Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 3, 2026

DEVELOPMENT OF ENERGY EFFICIENT ARTIFICIAL INTELLIGENCE MODELS FOR EDGE COMPUTING BASED REAL TIME APPLICATIONS IN LOW POWER AND RESOURCE LIMITED IOT ENVIRONMENT

Nisar Muhammad

MS Scholar, Department of Computer Science, Sarhad University of Science and Information Technology (SUIT) Peshawar (KPK), Pakistan

nisarmuhammad85@gmail.com

DOI: https://doi.org/10.5281/zenodo.21803224

Keywords
Edge Computing, Energy-Efficient Artificial Intelligence, Internet of Things, Lightweight Deep Learning, Model Optimization, Resource-Constrained Systems, Real-Time Applications

Article History
Received: 01 February 2026
Accepted: 17 March 2026
Published: 31 March 2026

Corresponding Author: *
Nisar Muhammad

Abstract
The rapid expansion of Internet of Things (IoT) applications has created a growing demand for real-time intelligent processing; however, conventional cloud-based artificial intelligence (AI) systems face challenges related to high latency, excessive energy consumption, bandwidth limitations, and dependency on continuous connectivity. This research addresses the problem of developing energy-efficient AI models capable of operating effectively in low-power and resource-constrained IoT environments through edge computing-based architectures. The study aims to design and evaluate optimized AI models that balance computational performance, energy efficiency, and real-time decision-making capabilities for edge-enabled IoT applications. A systematic methodology was adopted involving lightweight deep learning architectures, model compression techniques, quantization approaches, and performance evaluation using energy consumption, processing latency, memory utilization, and prediction accuracy as key metrics. Experimental analysis demonstrates that optimized edge AI models significantly reduce computational overhead and energy requirements while maintaining competitive accuracy levels compared with traditional AI models deployed on centralized cloud platforms. The findings reveal that techniques such as model pruning, efficient neural network structures, and adaptive resource management enhance the sustainability and scalability of IoT-based intelligent systems. The research highlights the importance of integrating energy-aware AI algorithms with edge computing frameworks to support real-time applications in smart cities, healthcare monitoring, industrial automation, and environmental sensing. The proposed approach provides a practical pathway for developing sustainable, autonomous, and efficient AI-driven IoT ecosystems, particularly in environments where computational resources and energy availability are limited.

Introduction

With its high-speed development, Industry 4.0 has revolutionized industrial systems through the application of artificial intelligence (AI), Internet of Things (IoT), cyber-physical systems and intelligent automation technologies. From boosting efficiency to optimizing resources, minimizing production failures, and enabling real-time decision-making, AI-powered solutions are increasingly being embraced by industries for their operational advantages. Today, there is a massive amount of real-time data being generated as a result of the massive deployment of IoT sensors and connected industrial devices. But the conventional cloud-based AI architectures have some drawbacks such as communication delays, high bandwidth requirement, reliance on stable network connection, data privacy issue and high energy consumption due to continuous data transmission to the centralized servers (Liang et al., 2020; Tuli et al., 2022).

In industrial applications like predictive maintenance, automated quality inspection, machine fault diagnostics and intelligent energy management, quick data processing and fast reaction are essential. In all such cases, cloud processing might increase latency and compromise the reliability and productivity of operations. Thus, industries need decentralized computational methods that can provide intelligent computing near the data sources. To overcome this challenge, Edge Computing has been introduced as a promising approach that enables data processing at the edge of the network, and minimizes reliance on remote cloud infrastructures (Zhu et al., 2022).

Edge computing is a decentralized computing model that degrades computations to the edge of the network, utilizing edge devices, gateways or local servers to close to the place of data creation. The combination of AI and edge computing has ushered in the era of edge intelligence, where machine learning models can be deployed in resource-limited environments. Edge AI systems can lower cost of communication, speed up data processing and increase the accuracy of real-time applications in industrial environments, in contrast with cloud-based systems (Liang et al., 2020).

For industrial IoT, edge AI can enable intelligent monitoring, autonomous control, and predictive analytics by processing sensor-generated data at the edge. For instance, using edge-based machine learning algorithms, it's possible to detect unusual equipment behaviour, forecast maintenance needs, and optimise production processes without moving all operational data to external servers. Some recent studies suggest that edge AI systems can enhance the scalability and efficiency of AI systems, thereby lessening network congestion and enabling quick decisions in industrial settings (Iftikhar et al., 2022).

Deploying edge AI, however, will face huge limitations in terms of computation and energy. Typical IIoT devices have limited processing powers, storage space and power. Creating efficient AI models that can remain highly predictive but require less computing power, therefore, is a great challenge for researchers.

While AI has proven to be a powerful tool in industrial automation, the growing need for computational power raises energy concerns. Many deep learning architectures are very large, with millions of parameters and demand a lot of computing power and are hard to deploy on low power edge devices. If an algorithm requires too many computations, it is likely to cost more to operate, impact device life and will restrict the scalability of an Industrial IoT system (Fanariotis et al., 2023).

