Big Data-Enabled Federated Learning for Secure and Collaborative Industrial IoT in Industry 4.0

Asia Bataineh
Computer Science Department
Faculty of Information Technology
University of Petra (UoP)
Amman, Jordan
asia.bataineh@uop.edu.jo

Hamzah Alqudah
Department of Data Science and Artificial Intelligence
Faculty of Information Technology
American University of Madaba (AUM)
Amman, Jordan
h.qudah@aum.edu.jo

Hazem Bani Abdoh
Computer Science Department
Faculty of Computer Science and Information Technology
Jerash university
Jerash, Jordan
Hazim.baniabdoh@jpu.edu.jo

Fuad Fataftah
Mutlimedia systems and Virtual Reality
School of Computer science
Universiti Sains Malaysia (USM)
Penang, Malaysia
fuadmanna@student.usm.my

***Abstract*—Integrating the Internet of Things (IoT), Artificial Intelligence and big data analytics in Industry 4.0 has revolutionized industrial processes, enabling enhanced operational efficiency, predictive maintenance, and innovation. However, the increasing volume of sensitive and decentralized data generated by Industrial IoT (IIoT) devices introduces significant challenges, including data fragmentation, privacy concerns, and interoperability issues. Traditional centralized data analysis methods often fail to address these challenges effectively. This paper proposes a novel privacy-preserving federated learning framework tailored for IIoT environments to bridge these gaps. The framework enables secure and decentralized distributed big data analysis while ensuring data sovereignty and minimizing communication overhead. The proposed approach enhances predictive maintenance and anomaly detection by integrating advanced deep learning models with edge-fog-cloud architectures, fostering cross-company collaboration and scalability. Experimental evaluations using real-world predictive maintenance datasets demonstrate the framework's effectiveness in achieving high accuracy, optimized resource utilization, and reduced runtime. Additionally, incorporating clustering techniques improves model personalization, enhancing performance without compromising data privacy. This research establishes FL as a transformative solution for secure, collaborative intelligence in Industry 4.0 ecosystems, paving the way for sustainable and intelligent manufacturing environments.**

***Index Terms*—Big data, federated learning, Industry 4.0, IoT, predictive maintenance, privacy preservation**

I. Introduction

Industry 4.0 represents a transformative era in manufacturing and industrial processes, driven by integrating digital technologies such as the Internet of Things (IoT), Artificial Intelligence (AI), and big data analytics [1]. This industrial revolution builds on the foundation of automation and smart systems to enhance efficiency, productivity, and sustainability. At the core of Industry 4.0 is the convergence of physical and digital domains, enabling real-time data collection, advanced analytics, and intelligent decision-making [2]. With interconnected devices and systems, businesses can optimize operations, reduce costs, and deliver higher-quality products and services. This paradigm shift is not just about adopting advanced technologies; it's about reimagining traditional industrial practices to meet the demands of a fast-evolving global market. The IoT is pivotal in realizing Industry 4.0, the backbone of interconnected industrial systems [3]. IoT enables devices, sensors, and machines to communicate seamlessly, creating a cohesive ecosystem that facilitates real-time monitoring, predictive maintenance, and automation. IoT empowers manufacturers to gather vast amounts of data from machinery, production lines, and supply chains in industrial settings [4]. This data provides valuable insights into operational performance, equipment health, and workflow optimization. By bridging the physical and digital worlds gap, IoT enhances operational efficiency and fosters innovation in smart factories, intelligent supply chains, and advanced manufacturing processes.

The immense data generated by IoT devices in Industry 4.0 brings unprecedented innovation opportunities but poses significant challenges [1]. Privacy concerns, data silos, and high computational costs often constrain traditional centralized approaches to data analysis. Big data-enabled federated learning emerges as a transformative solution [5], allowing organizations to collaboratively analyze decentralized datasets without compromising security or privacy. Federated learning leverages edge devices and distributed computing to enable secure, privacy-preserving data sharing and processing across multiple stakeholders [6]. This approach is particularly critical in industrial ecosystems, where sensitive data ownership

979-8-3315-2365-7/25/$31.00 ©2025 IEEE

and regulatory compliance are paramount. Incorporating big data-enabled federated learning into Industry 4.0 frameworks provides a secure and scalable pathway for achieving collaborative intelligence. By enabling multiple stakeholders—such as manufacturers, suppliers, and operators—to share insights derived from distributed datasets, federated learning helps overcome the challenges of data silos and fragmented information. This collaborative model enhances predictive maintenance, anomaly detection, and resource optimization and fosters innovation through cross-company partnerships. Integrating big data and federated learning ensures that industrial IoT systems remain robust, secure, and adaptable, paving the way for future sustainable and intelligent factories.

