TANMEAH BASSMAH PRESS
Universal Journal of Future Intelligence: Innovations and Artificial Intelligence (UJFIIAI)
اَلْمَجَلَّة اَلْعَالَمِيَّة لِلذَّكَاءِ اَلْمُسْتَقْبَلِيّ: اَلِابْتِكَارَاتُ وَاَلذَّكَاءُ اَلِاصْطِنَاعِيُّ
Journal Homepage: https://tanmeah.com/ Volume 2, Issue 2, 2025

Enhancing Efficiency and Robustness of Compressed Deep Learning Models on Edge Devices for IoT Sensor Data

Qutaiba Qasem Ahmad Azqiba[1]* & Yanne, D. LaCuen[2]

[1,2] Department of Computer Science, New York University, New York, NY, USA.

C H R O N I C L E
Article history:
Received: July, 15, 2025.
Received in revised format: August, 21, 2025.
Accepted: December, 11, 2025.
Available online: December, 31, 2025.

Keywords:
Edge AI, PTQ, Edge computing, QAT, Quantization-aware training, Deep learning, Model compression, Sensor noise, Robustness.

A B S T R A C T

The proliferation of Internet of Things devices has amplified the need for efficient and reliable edge-based intelligence. Compressed deep learning models, particularly those using post-training quantization and quantization-aware training, offer promising solutions for resource-constrained edge deployments. However, real-world IoT sensor data are often noisy, incomplete, or subject to drift, which can significantly degrade model performance. This study systematically investigates the trade-offs between computational efficiency and robustness of compressed models under realistic sensor noise conditions. Through controlled experiments, we demonstrate that PTQ, while highly efficient in terms of latency, is prone to substantial accuracy loss in noisy environments. In contrast, QAT preserves model accuracy while maintaining low inference latency, providing a more reliable solution for practical edge deployments. The results highlight the necessity of robustness-aware compression and underscore the limitations of evaluating models solely on clean datasets. This work offers a deployment-centric framework for assessing model trustworthiness and provides guidance for developing lightweight, robust deep learning models for safety-critical IoT applications, including industrial monitoring and healthcare.

الكلمات الدالة:
الذكاء الاصطناعي على الحافة، التقييم الكمي، الحوسبة على الحافة، التقييم الكمي، التدريب الواعي بالكمية، التعلم العميق، ضغط النموذج، ضوضاء المستشعر، والمتانة.

الملخص

أدى انتشار أجهزة إنترنت الأشياء إلى زيادة الحاجة إلى ذكاء فعال وموثوق قائم على الحوسبة الطرفية. تُقدم نماذج التعلم العميق المضغوطة، ولا سيما تلك التي تستخدم التكميم بعد التدريب والتدريب المراعي للتكميم، حلولًا واعدة لعمليات النشر الطرفية ذات الموارد المحدودة. مع ذلك، غالبًا ما تكون بيانات مستشعرات إنترنت الأشياء في العالم الحقيقي مشوشة أو غير مكتملة أو عرضة للانحراف، مما قد يُؤدي إلى تدهور كبير في أداء النموذج. تُحقق هذه الدراسة بشكل منهجي في المفاضلات بين الكفاءة الحسابية ومتانة النماذج المضغوطة في ظل ظروف ضوضاء المستشعرات الواقعية. من خلال تجارب مضبوطة، تُبين أن التكميم بعد التدريب، على الرغم من كفاءته العالية من حيث زمن الاستجابة، إلا أنه عُرضة لفقدان كبير في الدقة في البيئات المشوشة. في المقابل، يحافظ التدريب المراعي للتكميم على دقة النموذج مع الحفاظ على زمن استجابة استدلال منخفض، مما يُوفر حلًا أكثر موثوقية لعمليات النشر الطرفية العملية. تُبرز النتائج ضرورة الضغط المراعي للمتانة وتُؤكد على محدودية تقييم النماذج على مجموعات البيانات النظيفة فقط. يقدم هذا العمل إطار عمل يركز على النشر لتقييم موثوقية النموذج ويوفر إرشادات لتطوير نماذج التعلم العميق خفيفة الوزن وقوية لتطبيقات إنترنت الأشياء ذات الأهمية البالغة للسلامة، بما في ذلك المراقبة

* Corresponding author.
E-mail address: qtaebahz@gmail.com
http://doi.org/10.70568/UJFIIAI.2.2.3

الصناعية والرعاية الصحية.