Therefore, energy-efficient artificial intelligence becomes an important line of research, where the goal is to have a low computational complexity and at the same time maintain an acceptable accuracy. In contrast to traditional AI development methods, which focus mainly on model performance, energy-aware AI centers on minimizing power usage, processing requirements, and hardware usage. However, recent research has emphasized the need for the sustainable deployment of AI while ensuring high predictiveness and computational efficiency, especially in large-scale IoT deployments (Zhu et al., 2022).

There are a number of optimisation strategies that have been suggested for making AI more efficient for use on edge devices. These include model pruning, quantization, knowledge distillation, lightweight neural network architecture and adaptive inference mechanisms. These techniques can help to compress the model, lower memory usage, and decrease processing demands, allowing AI applications to function effectively in resource-constrained settings (Fanariotis et al., 2023).

There are specific challenges for developing industrial economies to implement advanced AI-driven industrial technologies, such as a lack of infrastructure, energy limitations, and low implementation costs. Despite the shift towards digitalization, the industrial sector in Pakistan still has a long way to go before the full advantages of intelligent automation can be realized due to technological and financial constraints.

Lahore is one of the big industrial centers of Pakistan which includes a wide range of industries including textile industry, pharmaceutical, food processing, electronics and engineering. As these industries integrate automated machinery, sensor-based monitoring systems and digital production technologies, they are increasingly employing digital tools. But, there are still many industrial units using traditional monitoring methods or cloud-based solutions that may not be able to offer real-time processing efficiently.

Adopting energy-efficient edge AI models can offer a viable solution for industries in Lahore, facilitating localized intelligence, cloud-free operations, cost savings, and enhanced industrial sustainability. These systems are especially beneficial for small to medium scale enterprises where computing power and energy resources are still key limitations.

With the growing proliferation of industrial IoT systems, there is a rising need for a real time decision making system with low energy consumption by using Artificial Intelligence models. But the state-of-the-art AI architectures are designed for high-performance computing systems and are not practical for most industrial edge devices because they have too much complexity, memory, and energy usage.

Additionally, traditional cloud-based AI systems present challenges such as higher communication delays, need for continuous network connectivity, and extra energy usage for data transmission. The constraints make it difficult to use AI applications that demand real-time responses, like industrial fault detection, predictive maintenance systems, and automated control systems.

These difficulties are further complicated in Lahore-based industrial environments because of lack of computational infrastructure, energy shortages and financial restrictions. Any industry that wants to use AI to process industrial data locally with accuracy and efficiency, without needing to transfer it to the cloud, will need affordable and scalable AI solutions. But very little work has been done to explore the practical development and testing of energy efficient AI models in such resource-limited industrial environments.

The present study, therefore, is aimed at addressing the lack of optimized AI models that can ensure high predictive performance while simultaneously consuming low energy and also operating in real time in edge computing-based industrial IoT environments.

Research Gap

While impressive advancements have been made in the field of industrial AI and edge computing, there are still several gaps that should be addressed.

First, research to date has largely been concerned with accuracy and automation of AI predictions and systems, with less attention paid to energy efficiency and sustainable deployment. Though some of the advanced AI models operate with high accuracy, they are not as efficient as low power industrial devices (Fanariotis et al., 2023).

Second, the test sets used in previous studies are often standard test sets or simulated environments, but are not necessarily the actual conditions of the industrial operational environment. This imparts restrictions on their practical implementation in manufacturing environments where system performance is affected by the availability of energy, hardware limitations and variability in energy supply.

Third, studies on energy-efficient edge AI have focused mainly on the technologically advanced areas, and there is a limited amount of evidence of the energy efficiency in developing industrial settings like in Pakistan. AI frameworks need to be tailored for local industries, taking into account factors such as affordability, infrastructure constraints, and energy issues.

Fourthly, although the optimization techniques like pruning, quantization and knowledge distillation have shown to be effective, little studies have addressed their combined effects on industrial metrics, such as energy savings, processing latency, memory usage, and predictive accuracy.

To overcome these challenges, this study proposes and tests optimized AI models with industrial IoT data from manufacturing units in Lahore.

Research Objectives

The study aims to develop energy-efficient artificial intelligence models for edge computing-based real-time applications in low-power and resource-limited IoT environments.

The specific objectives are:

1. To develop lightweight AI models suitable for deployment on industrial edge devices.
2. To optimize AI architectures through model compression and energy-efficient learning techniques.
3. To evaluate optimized AI models based on accuracy, processing latency, memory utilization, and energy consumption.
4. To compare the performance of edge-based AI models with conventional cloud-based AI approaches.
5. To investigate the applicability of energy-efficient AI solutions in Lahore-based industrial environments.
6. To propose an energy-efficient framework for sustainable industrial IoT applications.

Research Questions

The study addresses the following research questions:

RQ1: How can artificial intelligence models be optimized for deployment in low-power and resource-constrained industrial edge environments?

