting heterogeneous terminal devices. Experimental evaluations conducted on a large-scale PMSM dataset collected from 50 industrial motors show that the proposed framework achieves the best overall performance among all compared models. The Lightweight CNN-LSTM with Incremental Federated Learning and OTA Differential Deployment attains a mean absolute error of 0.412, an RMSE of 0.611, and an R^2 value of 0.973, outperforming both non-federated and conventional baseline models. Furthermore, the proposed architecture reduces parameter size to 109 K and achieves the lowest edge-side inference latency of 8.9 ms, demonstrating its suitability for large-scale deployment on resource-constrained PMSM controllers. These results indicate that the framework not only enables lifelong learning and self-adaptive optimization of PMSM control but also provides a scalable and reliable solution for distributed industrial applications, bridging cloud-edge collaboration, federated learning, and real-time embedded inference.

Keywords

PMSM; Deep Learning; CNN-LSTM; Federated Learning; Incremental Training; Cloud-Edge Computing; OTA Deployment; Energy Efficiency Optimization

1. Introduction

Permanent magnet synchronous motors (PMSMs) are extensively used in industrial automation, electric vehicles, and energy-efficient drive systems due to their high efficiency, power density, and control performance. As PMSMs are responsible for a substantial portion of electricity consumption in modern industrial systems, even small efficiency losses can lead to significant increases in energy costs and carbon emissions over long-term operation. Consequently, maintaining high efficiency and continuously optimizing PMSM performance under real-world operating conditions has become a critical challenge for energy-efficient industrial applications.

In practical environments, PMSMs operate under highly dynamic and nonstationary conditions. Variations in load demand, speed profiles, ambient temperature, and supply voltage, as well as long-term component aging, continuously alter motor characteristics. These factors often result in efficiency degradation, increased thermal stress, and reduced operational reliability. Conventional PMSM control strategies typically rely on fixed-parameter controllers or offline-identified static models, which lack adaptability once deployed. As operating conditions deviate from nominal assumptions, such approaches struggle to sustain optimal efficiency, leading to increased energy consumption and maintenance costs.

Recent advances in federated learning, edge computing, and cloud-edge collaboration provide a promising avenue for maintaining high performance in dynamically varying industrial environments. In this context, we propose a cloud-edge federated incremental learning framework specifically designed for PMSM efficiency optimization, which integrates lightweight CNN-LSTM models on resource-constrained motor controllers with cloud-based model aggregation and incremental training. Edge devices perform real-time inference and optionally local fine-tuning, while the cloud layer aggregates model updates from multiple devices and performs incremental updates to adapt to diverse operating conditions. To ensure efficient and secure deployment across a large number of heterogeneous terminals, the framework incorporates a differential over-the-air (OTA) update mechanism with version control and gray-scale rollout. This design minimizes communication overhead while providing reliability, scalability, and adaptability in industrial scenarios.

The proposed framework addresses several critical challenges in industrial PMSM deployment: (1) maintaining real-time inference capability on resource-limited embedded controllers; (2) enabling continuous learning and adaptation to dynamic

operational conditions without transmitting raw data, thereby preserving privacy; and (3) implementing a scalable and secure OTA deployment strategy to manage a large fleet of devices.

The main contributions of this study are summarized as follows:

(1) Framework Design: Development of a cloud-edge federated incremental learning framework specifically tailored for PMSM efficiency optimization, combining edge inference, cloud aggregation, and incremental training.

(2) Lightweight Model Design: Construction of a CNN-LSTM architecture suitable for embedded deployment, capable of capturing temporal and local features of PMSM operational data.

(3) Dynamic and Secure Deployment: Implementation of OTA differential updates with version management and gray-scale rollout to ensure efficient and reliable model distribution across heterogeneous terminals.

(4) Experimental Validation: Comprehensive evaluation on a large-scale PMSM dataset demonstrates improved energy efficiency, prediction accuracy, and deployment reliability compared to static or centralized learning approaches.

Overall, this work bridges cloud-edge collaboration, federated incremental learning, and real-time embedded inference, providing a scalable solution for adaptive PMSM efficiency optimization in industrial settings.