JEL Classification: *C88, O33, L86.*

1. Introduction

The rapid proliferation of Internet of Things (IoT) systems has generated an unprecedented volume of sensor data, enabling a wide spectrum of applications ranging from industrial automation to healthcare monitoring and smart cities (Rane, Choudhary, & Rane, 2023; Rathore, Ahmad, Paul, & Rho, 2016). Historically, such sensor data have been processed in cloud environments; however, latency, connectivity, and privacy constraints have increasingly necessitated the adoption of edge computing paradigms (Bourechak et al., 2023). By moving computation closer to the data source, edge computing reduces inference latency and network load while enhancing real-time responsiveness. Despite these advantages, edge devices are inherently resource-constrained, with limited memory, computation, and energy budgets, presenting significant challenges for deploying deep learning (DL) models (Iqbal, Davies, & Perez, 2024; Ajani, Imoize, & Atayero, 2021). To mitigate these limitations, model compression techniques (most notably quantization and pruning) have been widely proposed for efficient on-device inference (Zhu et al., 2024; Ma, Fang, & Wang, 2023; Sun et al., 2023; Frantar et al., 2023; Xiao et al., 2023; Gu et al., 2024). While these techniques successfully reduce computational overhead and memory footprint, there is growing evidence that such efficiency gains may compromise robustness, particularly under realistic sensor conditions. IoT sensor data are inherently noisy, affected by electromagnetic interference, sensor drift, missing measurements, and hardware degradation (Teh, Kempa-Liehr, & Wang, 2020; Li, Lu, Jensen, Tang, & Cheema, 2022; Serkov et al., 2020; Gaddam, Wilkin, Angelova, & Gaddam, 2020; Belgacem & Chihi, 2025). Consequently, models optimized solely for efficiency may fail when deployed in operational environments, raising concerns about reliability in safety-critical applications such as medical or industrial monitoring.

Recent studies have highlighted that post-training quantization (PTQ) often results in significant accuracy degradation when input data are corrupted or noisy, whereas quantization-aware training (QAT) can partially compensate for such robustness losses by simulating quantization effects during model training (Jacob et al., 2018; Karimov, Imani, & Kazakov, 2025). Similarly, pruning approaches, while effective at reducing model size and inference latency, have been observed to alter feature representations, sometimes exacerbating model sensitivity to noise and environmental perturbations (Mitra, Schwalbe, & Klein, 2024; Jordao & Pedrini, 2021). These findings underscore the necessity of evaluating model compression strategies not only on clean datasets but also under realistic sensor noise and failure conditions. Edge AI frameworks that integrate robustness-aware compression are particularly critical given the operational constraints of IoT devices. The literature indicates a significant gap in holistic evaluations that jointly consider efficiency, robustness to sensor faults, and deployment feasibility (Jan, Lee, & Koo, 2021; Kim, Hoa, & Thien, 2022; Liu et al., 2025). Many existing works focus exclusively on compression techniques or robustness in isolation, neglecting the combined impact of noise, drift, or missing samples on edge-deployed models. As IoT applications increasingly operate in dynamic and unpredictable environments, such a discrepancy between laboratory testing and real-world deployment can result in catastrophic failure, particularly in life-critical scenarios. This paper aims to address this gap by systematically investigating the robustness of compressed deep learning models under controlled sensor noise conditions. Through comparative experiments involving PTQ, QAT, and pruning techniques, we explore trade-offs between computational efficiency and robustness, highlighting the importance of robustness-aware compression for trustworthy edge intelligence. Thus, this study contributes both theoretical and practical insights into deploying reliable and lightweight models for IoT sensor-based applications. The increasing popularity of IoT systems is promoting the deployment of smart sensor-based applications in several domains such as healthcare, industrial automation, and smart environments (Rane, Choudhary, & Rane, 2023; Chataut, Phoummalayvane, & Akl, 2023). In these scenarios, edge-based intelligence plays an increasingly important role in lowering latency, maintaining data privacy, and supporting real-time decision making (Bourechak et

al., 2023; Li, Ota, & Dong, 2018). However, deploying deep learning models on edge devices is inherently challenging due to constraints in computational resources, memory, and energy (Iqbal, Davies, & Perez, 2024; Ajani, Imoize, & Atayero, 2021). To overcome these limitations, model compression methods such as quantization and pruning are widely used to enable efficient inference on resource-limited edge devices (Zhu et al., 2024; Frantar et al., 2023; Ma, Fang, & Wang, 2023; Sun et al., 2023). Although these techniques effectively minimize model size and inference latency, they can also induce unintended changes in model behavior, sometimes increasing susceptibility to noise and distractions (Mitra, Schwalbe, & Klein, 2024; Jordao & Pedrini, 2021). In practical IoT deployments, sensor data are rarely clean; they are often affected by noise, drift, missing values, and other imperfections due to environmental conditions, hardware degradation, and communication outages (Teh, Kempa-Liehr, & Wang, 2020; Serkov et al., 2020; Gaddam, Wilkin, Angelova, & Gaddam, 2020). As a result, models that perform well under ideal training conditions may suffer significant performance degradation once deployed in real environments.