RQ2: What impact do AI optimization techniques have on energy consumption and computational efficiency of edge-based industrial IoT systems?

RQ3: To what extent do energy-efficient edge AI models maintain predictive accuracy compared with conventional cloud-based AI models?

RQ4: How does the deployment of lightweight AI architectures influence processing latency and memory utilization in industrial IoT applications?

RQ5: What role can energy-efficient edge AI frameworks play in improving sustainable digital transformation within Lahore-based industries?

This study falls within the new research area of sustainable Artificial Intelligence and aims at the problem of deployment of intelligent systems in resource constrained industrial environments. The study also offers theoretical contributions by strengthening the current understanding of energy-aware AI optimization and the edge intelligence in Industrial IoT ecosystems.

Industry, in practice, will find the research of great value when considering cost-effective and energy efficient solutions for automation. The proposed framework can help industries to decrease computational costs, decrease the amount of energy used, implement intelligent monitoring systems, and operate with real-time intelligence.

This study is timely for the industrial sector of Lahore as it offers recommendations based on empirical evidence for the implementation of AI technologies in the context of limited infrastructure and resources in the city. The combination of lightweight AI models and edge computing architectures contributes to the creation of efficient, sustainable, and scalable industrial IoT environments.

Literature Review

Industrial Internet of Things (IIoT) has emerged in a fast-growing area, where there is a need for large-scale industrial data processing in real time with efficient computational frameworks. However, the constant transfer of data between devices and centralized servers introduces issues such as latency, bandwidth usage, energy needs, and reliance on the network with traditional cloud-based AI systems. The concept of edge computing is proposed to mitigate these drawbacks by performing data processing close to the data generation sources which results in faster response, reliability, and scalability of industrial applications (Zhou et al., 2021; Shi et al., 2020).

Edge computing enables applications in industrial use cases like predictive maintenance, automated monitoring, quality control, and intelligent energy management. Edge-based systems can process sensor data locally, thereby minimizing communication overhead and supporting real-time decision-making, which makes them applicable to industrial applications.The local processing of sensor data by the edge-based systems reduces the communication overhead and enables faster decision-making, which makes them applicable in industrial applications (Tuli et al., 2022).

AI is a significant technology in the realm of industrial automation, offering predictive analytics, anomaly detection, and autonomous decision-making capabilities. Machine learning and deep learning models are used to analyze complex industrial data captured by sensors and connected devices, thus enhancing operational efficiency and minimizing production failures (Wang et al., 2021).

But traditional AI models tend to consume a lot of power, which can be a challenge to run on edge devices with limited power. One of the key challenges in the application of complex AI models in industrial IoT systems is their limited processing power, memory capacity and energy supply (Zhang et al., 2022). Thus, light and optimized AI architectures are needed to achieve efficient edge deployment.

One of the major problems of AI-powered industrial systems is energy consumption. Deep learning models demand tremendous computational tasks, leading to high energy use and expenses. In the field of AI, Green AI strategies focus on minimizing the use of computational resources and optimizing algorithms and model performance (Schwartz et al., 2020).

For industrial IoT devices which have power and hardware constraints, energy-efficient AI models are especially significant. To deploy them efficiently on the edge devices, one can use optimization methods like model pruning, quantization, and knowledge distillation, which are used to simplify the model.One can simplify the model using optimization techniques such as model pruning, quantization, and knowledge distillation, to deploy them efficiently on the edge devices (Gou et al., 2021).

The goal of lightweight AI models is to deliver strong predictive results while minimizing computational load. Model pruning removes unused parameters, knowledge distillation distills knowledge from large models to smaller ones, and quantization reduces memory consumption. These techniques are effective for boosting processing speed, reducing energy consumption, and making AI deployment more feasible on resource-constrained devices (Cheng et al., 2020; Li et al., 2021).

Recent works show that the optimized AI models perform effectively in the industrial real-time application such as equipment monitoring, fault prediction, and automatic inspection with acceptable accuracy rates (Zhang et al., 2022).

The edge AI opportunity for industry is huge, as it allows for cost-effective and eco-friendly automation. Predictive maintenance, intelligent manufacturing, real-time quality assessment are all examples of applications that can help boost productivity and lower operational expenses.

Limited infrastructure, energy issues, and the high cost of technology are a major challenge for industrial digital transformation in developing economies, such as Pakistan. As a key industrial node, Lahore demands scalable AI solutions that can function with limited resources. The solution to this problem is Edge AI, which minimizes reliance on costly cloud resources and empowers local industrial intelligence.

While some previous works have explored the development of edge computing and AI optimization, there are still issues to be resolved. Studies conducted so far tend to concentrate on the accuracy of the model and not its use in the real world or energy efficiency. Many of the studies are conducted on benchmark data rather than on real industrial data, which affects their applicability for actual operations.

Also, the literature is scarce on the energy efficiency perspective of the adoption of edge AI in developing industrial settings like Pakistan.

The overall impact of AI optimization on energy use, processing time, memory usage and industrial productivity still needs to be examined. This study aims to fill these gaps by designing and testing energy-efficient AI models for Industrial IoT applications in Lahore.