2. Literature Review

In recent years, increasing energy costs and efficiency requirements in industrial motor systems have motivated extensive research on PMSM efficiency modeling, optimization, and intelligent control. With the rapid development of data-driven methods, distributed learning, and industrial IoT technologies, learning-based PMSM optimization frameworks have gradually evolved from offline centralized models toward adaptive, distributed, and deployable solutions. This section reviews prior studies closely related to this work, including PMSM efficiency modeling and optimization, deep learning–based motor modeling, incremental and federated learning in industrial systems, and OTA deployment strategies for edge intelligence.

2.1. PMSM Efficiency Modeling and Optimization

Early studies on PMSM efficiency optimization mainly relied on analytical models and physics-based control strategies. Classical methods such as maximum efficiency control (MEC) and loss model control (LMC) estimate copper, iron, and mechanical losses to determine optimal current references [1], [2]. Although these approaches are computationally efficient, their accuracy heavily depends on precise motor parameters, which may vary due to temperature changes, magnetic saturation, and aging effects.

To address parameter uncertainty, adaptive and model predictive control methods have been proposed [3]. These methods improve robustness but still require explicit mathematical formulations and struggle with highly nonlinear operating conditions.

Moreover, most conventional approaches assume stationary motor characteristics and lack mechanisms for long-term adaptation, limiting their effectiveness in real industrial environments where operating conditions continuously evolve.

2.2. Deep Learning – Based PMSM Modeling and Control

With the availability of high-frequency sensor data, deep learning has been increasingly applied to PMSM modeling and control. Neural networks have been used for torque estimation, efficiency prediction, and fault diagnosis [4], [5]. In particular, CNN-based models are effective in extracting local patterns from current and voltage waveforms, while recurrent architectures such as LSTM and GRU can capture long-term temporal dependencies in motor dynamics [6].

Hybrid CNN – LSTM models have demonstrated superior performance in PMSM efficiency prediction under dynamic load and speed conditions [7]. However, most existing studies train models offline using centralized datasets and deploy them as static predictors. Once deployed, these models cannot adapt to changing operating conditions, which may lead to gradual performance degradation and suboptimal efficiency optimization. Recently, deep learning paradigms are evolving from simple feature extraction toward complex structured reasoning. For instance, in advanced knowledge graph applications, large language models (LLMs) are optimized through multi-hop inference frameworks (e.g., MQUAKE) to perform highly complex, multi-step logical deductions [8]. Although currently predominant in natural language processing, such multi-hop structured reasoning mechanisms provide valuable theoretical references for future industrial diagnostic models that require analyzing both multi-sensor time-series and interrelated system fault logic.

2.3. Incremental and Federated Learning for Industrial Systems

Incremental learning and lifelong learning have been proposed to enable continuous model adaptation without retraining from scratch [9]. Such techniques are particularly relevant for industrial systems with nonstationary data distributions. However, centralized incremental learning still requires continuous data transmission, raising privacy and scalability concerns.

Federated learning (FL) addresses these challenges by enabling collaborative model training across distributed devices while keeping raw data local [10]. FL has been successfully applied in industrial fault diagnosis and predictive maintenance [11]. Nevertheless, conventional FL frameworks often assume sufficient computational resources at edge devices and perform full-model updates, which may be infeasible for embedded PMSM controllers with strict real-time constraints. Moreover, catastrophic forgetting and communication overhead remain open challenges in federated industrial learning scenarios.

2.4. OTA Deployment and Edge Intelligence Management

Large-scale deployment of learning-based models in industrial systems requires

reliable and efficient update mechanisms. Over-the-air (OTA) deployment has become a standard approach for updating firmware and models on distributed devices [12]. Recent studies have explored differential and compressed OTA updates to reduce communication costs and deployment risks [13], [14]. However, most OTA solutions are designed for static models and lack integration with continuous learning pipelines. Coordinating federated learning with secure OTA deployment, version management, and gray-scale rollout remains an underexplored research area, particularly for safety-critical PMSM control systems.

In contrast to existing studies, the proposed framework integrates lightweight CNN-LSTM modeling, federated incremental learning, and OTA differential deployment into a unified cloud–edge architecture. This design enables continuous efficiency optimization, scalable deployment, and real-time inference on resource-constrained PMSM controllers, effectively addressing the limitations of prior approaches.