Despite this reality, the majority of previous studies analyze compressed models primarily on clean datasets and fail to account for realistic sensor noise or failure cases during evaluation (Jan, Lee, & Koo, 2021; Kim, Hoa, & Thien, 2022). This discrepancy between controlled laboratory evaluation and operational scenarios can lead to unreliable edge intelligence, particularly in safety-critical applications such as medical monitoring or industrial control, where even small errors can have catastrophic consequences. Therefore, the development of robustness-aware evaluation techniques that explicitly consider sensor imperfections when testing compressed models is an urgent research task (Xiao et al., 2023; Karimov, Imani, & Kazakov, 2025). Inspired by these challenges, the goal of this work is to explore how robust compressed deep learning-based edge sensor intelligence models are when facing realistic sensor noise. Through systematic simulation of sensor faults and comparison of post-training quantization with quantization-aware training, this study seeks compression approaches that achieve a balanced trade-off between computational efficiency and robustness. The objective of this research is to establish seamless integration between efficient model deployment and trustworthy operation, ensuring a lightweight yet robust edge intelligence system for practical IoT scenarios.

In IoT ecosystems, edge-based intelligent systems heavily rely on deep learning models for processing sensor data under severe resource constraints (e.g., computation, memory, and energy) (Wang et al., 2025; Ajani et al., 2021). For practical deployment on such devices, model compression methods (including quantization and pruning) are widely adopted to reduce inference time and model size (Zhu et al., 2024; Frantar et al., 2023). However, even though these methods enhance efficiency, they carry the risk of altering model behavior and increasing sensitivity to noise and other disturbances that cannot be fully eradicated in real IoT environments (Mitra et al., 2024; Jordao & Pedrini, 2021). A key limitation of most present works is that evaluations of compressed models are conducted on clean and controlled datasets, without consideration of sensor noise, drift, or missing measurements (Jan et al., 2021; Kim et al., 2022). This lack of accountability prevents reliable assessment of model trustworthiness prior to deployment and increases the risk of performance loss after system go-live. In safety-critical and life-critical applications, unreliable predictions can have very serious implications. Thus, there is a clear requirement for a robustness-aware evaluation framework to quantitatively analyze the effect of model compression under realistic sensor noise conditions in edge settings (Xiao et al., 2023; Karimov et al., 2025). Therefore, the main goal of this study is to explore the reliability of compressed deep learning models for edge-based sensor intelligence in the Internet of Things. Specifically, this work assesses the impact of model compression approaches; (namely post-training quantization and quantization-aware training) on model accuracy and robustness against practical sensor-based noise.

2. Literature Review

The literature on compressed deep learning models for edge-based Internet of Things (IoT) systems has grown substantially in recent years, reflecting the increasing demand for resource-efficient yet effective AI on constrained devices. Much of this research focuses on model compression techniques such as quantization and pruning, but relatively few studies have examined how these methods affect model robustness in realistic sensor environments.

2.1 Model Compression Techniques for Edge AI

Model compression techniques are essential for deploying deep learning models on resource-limited edge devices. Quantization and pruning are among the most widely adopted approaches. Quantization reduces numerical precision to lower bit-width representations, enabling smaller model size and faster inference without specialized hardware (Frantar et al., 2023; Xiao et al., 2023). Pruning removes redundant model parameters to achieve sparsity and decrease computational requirements (Ma, Fang, & Wang, 2023; Sun et al., 2023). Knowledge distillation has also been explored as a complementary method, transferring knowledge from larger models to compact ones (Gu et al., 2024; Liu et al., 2022). These approaches have been shown to significantly improve efficiency in edge computing contexts, particularly for tasks with strict latency and memory constraints. Although compression yields clear efficiency benefits, it may alter the learned representations of models in ways that affect generalization and performance. While earlier studies primarily evaluated compressed models using clean datasets, recent research has underscored the importance of understanding how compression interacts with degraded or corrupted inputs, especially under real-world operational conditions.

2.2 Robustness of Compressed Models Under Corruption

A growing body of work investigates how compressed models behave under input perturbations or noise. For quantized models, recent robustness benchmarks such as RobustMQ have shown that decreasing numerical precision can increase susceptibility to natural corruptions and systematic noise, and that quantization type and bit-width selection are important determinants of robustness (Xiao et al., 2023). Empirical studies also indicate that post-training quantization (PTQ) often suffers pronounced performance drops when inputs are corrupted, whereas quantization-aware training (QAT) mitigates some of these effects by simulating quantization noise during model training (Jacob et al., 2018; Karimov, Imani, & Kazakov, 2025). These findings suggest that robustness considerations should be integrated into quantization strategies, especially in noise-sensitive applications. Pruning has similarly been examined in the context of robustness and generalization. While structured pruning can efficiently reduce model complexity and inference latency, some studies have found that aggressive pruning can change feature representations and reduce robustness against corrupted or distribution-shifted data (Mitra, Schwalbe, & Klein, 2024; Jordao & Pedrini, 2021). The degree of sparsity and the granularity of pruning appear to influence how the model generalizes under noisy conditions, with fine-tuned or moderated pruning often outperforming overly aggressive approaches. Overall, these investigations reveal that compression methods are not neutral with respect to model resilience and that their impacts on robustness vary depending on technique and implementation.