Methodology

This study adopted a quantitative experimental research design to develop and evaluate energy-efficient artificial intelligence models for edge computing-based industrial IoT applications. The research focused on examining how AI optimization techniques can improve computational efficiency, reduce energy consumption, and maintain predictive accuracy within low-power and resource-constrained environments. A performance-based evaluation framework was developed to compare optimized edge AI models with conventional AI approaches. The methodology combined industrial IoT data analysis, AI model development, optimization techniques, and performance evaluation. The research framework consisted of four major phases: (1) industrial data acquisition, (2) AI model development, (3) energy-efficient model optimization, and (4) performance assessment.

Industrial Context and Study Area

The study was conducted using industrial IoT data representing selected manufacturing environments in Lahore, Pakistan. Lahore was selected because it represents one of Pakistan's major industrial centers with diverse manufacturing activities, including textile production, engineering, pharmaceuticals, food processing, and electronics.

The selected industrial environments utilize sensor-based monitoring systems for tracking operational parameters such as machine temperature, vibration levels, power consumption, production speed, and equipment performance. These industrial parameters provide suitable datasets for developing AI-based predictive and monitoring applications.

The research focused on resource-constrained industrial settings where computational efficiency and energy optimization are essential due to limitations in hardware capacity, energy availability, and operational costs.

Sample and Data Collection

A dataset was developed from industrial IoT operational records collected from **120 industrial monitoring devices** installed across selected manufacturing units in Lahore. The dataset represented real-time industrial conditions and included operational measurements collected through IoT-enabled sensors.

The collected variables included:

Parameter	Description
Temperature	Machine operating temperature measurements
Vibration	Equipment vibration intensity
Energy Consumption	Power utilization patterns
Machine Status	Normal and abnormal operational conditions
Processing Time	AI inference response time
Production Output	Operational efficiency indicators
Device Resource Usage	Memory and computational utilization

The dataset was preprocessed before AI model development. Missing values were handled through statistical imputation, while normalization techniques were applied to ensure consistency among different measurement scales.

The proposed framework consisted of three major layers:

1. Data Acquisition Layer

The first layer involved IoT sensors installed within industrial environments. These sensors continuously collected operational information from machinery and production systems. The collected data were transmitted to local edge computing devices for immediate analysis.

2. Edge Intelligence Layer

The second layer consisted of optimized AI models deployed on edge devices. Instead of transferring all data to cloud servers, AI algorithms processed information locally to generate real-time predictions. This approach reduced latency, bandwidth consumption, and energy requirements.

3. Decision and Control Layer

The final layer involved automated decision-making based on AI predictions. The system generated alerts regarding equipment abnormalities, maintenance requirements, and energy optimization opportunities.

AI Model Development

Three machine learning and deep learning approaches were developed and evaluated:

1. Lightweight Convolutional Neural Network (CNN)

A lightweight CNN architecture was developed for industrial pattern recognition and anomaly detection. The model was optimized by reducing unnecessary parameters and computational operations.

2. Decision Tree-Based Model

A decision tree model was implemented as a low-complexity machine learning approach suitable for resource-limited environments. The model provided a baseline comparison for evaluating computational efficiency.

3. Optimized Deep Learning Model

An optimized deep learning architecture was developed using compression techniques, including parameter reduction and quantization. The purpose was to maintain prediction accuracy while reducing memory requirements and energy consumption.

AI Model Optimization Techniques

1. Model Pruning

Model pruning was applied to remove redundant parameters from neural networks. This reduced computational complexity and improved inference efficiency. Previous studies indicate that pruning can significantly decrease model size while maintaining predictive performance (Cheng et al., 2020).

2. Quantization

Quantization was applied to reduce numerical precision within AI models. This technique decreases memory requirements and improves processing speed, making models more suitable for deployment on edge devices (Li et al., 2021).

3. Knowledge Distillation

Knowledge distillation was used to transfer knowledge from larger AI models to smaller lightweight models. This approach enabled efficient deployment while preserving important predictive capabilities (Gou et al., 2021).

Performance Evaluation Metrics

The developed AI models were evaluated using multiple performance indicators:

1. Prediction Accuracy

Accuracy was measured to determine the effectiveness of AI models in identifying industrial conditions and predicting equipment behavior.

Accuracy=CorrectPredictionsTotalPredictions×100Accuracy = \frac{Correct Predictions}{Total Predictions} \times 100Accuracy=Total Predictions Correct Predictions×100

2. Processing Latency
Latency measured the time required by AI models to generate predictions after receiving industrial sensor data. Lower latency indicated better suitability for real-time applications.

3. Energy Consumption
Energy efficiency was evaluated by measuring power usage during AI inference operations. The objective was to identify models that achieved reliable performance with minimum energy requirements.

4. Memory Utilization
Memory consumption was analyzed to determine whether AI models could operate effectively on resource-limited edge devices.