Figure 1 illustrates the integration of the IoT and the Industrial Internet of Things (IIoT) within the context of Industry 4.0 [7]. It highlights various applications across sectors such as mining, catering, food, textiles, automobiles, and electronics. The figure also showcases the role of IoT in traffic management, personal area networks, sensor networks in mechanics, factory area networks, and intelligent transportation systems. These interconnected systems facilitate more innovative operations, enhanced automation, and real-time monitoring, contributing to transforming industrial processes and everyday services.

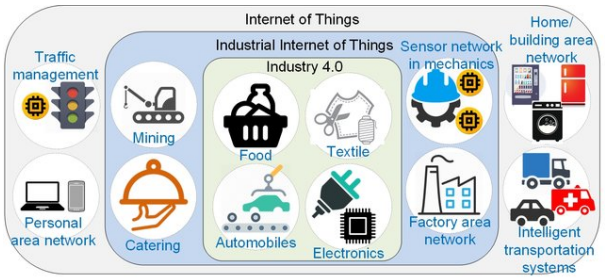

Fig. 1. Applications of IoT and IIoT in Industry 4.0

The vast amount of data generated by industrial IoT devices remains fragmented across stakeholders due to privacy concerns, data ownership issues, and security challenges [8] [9] [10]. This fragmented data landscape limits the potential of collaborative intelligence, which is essential for predictive maintenance, anomaly detection, and resource optimization in smart factories. Big data-enabled federated learning is a promising solution to address these challenges, enabling secure and decentralized data analysis without exposing sensitive information. By facilitating cross-company collaboration and ensuring data privacy, this approach can unlock the true value of IoT data, fostering innovation, improving efficiency, and driving the sustainable growth of Industry 4.0 ecosystems.

Federated Learning addresses the challenges of data privacy, fragmentation, and interoperability in Industry 4.0 by enabling decentralized AI model training across IoT devices without sharing sensitive data. As IoT devices in smart factories, industrial systems, and transportation networks generate vast amounts of big data, FL facilitates collaborative learning by allowing local models to be trained on distributed data and aggregated into a global model. This approach ensures privacy-preserving analysis, fosters cross-company collaboration and optimizes predictive maintenance and fault detection processes. By bridging big data, IoT, and AI, FL enables smart and secure solutions critical to realizing the full potential of Industry 4.0. Figure 2 illustrates the Federated Learning architecture, where IoT devices such as smartphones, mobile computers, and smart vehicles collaboratively train local models on their local data while ensuring privacy and sharing only model updates with a central federated server for global model aggregation and distribution.

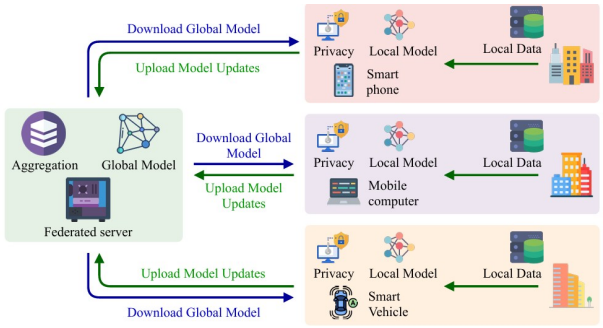

Fig. 2. Architecture of Federated Learning

Figure 3 presents the key research challenges in implementing Federated Learning in Big Data environments. These challenges include system and data heterogeneity, statistical variability, communication bottlenecks, privacy concerns, and potential security risks like membership inference attacks and poisoning issues. Additionally, the figure highlights algorithmic limitations, autonomy-related dilemmas, and efficiency bottlenecks that hinder large-scale adoption of federated learning across diverse and decentralized networks.