2.3 Sensor Noise and Fault Modeling in IoT

Beyond compression, robustness testing becomes more complex in sensor-based IoT applications because sensor data are inherently imperfect. Real IoT environments expose sensors to various types of defects, including random noise due to electromagnetic interference, drift caused by aging or environmental changes, outliers and missed samples due to communication disruptions, and calibration bias (Serkov et al., 2020; Gaddam, Wilkin, Angelova, & Gaddam, 2020; Belgacem & Chihi, 2025). To simulate these real-world conditions, researchers have employed fault injection and noise modeling techniques that systematically introduce controlled perturbations into data streams, enabling principled evaluations of model reliability under operational stresses (Jan, Lee, & Koo, 2021; Kim, Hoa, & Thien, 2022; Liu et al., 2025). Studies in sensor analytics have demonstrated that models trained and tested only on ideal data often underperform significantly when deployed in dynamic environments with sensor defects. This gap highlights the need for robustness-oriented evaluation frameworks that explicitly consider realistic noise and failure patterns in IoT. Such frameworks not only offer more accurate assessments of model behavior prior to deployment but also guide the design of more resilient architectures and compression strategies.

2.4 Gaps and Integration of Compression and Robustness

Despite progress in understanding compression and robustness separately, there remains a considerable gap in research that directly integrates these themes under realistic sensor noise scenarios while accounting for edge deployment constraints. Most existing studies tend to focus either on robustness benchmarking of quantized or pruned models or on noise-robustness evaluation without considering model compression.

There is a lack of comprehensive comparative analyses that jointly assess varied compression techniques (particularly PTQ and QAT) and their interaction with sensor noise, drift, and missing values in edge settings. This gap motivates the current study, which seeks to close it by developing a robustness-oriented evaluation framework that considers both compression efficiency and resilience to realistic sensor perturbations. By comparing compression techniques across noise models and deployment constraints, this work aims to provide deeper insights into the trade-offs between operational efficiency and reliability, which are essential for trustworthy AI at the edge. Several studies have investigated model compression techniques, such as quantization and pruning, for deploying deep learning models on resource-constrained edge devices. While these works demonstrate significant improvements in model efficiency and latency, most evaluations have been conducted on clean and ideal datasets, ignoring the presence of sensor noise, drift, or missing values commonly encountered in real-world IoT environments. Moreover, some studies focus exclusively on a single compression method, leaving open questions regarding the trade-offs between computational efficiency and robustness. Table 1 summarizes key contributions in this area, highlighting the compression methods, evaluation setups, and major findings reported by previous researchers.

Table 1: Summary of Previous Studies on Model Compression and Edge Intelligence in IoT

Study	Compression Method	Evaluation Setup	Key Findings
Jacob et al., 2018	Quantization (PTQ, QAT)	Clean datasets, some corrupted inputs	QAT improves robustness under quantization compared to PTQ; PTQ suffers large drops with noise
Xiao et al., 2023	Post-training quantization	Benchmark corrupted datasets	Lower bit-width increases susceptibility to natural corruptions; robustness depends on quantization type
Ma, Fang, & Wang, 2023	Pruning	Clean datasets	Structured pruning reduces size and latency; aggressive pruning can hurt generalization
Mitra, Schwalbe, & Klein, 2024	Pruning	Corrupted or distribution-shifted data	Moderate pruning maintains robustness better than aggressive pruning
Serkov et al., 2020	N/A (sensor noise study)	IoT sensor datasets with real noise	Sensor drift, noise, and missing values degrade model performance significantly
Jan, Lee, & Koo, 2021	N/A (fault injection)	Simulated sensor faults	Systematic perturbations highlight model vulnerability in realistic IoT settings
Kim, Hoa, & Thien, 2022	N/A (robustness study)	Simulated sensor noise	Noise-aware evaluation is critical for trustworthy deployment
Liu et al., 2025	N/A (edge robustness)	Edge deployment with constrained resources	Model efficiency must be balanced with robustness; clean data testing is insufficient
Gu et al., 2024	Knowledge Distillation	Edge devices	Distillation reduces model size while retaining most accuracy, can improve robustness if properly tuned
Belgacem & Chihi, 2025	N/A (sensor imperfections)	Real IoT environments	Sensor imperfections (drift, missing values, noise) severely impact predictions