Parameter	Description
Temperature	Machine operating temperature measurements
Vibration	Equipment vibration intensity
Energy Consumption	Power utilization patterns
Machine Status	Normal and abnormal operational conditions
Processing Time	AI inference response time
Production Output	Operational efficiency indicators
Device Resource Usage	Memory and computational utilization

5. Data Analysis Procedure
The collected industrial dataset was analyzed using statistical and computational approaches. Descriptive statistics were applied to summarize operational characteristics, while comparative analysis was performed to evaluate different AI models.
The performance of conventional and optimized AI models was compared based on:

- Prediction accuracy
- Processing latency
- Energy consumption
- Memory utilization

Statistical comparison was conducted to determine whether optimization techniques significantly improved AI efficiency in industrial edge environments.

Ethical Considerations
Confidentiality of the information was preserved in the research process in the industrial context. Only necessary operational parameters were included in the company-specific information and this data was anonymized. All research was conducted in an ethical manner regarding the responsibility of managing data, protecting privacy, and handling industrial information securely.

The methodology is systematic, providing a method for creating energy-efficient AI models for edge computing settings in industry. The study demonstrated ways in which edge intelligence can aid sustainable real-time industrial applications by leveraging data from the industrial Internet of Things, lightweight AI architectures, and optimization techniques.
The developed methodology allows for the evaluation of the correlation between AI optimization, energy efficiency and industrial performance, giving proof of the practical applicability of the sustainable systems based on AI optimization in resource-limited environments.

Results and Findings
The study compared the performance of various energy-efficient AI models for use in edge computing applications of industrial IoT systems on real data obtained from 120 monitoring devices installed in Lahore, Pakistan. Four major performance indicators were studied: prediction accuracy, processing latency, energy consumption, and memory utilization.
Results showed that optimized edge AI models were more efficient at prediction and computational tasks, without sacrificing

prediction accuracy. The energy efficiency, response time, and real-time processing capabilities of the proposed edge-based models were found to be superior compared to the conventional cloud-based AI models.
The data set comprised information from operation gathered from the industrial monitoring systems such as machine temperature, vibration characteristics, energy consumption, production status, and equipment monitoring and performance indicators. The data set contained various operating conditions, including normal production and abnormal machine actions.

Table 1
Characteristics of Industrial IoT Dataset

Variable	Mean Value	Standard Deviation	Measurement Range
Machine Temperature (°C)	62.4	8.6	40–85
Vibration Level (mm/s)	3.8	1.2	1–7
Energy Consumption (kWh)	18.7	4.5	10–30
Processing Load (%)	64.3	12.8	35–90
Production Efficiency (%)	86.5	7.4	65–98

The dataset demonstrated significant variations in industrial operating conditions, making it suitable for evaluating AI-based prediction and optimization models.

Performance Comparison of AI Models

Three AI approaches were evaluated:

1. Conventional deep learning model
2. Lightweight optimized AI model
3. Edge-based compressed AI model

The comparison was conducted using accuracy, latency, energy consumption, and memory utilization.

Table 2
Performance Comparison of Developed AI Models

Model	Accuracy (%)	Processing Latency (ms)	Energy Consumption (J/inference)	Memory Usage (MB)
Conventional Deep Learning Model	94.2	185	42.6	245
Lightweight AI Model	92.8	112	29.4	138
Optimized Edge AI Model	93.6	109	27.8	96

The findings indicate that the optimized edge AI model achieved comparable prediction accuracy while significantly reducing computational requirements. Although the conventional deep learning model achieved slightly higher accuracy, it required considerably greater energy consumption and memory resources.

Energy efficiency was one of the primary objectives of the study. The optimized edge AI model demonstrated substantial reductions in power requirements compared with conventional AI architectures

Table 3
Energy Efficiency Improvement of Edge AI Model

Model	Energy Consumption (J/inference)	Energy Reduction
Conventional AI Model	42.6	—
Lightweight AI Model	29.4	30.9%
Optimized Edge AI Model	27.8	34.7%

The results demonstrate that AI optimization techniques reduced energy consumption by approximately **34.7%** compared with the conventional model. This improvement resulted from reduced model parameters, optimized computational operations, and efficient inference mechanisms.
These findings support previous research emphasizing that lightweight AI architectures can significantly improve sustainability in edge computing environments by minimizing computational energy requirements (Schwartz et al., 2020; Fanariotis et al., 2023).

Processing Latency Analysis
Real-time industrial applications require rapid decision-making; therefore, processing latency was evaluated for each AI model.

Table 4
Processing Latency Comparison

Model	Average Latency (milliseconds)
Conventional AI Model	185
Lightweight AI Model	112
Optimized Edge AI Model	109

The optimized edge AI model reduced processing latency by approximately **41.1%** compared with the conventional AI model. This improvement indicates that localized AI processing enables faster industrial responses by eliminating unnecessary communication delays associated with cloud-based processing.
The results are consistent with previous studies demonstrating that edge intelligence improves real-time performance by reducing dependence on centralized computing infrastructures (Zhou et al., 2021).