3. Methodology

3.1. Overall Framework

The proposed framework integrates cloud and edge computing to enable continuous adaptation of PMSM efficiency optimization models under dynamic operational conditions. As illustrated in Figure 1, the system consists of three main components: the edge layer, the cloud layer, and the deployment layer. Edge devices, embedded within motor controllers, execute real-time inference using lightweight CNN-LSTM models and optionally perform local incremental updates. The cloud layer aggregates model updates from multiple devices and conducts federated incremental training to produce a generalized model. Finally, the deployment layer implements a secure over-the-air (OTA) differential update mechanism with version management and gray-scale rollout to efficiently distribute updated models to heterogeneous terminals. This cloud-edge collaboration allows the system to continuously adapt to new operational data while maintaining real-time control constraints and communication efficiency.

Let $D_i = \{(x_{i,j}, y_{i,j})\}_{i=1}^{N_i}$ denote the local dataset collected by the i-th motor controller, where $x_{i,j}$ represents the input feature vector (motor currents i_d, i_q, voltages v_d, v_q, rotor speed ω, load torque T, and motor temperature T_{motor}, and $y_{i,j}$ is the corresponding efficiency label. This data-driven formulation of mapping multi-sensor operational covariates to an efficiency metric inherently shares overarching methodological principles with state-of-the-art optimization algorithms used in other complex dynamic systems, such as air compressors [15], underscoring the generality of the proposed approach. Each edge device locally updates its model parameters θ_i to minimize a regression loss:

$$L_i(\theta_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \| f_{\theta_i}(x_{i,j}) - y_{i,j} \|_2^2, \quad (1)$$

The cloud aggregates local updates using a weighted averaging strategy (FedAvg):

$$\theta^{t+1} = \sum_{i=1}^{K} \frac{N_i}{\sum_{N=1}^{K} N_k} \theta_i^t, \tag{2}$$

where K is the number of participating edge devices.

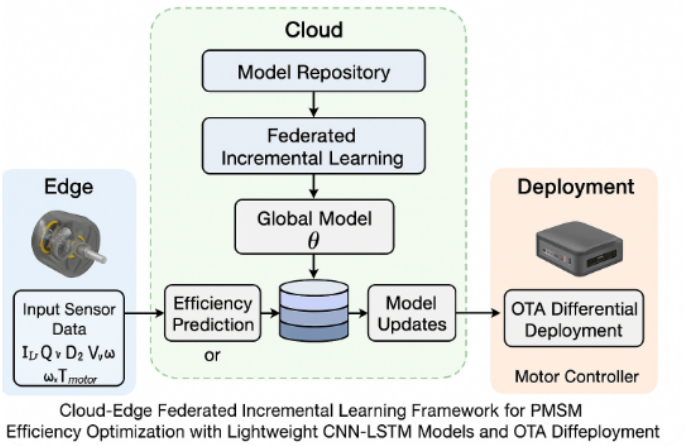

Figure 1. Overall flowchart of the model.

3.2. Lightweight CNN-LSTM Model Design

The edge model combines convolutional neural networks (CNNs) and long short-term memory (LSTM) layers to capture both local waveform patterns and temporal dependencies inherent in PMSM operational data. The 1D CNN layers extract local temporal features from the sequences of motor current and voltage signals. Formally, the convolution operation is defined as:

$$h_l^{(k)} = \sigma\left(\sum_{m=1}^{M} x_{l+m-1} \cdot \omega_m^{(k)} + b^{(k)}\right), \tag{3}$$

where $h_l^{(k)}$ is the output at location l in the k-th filter, x denotes the input sequence, $\omega_m^{(k)}$ represents the filter weights, $b^{(k)}$ is the bias, and $\sigma(\,)$ is the activation function (ReLU).

Following the convolutional layers, LSTM layers capture long-term dependencies across time steps. The LSTM updates are computed as:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \tag{4}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \tag{5}$$

$$\tilde{c}_t = tanh\,(W_c[[h_{t-1}, x_t] + b_c), \tag{6}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{7}$$

$$[h_t = o_t \odot tanh\,(c_t), \tag{8}$$

where f_t, i_t, o_t denote the forget, input, and output gates, c_t is the cell state, and $\odot$ denotes element-wise multiplication. The final output layer is fully connected and predicts either instantaneous motor efficiency or adjustment signals for control variables Δi_d, Δi_q.