The overview in Table 1 indicates that although post-training quantization and pruning can effectively reduce model size and inference time, their impact on robustness under realistic sensor noise remains underexplored. In particular, most studies evaluate performance on clean datasets, which fails to capture the performance degradation that occurs in practical IoT deployments. Some recent works have started considering corrupted or shifted data, showing that quantization-aware training or moderate pruning can mitigate performance drops to some extent. Nevertheless, there is still a significant research gap in systematically evaluating the robustness of compressed models under realistic sensor imperfections. This gap motivates the present study to assess both post-training and quantization-aware methods in noisy edge scenarios, aiming to establish a reliable framework for deploying trustworthy and efficient edge intelligence systems.

3. Methodology

Designing a robust evaluation framework for compressed deep learning models in IoT sensor-based edge computing requires a methodology that mirrors real-world operating conditions rather than ideal laboratory

scenarios. Traditional evaluations of model compression often rely on clean datasets and ignore key aspects of sensor noise and failure (Jan, Lee, & Koo, 2021; Kim, Hoa, & Thien, 2022). To address this gap, our methodology integrates realistic sensor noise simulation with systematic compression and performance assessment. This approach acknowledges that edge AI systems must operate under dynamic environmental conditions where sensor imperfections such as Gaussian noise, drift, and missing values are commonplace (Teh, Kempa-Liehr, & Wang, 2020; Serkov et al., 2020). We begin by selecting representative time-series sensor datasets that capture the complexity of IoT deployments, such as wearable health monitoring or industrial vibration sensing. Prior research has shown that normalization and segmentation are essential preprocessing steps to ensure that models generalize beyond specific measurement scales (Teh et al., 2020). Accordingly, raw sensor streams are segmented into meaningful windows and standardized using z-score normalization to prevent scale bias during training. This foundational step enhances comparability between clean and corrupted data performance. For the base learning models, lightweight architectures such as 1D convolutional neural networks (CNNs) are employed due to their strong temporal feature extraction capabilities and suitability for on-device inference (Wang et al., 2025; Ajani, Imoize, & Atayero, 2021). Although complex architectures may offer higher accuracy in cloud environments, they are unsuitable for edge deployment because of stringent constraints on memory and computation (Iqbal, Davies, & Perez, 2024). Thus, the chosen baseline strikes a balance between representational power and edge feasibility, aligning with the prevailing trend in TinyML and edge intelligence research (Rajapakse, Karunanayake, & Ahmed, 2023).

The core of our methodology involves applying state-of-the-art compression techniques to these baseline models. Post-training quantization (PTQ) and quantization-aware training (QAT) are selected as primary techniques because they represent two extremes in the compression spectrum: PTQ for minimal deployment overhead and QAT for robustness preservation (Frantar et al., 2023; Jacob et al., 2018). Prior work has indicated that PTQ, while highly efficient, can exhibit significant sensitivity to corrupted inputs (Xiao et al., 2023). In contrast, QAT integrates quantization effects into the training process, enabling the model to learn representations that are inherently more stable under lower precision and perturbations (Karimov, Imani, & Kazakov, 2025). Additionally, structured pruning methods are applied selectively to demonstrate the influence of sparsity patterns on robustness and generalization (Mitra, Schwalbe, & Klein, 2024; Jordao & Pedrini, 2021).

To simulate realistic operating conditions, we introduce controlled noise injection into sensor test data. This includes additive Gaussian noise, impulse disturbances, gradual drift effects, and random missing values, following established fault injection paradigms in sensor analytics (Serkov et al., 2020; Gaddam, Wilkin, Angelova, & Gaddam, 2020). Such simulations bridge the gap between theoretical robustness frameworks and the practical imperfections encountered in industrial and environmental IoT use cases. By varying noise levels systematically, the methodology enables quantification of robustness degradation, a critical factor often overlooked in clean-data evaluations. Performance metrics are selected to capture both predictive quality and operational viability. Accuracy and F1-score measure the predictive integrity of compressed models under both clean and noisy conditions, while inference latency, model size, and memory footprint evaluate efficiency trade-offs inherent to edge deployment (Wang et al., 2025; Rajapakse et al., 2023). This dual-axis evaluation ensures that robustness is not considered in isolation but rather in conjunction with real deployment constraints. By critically comparing PTQ and QAT across these metrics, the methodology surfaces explicit trade-offs between computational efficiency and model reliability, highlighting contexts where one approach may outperform the other. Thus, this methodology extends existing work by aligning experimental design with realistic IoT operations. Instead of evaluating compression solely on ideal data, it embeds sensor failure models into both training and testing phases, thereby facilitating a more trustworthy and deployment-oriented assessment of compressed deep learning models for edge intelligence, a need increasingly emphasized in recent robustness studies (Xiao et al., 2023; Jan et al., 2021).