Memory Utilization Analysis
Memory efficiency is an important factor for deploying AI models on resource-constrained industrial devices.

Table 5
Memory Utilization of AI Models

Model	Memory Requirement (MB)	Reduction Compared with Conventional Model
Conventional AI Model	245	—
Lightweight AI Model	138	43.7%
Optimized Edge AI Model	96	60.8%

The optimized edge AI model required approximately **60.8% less memory** compared with the conventional model. This improvement indicates that compression techniques effectively reduced model complexity and improved suitability for deployment on low-power industrial devices.

Predictive Performance Evaluation

The predictive capability of AI models was assessed through classification of industrial equipment conditions, including normal operation and potential failure conditions.

Table 6

Prediction Performance Metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Conventional AI Model	94.2	93.8	94.1	93.9
Lightweight AI Model	92.8	92.4	92.6	92.5
Optimized Edge AI Model	93.6	93.2	93.5	93.3

The optimized edge AI model achieved only a minor reduction in predictive accuracy compared with the conventional model while providing significant improvements in energy efficiency and computational requirements. This demonstrates that optimization techniques successfully maintained model effectiveness while improving deployment feasibility.

Statistical Analysis of Model Performance

A comparative statistical analysis was conducted to examine differences among AI models.

Table 7

Statistical Comparison of AI Models

Performance Indicator	Mean Difference	Significance Level
Energy Consumption	14.8 J reduction	$p < 0.05$
Processing Latency	76 ms reduction	$p < 0.05$
Memory Usage	149 MB reduction	$p < 0.05$
Accuracy Difference	0.6% reduction	$p > 0.05$

The results indicate statistically significant improvements in energy consumption, latency, and memory utilization. However, the reduction in accuracy was statistically insignificant, demonstrating that optimization techniques improved efficiency without substantially affecting predictive capability.

Key Findings Summary

The major findings of the study are summarized as follows:

1. The optimized edge AI model reduced energy consumption by **34.7%** compared with conventional AI models.
2. Processing latency improved by approximately **41.1%**, supporting real-time industrial decision-making.
3. Memory requirements decreased by **60.8%**, making the model suitable for resource-limited IoT devices.
4. Predictive accuracy remained above **93%**, indicating that optimization did not significantly compromise performance.
5. Edge-based AI models demonstrated greater suitability for Lahore-based industrial environments where computational resources and energy efficiency are important considerations.

Overall, the findings confirm that energy-efficient AI optimization strategies can enable sustainable deployment of intelligent industrial IoT systems by balancing accuracy, speed, and resource utilization.

Discussion

The focus of this study was on the development of AI models for energy-efficient edge computing applications in the field of industry 4.0. The results show that optimizing AI models for edge can yield substantial computational efficiency with minimal loss of predictive performance. The optimized model saved 34.7% energy consumption, 41.1% processing latency and 60.8% memory usage over traditional AI models.

The findings suggest that while accuracy is significant in industrial applications, other factors such as energy efficiency, computational needs, and real-time response must also be taken into account when implementing AI. The results align with a growing trend to focus on more sustainable methods of AI that consider the efficiency of resources (Schwartz et al., 2020).

The optimized edge AI model was able to reach 93.6% accuracy, of which is very near to the conventional deep learning model (94.2%) but with much fewer computational resources. This shows that optimization methods like model compression, pruning and quantization can reduce complexity without significantly compromising predictive power.

Previous research has shown that light-weight AI architectures can enhance the feasibility of deploying machine learning models on resource-constrained devices, which can reduce the amount of memory needed and the computational load (Cheng et al., 2020). The results of this research also show that optimized AI models can be used for industrial IoT applications that demand efficient real-time processing.

One aspect that stood out as a benefit of the proposed edge AI framework was the potential for energy efficiency. This decrease of 34.7% in energy consumption clearly indicates that optimized models can contribute to sustainable industrial automation by lowering the computational power needed.

The results follow the Green AI principles of developing efficient algorithms that demonstrate a balance of energy efficiency and performance (Schwartz et al., 2020). In the energy and infrastructure industries, including many manufacturing facilities in Lahore, AI solutions can lower energy consumption and promote the adoption of new technologies.

The latency decrease for processing highlighted the edge AI's potential in time-sensitive industrial applications. Edge-based models process data locally, minimizing reliance on cloud communication and facilitating rapid decision-making.

The results align with earlier studies on the benefits of edge intelligence for responsiveness, scalability, and reliability of industrial IoT systems (Zhou et al., 2021). Machine fault detection, predictive maintenance, and automated quality monitoring are among the applications that can benefit greatly from the decreased response time.

These findings have significant implications for Lahore's industrial sector such as textile, pharmaceutical, engineering, and manufacturing industries, among others. Limited computational infrastructure, high technology investment and high energy cost are some of the challenges for many industries in the region. The proposed edge AI framework offers an economical solution for intelligent automation without undue reliance on the cloud.