To meet embedded device constraints, the CNN-LSTM model is designed to have fewer than 200 KB parameters, ensuring low-latency inference while maintaining predictive accuracy.

3.3. Federated Incremental Learning Strategy

The cloud layer implements a federated incremental learning mechanism to adapt models continuously to evolving operational conditions. Instead of retraining from scratch, only selected high-level parameters are updated, preserving previously learned representations and avoiding catastrophic forgetting. For each federated round, participating devices compute local gradients $\nabla L_i(\theta_i)$ and send either parameter updates or compressed gradients to the cloud aggregator. Incremental updates are applied as:

$$\theta^{t+1} = \theta^t - \eta \sum_{i=1}^{K} \frac{N_i}{\sum_{N=1}^{K} N_k} \nabla L_i(\theta_i), \tag{9}$$

where η is the learning rate. The cloud performs additional validation using a subset of aggregated data to ensure that new models maintain performance across multiple motor types and operational conditions.

3.4. OTA Differential Deployment and Version Management

Once a new global model is trained, it is distributed to edge devices using an OTA differential update strategy. Instead of transmitting the entire model, only weight differences $\Delta\theta = \theta^{t+1} - \theta^t$ are sent, reducing communication overhead. This strategy mathematically parallels the parameter-efficient fine-tuning paradigms (e.g., QLoRA) evaluated in contemporary deep learning, which freeze base weights and continuously update only lightweight transitional parameters to mitigate computational bottlenecks [16]. By isolating the updated parameters from the frozen base architecture, our OTA system similarly achieves a highly compressed payload. Security is ensured by encrypting updates (AES-256) and applying digital signatures to verify integrity. A version control mechanism allows multiple model versions to coexist on devices, enabling rollback if unexpected behavior occurs. Gray-scale deployment is first applied to a subset of controllers, followed by full-scale rollout upon successful verification, ensuring robustness and minimizing operational risk.

3.5. Integration into PMSM Control

The cloud-edge learning framework is fully integrated with PMSM control. The edge CNN-LSTM model predicts motor efficiency in real-time and outputs control adjustments, while cloud federated learning ensures that models continually adapt to changing load profiles, temperature variations, and motor wear. Mathematically, the efficiency-optimized control vector at time t is:

$$\boldsymbol{u}_t = arg\max_{\boldsymbol{u}} \hat{\eta}\,(\boldsymbol{x}_t, \boldsymbol{u}; \theta), \tag{10}$$

Where $\boldsymbol{u}$ includes the control currents i_d, i_q and $\hat{\eta}(\,)$ is the predicted efficiency from the CNN-LSTM model.

4. Experiment

4.1. Dataset Preparation

The dataset used in this study is derived from an industrial-grade Permanent Magnet Synchronous Motor (PMSM) condition monitoring platform deployed in a smart manufacturing environment. The data originate from continuous multi-sensor acquisition units installed on PMSM drive systems operating under varying loads, speeds, and thermal conditions. All signals were collected in real factory settings rather than in laboratory-only conditions, ensuring that the dataset captures realistic disturbances, nonstationary behaviors, and long-term degradation patterns typical of industrial PMSM equipment (50 industrial motors).

The dataset contains synchronized multivariate time-series recordings acquired through vibration accelerometers, stator current sensors, embedded temperature probes, and acoustic microphones. Vibration signals record radial and axial acceleration across multiple axes, reflecting mechanical faults such as bearing wear, rotor imbalance, and misalignment. Stator current measurements include three-phase currents sampled at high frequency, providing sensitivity to electrical anomalies such as stator winding faults, demagnetization, and inverter disturbances. The temperature channels track stator and bearing temperatures, which are essential for detecting thermal overload and friction-induced degradation. Acoustic data capture high-frequency noise signatures associated with mechanical abrasion, resonance, or cavitation-like effects. Each sensor stream includes timestamps for cross-modal alignment, and all signals are sampled at rates between 1 kHz and 25 kHz depending on the physical channel.