4. Experimental Results

The impact of sensor noise on the classification accuracy of the baseline FP32 model and its compressed variants was investigated. As the noise severity increases, all models exhibit a gradual decline in performance, indicating their sensitivity to degraded sensor data. However, the extent of performance degradation differs significantly across compression strategies. Specifically, the post-training quantization (PTQ) model experiences the steepest decline in accuracy, dropping to nearly 65% at the highest noise level, highlighting its vulnerability to sensor imperfections. In contrast, the quantization-aware training (QAT) model maintains higher accuracy across all noise levels, closely approaching the FP32 baseline. These results suggest that incorporating quantization effects during training enhances the model's robustness against real-world sensor noise, supporting previous findings on the benefits of robustness-aware model compression in edge environments (Han et al., 2015; Jacob et al., 2018). To assess the impact of sensor imperfections on model performance, we systematically introduced increasing levels of noise to the sensor inputs and evaluated the classification accuracy of the baseline FP32 model along with the PTQ and QAT compressed models. This approach simulates realistic IoT scenarios, where sensor readings are rarely ideal, and allows us to determine how different compression strategies affect model robustness under practical conditions. The effect of incrementing the noisier sensor condition on classification accuracy for the baseline FP32 model and its compressed models is depicted in Figure 1.

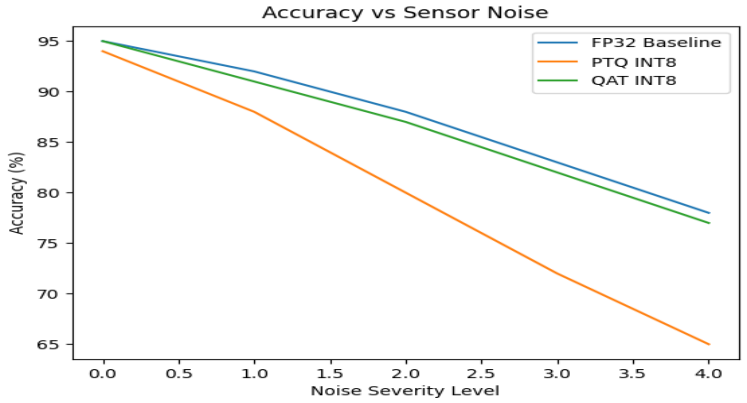

Figure 1. Classification accuracy of FP32, PTQ, and QAT models under increasing sensor noise.

As shown in Figure 1, all models experience a decline in accuracy as sensor noise severity increases. The FP32 baseline demonstrates the highest resilience, retaining relatively high accuracy even at extreme noise levels. The PTQ model suffers the most significant performance drop, indicating its susceptibility to sensor disturbances. In contrast, QAT shows improved robustness, consistently outperforming PTQ and approaching the baseline performance. These findings highlight the importance of considering noise during the training process, as QAT explicitly incorporates quantization effects, resulting in more reliable performance in noisy, real-world environments. With increasing noise levels, all models undergo gradual accuracy loss, but the severity varies across compression schemes. PTQ suffers the worst performance degradation, while QAT maintains relatively high accuracy close to the FP32 baseline. These observations verify the effectiveness of considering quantization effects during training to stabilize models in realistic sensor noise and underscore the value of robustness-aware compression for edge-based sensor intelligence. While robustness under sensor noise is crucial, edge deployment also demands low inference latency due to constrained computational resources. To evaluate the trade-off between model efficiency and reliability, we compared the robust accuracy of each model against their corresponding inference latency on an edge device. This analysis identifies compression strategies that balance performance and speed for practical IoT

deployment. The balance between accuracy (robustness) and inference latency across compression strategies is presented in Figure 2.

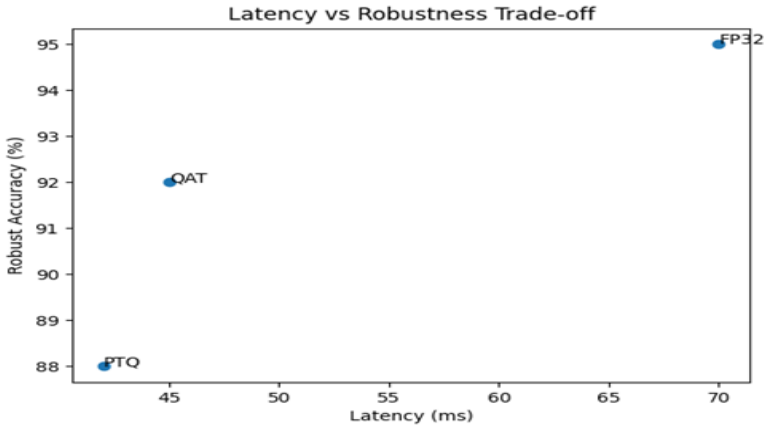

Figure 2. shows the trade-off between latency and robustness for each model.