The adoption of lightweight AI models on local edge devices can help industries become more productive, cost-effective, and sustainable. It is an excellent method for small and medium-sized businesses (SMBs) looking for a more progressive digital transformation.

In this study, they provided empirical evidence for deploying edge AI in a developing industrial context, which adds to the existing body of literature. This research uses real data from Industrial IoT applications in Lahore to test the performance of AI rather than simulated data, as has been done in many other studies.

The results show the possibility of optimizing AI models for accuracy, energy efficiency, and computational demands. Moving forward, there is a need for further research on cutting-edge technologies like federated learning, adaptive AI models, and hybrid edge-cloud solutions to enhance industrial intelligence and sustainability.

References

Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2020). Model compression and acceleration for deep neural networks: A survey. *IEEE Signal Processing Magazine.*

Schwartz, R., Dodge, J., Smith, N., & Etzioni, O. (2020). Green AI. *Communications of the ACM, 63*(12), 54–63.

Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2021). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE.*

Tuli, S., Mirhakimi, F., Pallewatta, S., Casale, G., & Buyya, R. (2022). AI augmented edge and fog computing: Trends and challenges.

Conclusion and Recommendations

The present study developed and evaluated energy-efficient artificial intelligence models for edge computing-based real-time industrial IoT applications in low-power and resource-constrained environments. The increasing complexity of industrial data and the limitations of cloud-based AI systems highlight the need for localized, efficient, and sustainable computing solutions. The findings demonstrate that optimized edge AI models can effectively address challenges related to energy consumption, processing latency, and hardware limitations.

The results revealed that the proposed optimized edge AI model achieved **93.6% predictive accuracy** while reducing energy consumption by **34.7%**, processing latency by **41.1%**, and memory utilization by **60.8%** compared with conventional AI models. These improvements indicate that AI optimization techniques, including model compression, pruning, and quantization, can significantly enhance the feasibility of deploying intelligent systems on resource-limited industrial devices.

The study further highlights the importance of edge intelligence for industrial sectors seeking real-time monitoring, predictive maintenance, and automated decision-making capabilities. By processing industrial data closer to the source, edge AI reduces dependence on cloud infrastructures, minimizes communication delays, and supports sustainable digital transformation.

In the context of Lahore-based industries, the proposed framework provides a practical pathway for adopting AI-driven automation despite limitations related to energy availability, computational resources, and technology costs. The findings contribute to the growing field of sustainable AI by demonstrating that industrial intelligence can be achieved through efficient and resource-aware AI architectures.

This study makes several important contributions:

1. **Development of an energy-efficient edge AI framework:** The research proposes an optimized AI architecture suitable for real-time industrial IoT applications under resource constraints.
2. **Performance-based evaluation:** The study evaluates AI models using multiple indicators, including accuracy, latency, energy consumption, and memory utilization.
3. **Industrial context contribution:** The research provides empirical evidence from Lahore-based industrial environments, addressing the limited availability of studies from developing economies.
4. **Sustainability contribution:** The findings demonstrate how AI optimization techniques can support environmentally sustainable industrial automation.

Based on the findings, the following recommendations are proposed:

1. Adoption of Lightweight AI Models

Industries should prioritize lightweight and optimized AI architectures instead of highly complex models requiring extensive computational resources. Such models provide efficient performance while reducing energy requirements.

2. Integration of Edge-Based Processing

Manufacturing organizations should implement edge computing solutions for applications requiring rapid responses, including predictive maintenance, machine monitoring, and automated quality control.

3. Implementation of AI Optimization Techniques
AI developers should incorporate pruning, quantization, and knowledge distillation during model development to improve computational efficiency and reduce deployment costs.
4. Investment in Industrial IoT Infrastructure
Industrial organizations should gradually enhance sensor networks and edge computing infrastructure to support sustainable digital transformation.
5. Development of Industry-Specific AI Solutions
Future industrial AI systems should be designed according to specific operational requirements rather than adopting generalized models that may require excessive resources.
6.4 Policy and Industrial Implications
The findings provide important implications for industrial policymakers and technology planners. Government and industrial stakeholders should encourage the adoption of energy-efficient AI technologies through digital transformation programs, research collaborations, and technical training initiatives.
For developing economies such as Pakistan, sustainable AI adoption requires affordable technological solutions that consider energy limitations and industrial diversity. Supporting local development of edge AI systems can improve industrial competitiveness, reduce operational costs, and promote smart manufacturing practices.
Future research may extend this study by exploring:

- **Federated learning-based edge AI systems** to improve data privacy and collaborative industrial intelligence.
- **Hybrid edge-cloud architectures** combining local processing efficiency with cloud computational capabilities.
- **Real-time deployment studies** involving larger industrial networks and diverse manufacturing sectors.
- **Advanced energy-aware AI algorithms** for further reducing computational requirements.
- **Integration of renewable energy-powered edge devices** for sustainable industrial automation.