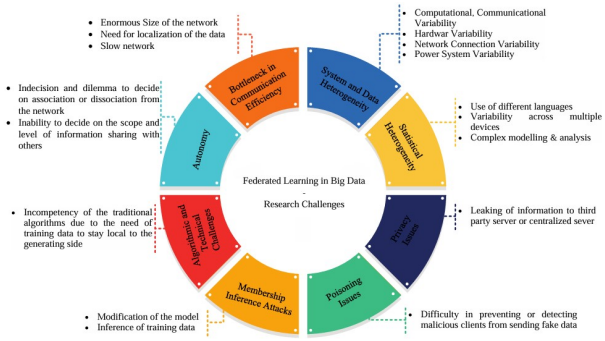

Fig. 3. Research Issues in Big Data and Federated Learning

The main objectives of this paper are as follows:

- To develop a privacy-preserving federated learning framework that enables cross-company collaboration within Industrial IoT systems, ensuring secure and decentralized big data analysis without compromising sensitive information.
- To address the fragmentation of industrial data by creating a federated data space that adheres to principles

such as interoperability, data sovereignty, and security, enabling the sharing and collaborative use of big data across multiple stakeholders in Industry 4.0.
- To evaluate the proposed framework through predictive maintenance and condition monitoring use cases, demonstrating its ability to enhance AI model performance, reduce data silos, and optimize industrial processes using distributed edge-fog-cloud architectures.

The organization of this paper is as follows: Section II presents a comprehensive review of related work, highlighting existing solutions, their limitations, and the research gaps addressed by this paper. Section III describes the proposed privacy-preserving federated learning framework tailored for IIoT environments, integrating advanced deep learning models with edge-fog-cloud architectures to enable secure and scalable data analysis. Section IV details the experimental evaluations and analysis, demonstrating the effectiveness of the proposed framework through predictive maintenance and anomaly detection use cases, along with performance comparisons against state-of-the-art methods. Section V concludes the paper by summarizing the findings, emphasizing the contributions of this research, and outlining future research directions to address challenges in large-scale federated learning deployments.

II. Related Work

FL has emerged as a highly effective solution for developing secure and cost-efficient IIoT applications. By enabling the integration of large datasets and computational resources from diverse IIoT devices, FL facilitates the training of AI models while preserving data privacy. This approach significantly enhances the quality of IIoT training data, which is often unattainable using traditional AI methods.

As described in [11], data is initially generated by IoT devices across various smart industries and transmitted to an IoT sink. The sink serves as a repository that collects data from multiple IoT nodes within the sector via wired and wireless communication channels while encrypting the data before sending it to a centralized server. The server then aggregates the information from multiple IoT sinks and federates the data. Finally, the smart industry decrypts the aggregated knowledge into a comprehensible format. Both eavesdropping and hacking attempts are mitigated in this process, as the data remains encrypted throughout the transmission and storage stages.

Integrating deep learning models with IoT and edge devices has recently gained significant popularity, enabling real-time analytics with limited resources [12]. Federated Deep Learning empowers Industry 4.0 companies to incorporate deep learning into IoT devices while ensuring a secure framework through Federated Learning, as illustrated in Figure 4. The primary objective of FDL is to equip IIoT systems with advanced capabilities using optimized DL models, thereby transforming Industry 4.0 factories into smart, efficient, and intelligent manufacturing environments.