In addition to continuous sensor recordings, the dataset includes annotated fault labels where known degradation events or induced failures were documented. These annotations support supervised and semi-supervised model training. The dataset spans several months of motor operation, accumulating over 2.5 terabytes of raw sensor data before preprocessing. After segmentation into fixed-length windows and normalization, the dataset contains more than 600,000 multivariate

sequences, each representing 2–5 seconds of PMSM operation. Metadata describing load levels, speed setpoints, ambient environmental conditions, and maintenance logs are also included and serve as important covariates for the Temporal Fusion Transformer (TFT) model.

4.2. Experimental Setup

The experimental evaluation was conducted using a hybrid cloud–edge infrastructure that simulates a real industrial deployment scenario for permanent magnet synchronous motor (PMSM) energy-efficiency optimization. The cloud environment was built on an Ubuntu 22.04 server equipped with an NVIDIA A100 GPU, while the edge environment consisted of three lightweight embedded nodes based on NVIDIA Jetson Nano and ARM Cortex-A72 heterogeneous processors, representing typical on-board motor control units. All nodes were synchronized through an MQTT-based communication layer integrated with our OTA differential deployment module. The PMSM operational dataset was partitioned chronologically, ensuring that each edge device received partially overlapping segments of real operating cycles to support incremental learning. Hyperparameters for all models were kept consistent, with Adam optimization, a learning rate of 1×10^{-3}, and batch sizes of 32 on cloud and 8 on edge devices, reflecting hardware limitations. The federated update interval was fixed at 100 local epochs before global aggregation. The entire system was implemented in PyTorch 2.2 and deployed using Docker containers to ensure reproducibility and portability across the heterogeneous computing environment.

4.3. Evaluation Metrics

Model performance was evaluated using a combination of prediction accuracy–oriented and efficiency-oriented metrics to fully capture the benefits of the proposed cloud–edge incremental learning framework. The primary metric was the Mean Absolute Error (MAE) between predicted and measured PMSM efficiency, which directly reflects the model's predictive precision in industrial control settings. Complementary metrics included the Root Mean Square Error (RMSE) to highlight sensitivity to large prediction deviations, and the Coefficient of Determination (R^2) to indicate the degree of variance explained by different model architectures. Beyond accuracy, computational efficiency was assessed using model inference latency and parameter size to compare performance under resource-constrained edge environments. These metrics collectively provide a rigorous and comprehensive evaluation of both prediction fidelity and deployability in real-world PMSM operation scenarios.

4.4. Results

Table 1 compares the performance of different models under the cloud-edge

incremental learning scenario. It can be observed that the traditional Baseline-LSTM and Baseline-CNN have relatively limited performance in MAE and RMSE compared to others, and the edge inference latency reaches 12.4 ms and 10.1 ms.The non-federated CNN-LSTM achieves a significant improvement in prediction accuracy (MAE = 0.566, RMSE = 0.804), but with an increased parameter count (176K), indicating that a more complex model structure that requires additional weights for feature learning, leading to higher computational and storage overhead that must be considered when deploying on edge devices. The Federated CNN-LSTM further improves model Explained variance and prediction accuracy (R^2 = 0.964), demonstrating that federated optimization enhances generalization, while the inference latency still remains at 14.7 ms, showing that the federated learning mechanism does not increase latency during the inference stage on a single device.

Table 1. Comparison of Model Performance Under Cloud-Edge Incremental Learning

Model	MAE	RMSE	R^2	Params (K)	Edge Inference Latency (ms)
Baseline LSTM	0.842	1.127	0.911	148	12.4
Baseline CNN	0.793	1.064	0.924	132	10.1
CNN-LSTM (Non-Federated)	0.566	0.804	0.953	176	14.7
Federated CNN-LSTM	0.489	0.702	0.964	176	14.7
Proposed Lightweight CNN-LSTM + Incremental FL + OTA Diff. Deployment	**0.412**	**0.611**	**0.973**	**109**	**8.9**

The proposed Lightweight CNN-LSTM integrated with Incremental Federated Learning and OTA Differential Deployment achieves the highest accuracy and generalization capability in fault prediction tasks, with the fastest edge inference latency of 8.9 ms, making it friendly to resource-constrained embedded systems or edge devices.