Final Conclusion
Energy-efficient artificial intelligence models represent a promising solution for enabling sustainable and intelligent industrial systems. This study demonstrates that optimized edge AI architectures can provide high predictive performance while significantly reducing energy consumption and computational requirements. The proposed framework offers a practical approach for Lahore-based industries and other resource-constrained environments to adopt AI-driven automation, improve operational efficiency, and advance toward sustainable Industry 4.0 transformation.

REFERENCES
Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2020). A survey of model compression and acceleration for deep neural networks. *IEEE Signal Processing Magazine, 35*(1), 126–136.
Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S., & Zomaya, A. Y. (2022). Edge intelligence: The confluence of edge computing and artificial intelligence. *IEEE Internet Computing, 26*(1), 7–12.
Fanariotis, A., Orphanoudakis, T., Kotrotsios, K., Fotopoulos, V., Keramidas, G., & Karkazis, P. (2023). Power efficient machine learning models deployment on edge IoT devices. *Sensors, 23*(3), 1595.
Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision, 129*, 1789–1819.
Iftikhar, S., Gill, S. S., Song, C., Xu, M., Aslanpour, M. S., Toosi, A. N., & Buyya, R. (2022). AI-based fog and edge computing: A systematic review, taxonomy and future directions. *Internet of Things, 18*, 100487.

Li, G., Qian, C., & Han, S. (2021). Efficient deep learning for edge computing: Model compression and optimization approaches. *IEEE Access, 9*, 127100–127115.

Liang, F., Yu, W., Lu, X., Griffith, D., & Golmie, N. (2020). Towards edge-based deep learning in industrial Internet of Things. *IEEE Internet of Things Journal, 7*(5), 4320–4334.

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM, 63*(12), 54–63.

Shi, W., Zhou, H., Zhang, Q., & Li, Y. (2020). Edge computing in industrial Internet of Things: Architecture, challenges, and applications. *IEEE Network, 34*(5), 192–198.

Tuli, S., Mirhakimi, F., Pallewatta, S., Casale, G., & Buyya, R. (2022). AI augmented edge and fog computing: Trends and challenges. *Journal of Systems Architecture, 128*, 102562.

Wang, J., Zhang, Y., Chen, Y., & Li, X. (2021). Artificial intelligence enabled industrial IoT: Applications, challenges, and opportunities. *IEEE Transactions on Industrial Informatics, 17*(8), 5455–5464.

Zhang, C., Lu, Y., & Liu, J. (2022). Efficient deep learning models for edge intelligence applications. *Future Generation Computer Systems, 128*, 1–12.

Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2021). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE, 109*(8), 1228–1256.

Zhu, S., Ota, K., & Dong, M. (2022). Green AI for IIoT: Energy-efficient intelligent edge computing for Industrial Internet of Things. *IEEE Transactions on Green Communications and Networking, 6*(1), 79–88.

Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2020). Model compression and acceleration for deep neural networks: A survey. *IEEE Signal Processing Magazine.*

Gou, J., Yu, B., Maybank, S., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision, 129*, 1789–1819.

Li, G., Qian, C., & Han, S. (2021). Efficient deep learning for edge computing: Model compression and optimization approaches. *IEEE Access.*

Schwartz, R., Dodge, J., Smith, N., & Etzioni, O. (2020). Green AI. *Communications of the ACM, 63*(12), 54–63.

Shi, W., Zhou, H., & Zhang, Q. (2020). Edge computing in industrial IoT: Architecture and applications. *IEEE Network.*

Tuli, S., Mirhakimi, F., Pallewatta, S., Casale, G., & Buyya, R. (2022). AI augmented edge and fog computing: Trends and challenges.

Wang, J., Zhang, Y., & Chen, Y. (2021). Artificial intelligence enabled industrial IoT: Applications and challenges. *IEEE Transactions on Industrial Informatics.*

Zhang, C., Lu, Y., & Liu, J. (2022). Efficient deep learning models for edge intelligence applications. *Future Generation Computer Systems.*

Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2021). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE.*

Fanariotis, A., Orphanoudakis, T., Kotrotsios, K., Fotopoulos, V., Keramidas, G., & Karkazis, P. (2023). Power efficient machine learning models deployment on edge IoT devices. *Sensors, 23*(3), 1595.

Iftikhar, S., Gill, S. S., Song, C., Xu, M., Aslanpour, M. S., Toosi, A. N., et al. (2022). AI-based fog and edge computing: A systematic review, taxonomy and future directions.

Liang, F., Yu, W., Lu, X., Griffith, D., & Golmie, N. (2020). Towards edge-based deep learning in industrial Internet of Things. *IEEE Internet of Things Journal,* 7(5), 4320–4334.

Tuli, S., Mirhakimi, F., Pallewatta, S., Zawad, S., Casale, G., Buyya, R., et al. (2022). AI augmented edge and fog computing: Trends and challenges.

Zhu, S., Ota, K., & Dong, M. (2022). Green AI for IIoT: Energy-efficient intelligent edge computing for Industrial Internet of Things. *IEEE Transactions on Green Communications and Networking,* 6(1), 79–88.