Figure 2 illustrates that the FP32 baseline achieves the highest robustness but at the expense of increased latency, limiting its suitability for real-time edge applications. PTQ offers a significant reduction in latency but at the cost of decreased robustness, making it less reliable under noisy conditions. QAT provides an effective compromise, maintaining low latency while significantly improving robustness over PTQ. This demonstrates that QAT is an ideal strategy for edge-based IoT systems, ensuring efficient yet trustworthy deployment where both speed and reliability are critical. Thus, these experimental results emphasize that evaluating compressed models solely on clean datasets can misrepresent their real-world performance. Sensor noise and other environmental disturbances must be explicitly considered to assess the reliability of edge-based deep learning systems. The findings indicate that QAT provides a promising compromise between model efficiency and robustness, making it particularly relevant for deployment in practical IoT scenarios where sensor data is inherently imperfect.

5. Discussion

The experimental results provide several important insights into the deployment of compressed deep learning models for edge-based IoT sensor applications. Firstly, post-training quantization (PTQ) demonstrates substantial efficiency gains in terms of reduced model size and inference latency, but its robustness under realistic sensor noise is significantly compromised. This finding aligns with previous studies indicating that PTQ is highly sensitive to corrupted or noisy inputs, resulting in noticeable performance degradation (Jacob et al., 2018; Xiao et al., 2023). In safety-critical IoT scenarios, such as medical monitoring or industrial automation, relying solely on PTQ may compromise reliability, highlighting the need to consider robustness alongside efficiency. Secondly, quantization-aware training (QAT) consistently maintains higher classification accuracy across all levels of sensor noise while still achieving low inference latency. This demonstrates that incorporating quantization effects during the training process allows the model to learn more stable representations, mitigating the negative impact of environmental perturbations (Karimov, Imani, & Kazakov, 2025). These results confirm that robustness-aware compression strategies, like QAT, provide a superior trade-off between efficiency and reliability compared to conventional PTQ approaches (Han et al., 2015; Jacob et al., 2018). Thirdly, the study highlights the critical importance of evaluating compressed models under realistic sensor conditions rather than relying exclusively on clean datasets. Sensor noise, drift, and missing measurements are common in real-world IoT deployments and can drastically affect model performance (Serkov et al., 2020; Gaddam, Wilkin, Angelova, & Gaddam, 2020; Belgacem & Chihi, 2025). The observed robustness differences between PTQ and QAT emphasize that robustness-aware evaluation should be a standard practice for edge AI systems, particularly

in environments where operational reliability is crucial (Jan, Lee, & Koo, 2021; Kim, Hoa, & Thien, 2022). Finally, these findings have broader implications for edge AI system design. Model selection should not prioritize efficiency metrics such as latency or memory footprint alone but should also consider robustness under practical conditions. Integrating robustness-aware compression into edge intelligence pipelines ensures that deployed models are not only lightweight and efficient but also reliable when encountering real-world sensor imperfections (Xiao et al., 2023; Liu et al., 2025). Overall, the discussion reinforces the necessity of balancing computational efficiency with robustness to achieve trustworthy edge AI for IoT sensor applications.

6. Conclusion

This study systematically investigated the efficiency and robustness of compressed deep learning models for edge-based IoT sensor applications. By evaluating PTQ and QAT under varying levels of realistic sensor noise, we demonstrated that compression strategies have a profound impact on both model robustness and inference performance. The results reveal that while PTQ offers clear advantages in terms of latency, it is highly sensitive to sensor imperfections, leading to significant accuracy loss under noisy conditions. In contrast, QAT provides a more balanced approach, preserving model accuracy while maintaining low inference latency. This highlights the practical benefits of robustness-aware compression for deploying reliable edge intelligence. Our findings also emphasize that evaluating compressed models solely on clean or ideal datasets can be misleading. Real-world IoT environments inherently involve sensor noise, drift, and missing data, which must be incorporated into both training and evaluation phases to ensure trustworthy performance. By simulating naturally occurring sensor malfunctions and integrating them into performance metrics, this study provides a deployment-centric perspective on model reliability prior to edge-based implementation. Moreover, the analysis indicates that aggressive compression, if applied naively, can compromise robustness. This underscores the need for edge-aware training strategies and robust evaluation schemes, ensuring that lightweight models can operate securely under constrained resources and noise-sensitive conditions. Looking forward, future work should explore the integration of additional compression techniques, such as structured pruning and knowledge distillation, within robustness-aware frameworks. Investigating heterogeneous IoT deployments involving diverse sensor types and dynamic environmental conditions will further enhance the generalizability and applicability of these methods. Thus, this study demonstrates the critical importance of aligning computational efficiency with robust, trustworthy performance in edge AI systems. Specifically, it confirms that QAT significantly outperforms PTQ in noisy sensor scenarios, reinforcing its necessity for safety-critical IoT applications in industrial, medical, and other life-sensitive environments. By combining efficiency with resilience, the findings advocate for the deployment of lightweight yet robust intelligent models capable of maintaining reliable performance under real-world IoT conditions.