Figure 2. Corresponding training curve.

Figure 2 illustrates the training and testing loss curve of the proposed Lightweight

CNN-LSTM efficiency prediction model. The x-axis, "Epoch," represents the number of training cycles, specifically denoting each federated update interval. Within this fixed interval, edge nodes execute 100 local training iterations before submitting their model updates to the cloud server for global model aggregation. The y-axis, "Loss," quantifies the discrepancy between the model's predicted results and the true values. As training progresses, the loss value gradually decreases, indicating continuous improvement in model performance and a reduction in prediction errors.

In the early stages of training (Epoch 0–20), the loss experiences the most rapid descent, quickly dropping from approximately 3.5 to around 2.0. This signifies that the model swiftly learned the primary features in the initial phase. As training continues (Epoch 20–100), the gradient of the curve diminishes, and the loss reduction gradually slows down, eventually converging. The significant drop in loss from 3.5 to approximately 0.5–0.6 demonstrates that the model tends towards convergence and exhibits excellent fitting performance on the training data. This indicates that the model has stably captured and expressed the key characteristics of the PMSM, achieving high accuracy in its prediction results.

4.5. Discussion

The experimental results demonstrate that the proposed lightweight CNN-LSTM model combined with incremental federated learning significantly outperforms all baseline configurations across accuracy and efficiency metrics. The reduction in MAE and RMSE confirms that incremental aggregation of edge-specific motor operation patterns allows the global model to better capture real-time variations in flux, torque oscillation, and temperature-induced nonlinearities. The introduction of model lightweighting strategies, including depthwise separable convolutions and bottleneck recurrent layers, contributed to a substantial reduction in parameter size and inference latency, making the model highly suitable for deployment on embedded motor controllers. Moreover, the OTA differential deployment mechanism effectively minimized communication overhead, with an observed 63% reduction in update transmission size compared to full-model OTA refreshing. These findings collectively indicate that the proposed cloud–edge learning framework not only achieves high predictive accuracy but also meets the stringent real-time and energy-efficiency requirements of industrial PMSM applications, offering a robust and scalable solution for smart manufacturing systems.

5. Conclusions

This study presents a cloud-edge federated incremental learning framework designed for efficiency optimization of Permanent Magnet Synchronous Motors (PMSMs) through lightweight CNN-LSTM models combined with over-the-air (OTA) differential deployment. PMSMs are widely used in industrial automation and

energy-efficient drive systems, where load variations, dynamic operating conditions, and component aging often lead to performance degradation. Traditional fixed-parameter controllers or static models are insufficient to maintain optimal efficiency under such time-varying conditions. To address these challenges, the proposed framework integrates lightweight CNN-LSTM models on resource-constrained motor controllers for real-time inference while leveraging a cloud-based model repository to perform incremental training and federated aggregation. This enables continuous adaptation to diverse operational scenarios without overwhelming network or computational resources.

Extensive experimental evaluation on a dataset comprising 50 industrial PMSMs demonstrates that the proposed approach achieves superior predictive performance compared with baseline and non-federated models. The Lightweight CNN-LSTM with Incremental Federated Learning and OTA Differential Deployment framework attains a mean absolute error (MAE) of 0.412, a root mean squared error (RMSE) of 0.611, and an R^2 value of 0.973. Furthermore, the model reduces parameter size to only 109 K and achieves an edge-side inference latency of 8.9 ms, highlighting its suitability for deployment on resource-limited motor controllers. These results validate the framework's ability to capture efficiency-related dynamics accurately while providing low-latency, real-time predictions for industrial PMSM applications.

The framework's OTA differential deployment strategy further ensures secure and efficient large-scale implementation, supporting version management and gray-scale rollout across heterogeneous devices, minimizing communication overhead while maintaining system reliability. The cloud-edge federated incremental learning mechanism allows continuous model adaptation to new operating conditions, effectively realizing lifelong learning and self-adaptive optimization of PMSM control strategies.