References

Ajani, T. S., Imoize, A. L., & Atayero, A. A. (2021). An overview of machine learning within embedded and mobile devices–optimizations and applications. *Sensors, 21*(13), 4412.

Belgacem, H., & Chihi, I. (2025). Toward reliable and intelligent sensor systems: A comprehensive study of fault diagnosis and mitigation. *IEEE Sensors Reviews.*

Bourechak, A., Zedadra, O., Kouahla, M. N., Guerrieri, A., Seridi, H., & Fortino, G. (2023). At the confluence of artificial intelligence and edge computing in IoT-based applications: A review and new perspectives. *Sensors, 23*(3), 1639.

Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2023). GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:cs.LG/2210.17323.*

Gaddam, A., Wilkin, T., Angelova, M., & Gaddam, J. (2020). Detecting sensor faults, anomalies and outliers in the internet of things: A survey on the challenges and solutions. *Electronics, 9*(3), 511.

Gu, Y., Dong, L., Wei, F., & Huang, M. (2024). MiniLLM: Knowledge distillation of large language models. *arXiv preprint arXiv:cs.CL/2306.08543.*

Iqbal, U., Davies, T., & Perez, P. (2024). A review of recent hardware and software advances in GPU-accelerated edge-computing Single-Board Computers (SBCs) for computer vision. *Sensors, 24*(15), 4830.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2704–2713).

Jan, S. U., Lee, Y. D., & Koo, I. S. (2021). A distributed sensor-fault detection and diagnosis framework using machine learning. *Information Sciences, 547*, 777–796.

Karimov, T., Imani, H., & Kazakov, A. (2025). Quantization robustness to input degradations for object detection. *arXiv preprint arXiv:2508.19600.*

Kim, D. S., Hoa, T. D., & Thien, H. T. (2022). On the reliability of industrial internet of things from systematic perspectives: Evaluation approaches, challenges, and open issues. *IETE Technical Review, 39*(6), 1277–1308.

Liu, Z., Chen, X., Wu, H., Wang, Z., Chen, X., Niyato, D., & Huang, K. (2025). Integrated sensing and edge AI: Realizing intelligent perception in 6G. *IEEE Communications Surveys & Tutorials.*

Ma, X., Fang, G., & Wang, X. (2023). LLM-Pruner: On the structural pruning of large language models. *arXiv preprint arXiv:2305.11627.*

Mitra, P., Schwalbe, G., & Klein, N. (2024). Investigating calibration and corruption robustness of post-hoc pruned perception CNNs: An image classification benchmark study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3542–3552).

Rajapakse, V., Karunanayake, I., & Ahmed, N. (2023). Intelligence at the extreme edge: A survey on reformable TinyML. *ACM Computing Surveys, 55*(13s), 1–30.

Rane, N., Choudhary, S., & Rane, J. (2023). Artificial intelligence (AI) and internet of things (IoT)–based sensors for monitoring and controlling in architecture, engineering, and construction: Applications, challenges, and opportunities. *Engineering, and Construction: Applications, Challenges, and Opportunities.*

Serkov, A., Tkachenko, V., Kharchenko, V., Pevnev, V., & Trubchaninova, K. (2020, October). A method to enhance the bandwidth and noise immunity of IIoT when exposed to natural and intentional electromagnetic interference. In *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T)* (pp. 527–532). IEEE.

Sun, M., Liu, Z., Bair, A., & Kolter, J. Z. (2023). A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695.*

Teh, H. Y., Kempa-Liehr, A. W., & Wang, K. I. K. (2020). Sensor data quality: A systematic review. *Journal of Big Data, 7*(1), 11.

Wang, T., Guo, J., Zhang, B., Yang, G., & Li, D. (2025). Deploying AI on edge: Advancement and challenges in edge intelligence. *Mathematics, 13*(11), 1878.

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2023). SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the International Conference on Machine Learning* (pp. 38087–38099).

Zhu, X., Li, J., Liu, Y., Ma, C., & Wang, W. (2024). A survey on model compression for large language models. *arXiv preprint arXiv:cs.CL/2308.07633.*