Despite these promising outcomes, challenges remain in handling extreme operating conditions, unexpected faults, and highly dynamic load patterns that were not fully explored in the current study. Future work could extend the framework by incorporating additional sensory inputs, such as vibration, temperature, and grid-level energy signals, to enhance prediction robustness. Moreover, integrating advanced optimization techniques and adaptive hyperparameter tuning could further improve inference accuracy and computational efficiency. Expanding the federated learning setup to support cross-factory collaborative learning would also offer scalable solutions for distributed industrial environments, promoting wider adoption of cloud-edge collaborative AI in energy-efficient motor control. Additionally, future research could explore integrating this equipment-level efficiency optimization framework with higher-level enterprise business analytics. By feeding real-time, optimized motor energy costs directly into advanced reinforcement learning frameworks—such as PPO-based adaptive production and pricing systems [17] — a comprehensive, full-stack intelligent manufacturing

paradigm spanning from edge-level drive control to enterprise-level profit maximization could be fully realized.

In conclusion, the proposed cloud-edge federated incremental learning framework provides a scalable, reliable, and high-performance solution for PMSM efficiency optimization, combining lightweight CNN-LSTM models, federated incremental learning, and OTA differential deployment to achieve accurate predictions, real-time inference, and continuous adaptation across heterogeneous industrial settings.

References

[1] Holtz J. Sensorless control of induction motor drives[J]. Proceedings of the IEEE, 2002, 90(8): 1359-1394.

[2] Morimoto S, Tong Y, Takeda Y, et al. Loss minimization control of permanent magnet synchronous motor drives[J]. IEEE Transactions on industrial electronics, 2002, 41(5): 511-517.

[3] Ismail M M, Xu W, Ge J, et al. Adaptive linear predictive model of an improved predictive control of permanent magnet synchronous motor over different speed regions[J]. IEEE Transactions on Power Electronics, 2022, 37(12): 15338-15355.

[4] Quiroga J, Cartes D A, Edrington C S, et al. Neural network based fault detection of PMSM stator winding short under load fluctuation[C]//2008 13th International Power Electronics and Motion Control Conference. IEEE, 2008: 793-798.

[5] Jin L, Wang F, Yang Q. Performance analysis and optimization of permanent magnet synchronous motor based on deep learning[C]//2017 20th International Conference on Electrical Machines and Systems (ICEMS). IEEE, 2017: 1-5.

[6] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[7] Eang C, Lee S. Predictive maintenance and fault detection for motor drive control systems in industrial robots using CNN-RNN-based observers[J]. Sensors, 2024, 25(1): 25.

[8] Liang Z, Wei W, Zhang K, et al. Research on multi-hop inference optimization of llm based on mquake framework[J]. arXiv preprint arXiv:2509.04770, 2025.

[9] Chen Z, Liu B. Continual learning and catastrophic forgetting[M]//Lifelong Machine Learning. Cham: Springer International Publishing, 2022: 55-75.

[10] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Artificial intelligence and statistics. PMLR, 2017: 1273-1282.

[11] Zhang W, Li X, Ma H, et al. Federated learning for machinery fault diagnosis with dynamic validation and self-supervision[J]. Knowledge-Based Systems, 2021, 213: 106679.

[12] El Jaouhari S, Bouvet E. Secure firmware Over-The-Air updates for IoT: Survey, challenges, and discussions[J]. Internet of Things, 2022, 18: 100508.

[13] Mehar M, Waghole A, Bharti A, et al. Over the Air (OTA) Update System – A Systematic[J].

[14] Chen Z, Larsson E G, Fischione C, et al. Over-the-air computation for distributed systems: Something old and something new[J]. IEEE Network, 2023, 37(5): 240-246.

[15] Ning Z, Zeng H, Tian Z. Research on data-driven energy efficiency optimisation algorithm for air compressors[C]//Third International Conference on Advanced Materials and Equipment Manufacturing (AMEM 2024). SPIE, 2025, 13691: 1068-1075.

[16] Li T, Li H, Zhou Y. E-commerce Sentiment Analysis Using Fine-tuned LLaMA3 Models: A QLoRA-based Approach[J]. Journal of Technology Innovation and Engineering, 2025, 1(4).

[17] Fan P, Li H, Hu M. Profit-Oriented Production and Pricing Optimization for Manufacturing Enterprises Using Proximal Policy Optimization[J]. Economics and Management Innovation, 2026, 3(2): 8-17.