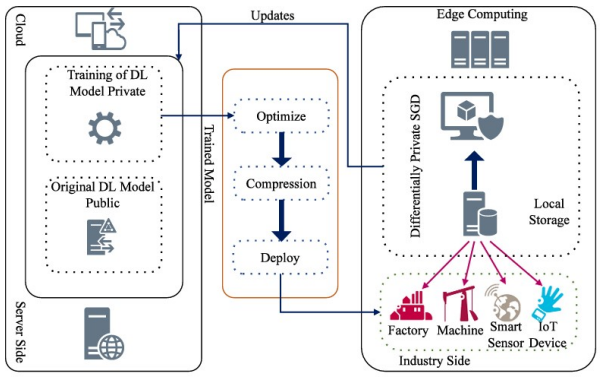

Fig. 4. Fedrtaed Learning in IIoT

Lim et al. [13] highlighted various security threats, which FDL effectively addresses by sharing deep learning models from the cloud to end devices. Security and privacy concerns can be mitigated through data encryption [55]. However, IIoT-based systems still face challenges related to security and privacy during data processing and analytics. The core principle of FDL involves training local deep learning models on localized data and exchanging parameters such as weights and biases of the neural network. These updates are periodically shared between local nodes to generate a global model collaboratively without exposing the underlying data to the cloud. On the server side, security issues arise from sharing DL models on the cloud, posing confidentiality and data security risks. On the client side, data encryption ensures that sensitive information remains secure during training before it is transmitted to the cloud server. Techniques such as Homomorphic Encryption further regulate and protect the amount of data shared, thereby addressing these security concerns effectively.

Given end devices' limited memory and computational capabilities, deep learning models must be optimized to enable efficient deployment on IIoT and edge devices. Optimizing DL models across IIoT nodes reduces memory and computational requirements, enhancing overall system performance. GPUs provide low-power computation for hardware optimisation, significantly reducing processing time, while devices like FPGAs and Google's TPU [14] further accelerate DL network processing. In terms of memory optimization, techniques such as shared memory allocation algorithms can be applied to improve efficiency. Additionally, dynamic scheduling [15] is critical in enhancing performance on cloud servers. Researchers [16] - [19] have recently proposed DL models integrated with Federated Learning for IIoT networks across various applications, including automobiles, mobile networks, traffic systems, and image processing.

III. Proposed Model

To address the challenges of fragmented data, privacy concerns, and the need for scalable AI solutions in IIoT environments, this paper presents a novel privacy-preserving federated learning framework tailored for predictive maintenance and collaborative industrial applications. However, large-scale

deployments of FL introduce critical challenges such as system heterogeneity and communication bottlenecks, which can significantly impact model performance and scalability. System heterogeneity arises due to the diverse hardware and software configurations across industrial IIoT devices, leading to variations in computational power, memory constraints, and data distributions. To mitigate these effects, adaptive federated learning techniques, including model compression, personalized FL, and asynchronous updates, can be employed to tailor model training to each device's capabilities.

The proposed model integrates advanced deep learning techniques with federated learning to enable secure, decentralized analysis of distributed big data across multiple stakeholders. By leveraging edge-fog-cloud computing architectures, the model efficiently extracts meaningful insights while maintaining data sovereignty and minimizing communication costs. This approach ensures that sensitive industrial data remains localized while only model updates are shared for global aggregation. The proposed model enhances predictive accuracy, anomaly detection, and resource optimization in Industry 4.0 ecosystems, fostering cross-company collaboration and driving innovation in smart manufacturing.

The proposed Federated Learning algorithm, depicted in Algorithm 1, addresses the challenge of decentralized data analysis within IIoT environments. As industrial systems generate vast amounts of sensitive data, traditional centralized machine-learning approaches face significant privacy concerns, regulatory restrictions, and high communication costs. To overcome these issues, the algorithm employs a federated learning framework that allows multiple participants (e.g., factories or edge devices) to train a global AI model collaboratively without sharing their raw data. Instead of transferring datasets to a central server, participants train locally on their data and share only model updates (parameters) with the server, ensuring data privacy and ownership.

Initialization and model distribution:
The algorithm begins with an initial global model θ^0, shared by a central server with all participating clients (e.g., industrial IoT devices or factories). Each client holds its local dataset D_i, which remains private throughout the training process. This initialization phase sets the foundation for decentralized training, where the clients will update the global model based on their individual data distributions. This method ensures that sensitive information, such as proprietary machine data or operational metrics, is never exposed.

Local training on participants' devices:
In each training round t, the global model θ^{t-1} is distributed to all participants. Each participant independently trains the model on its local dataset D_i using an optimization algorithm such as Stochastic Gradient Descent (SGD) or Adam. The local model updates θ_i^t are computed based on the loss function, typically designed to minimize prediction errors in failure detection or maintenance tasks. Since each client trains on its local data, this step preserves privacy and mitigates risks associated with centralized data collection, such as breaches or leaks.

Algorithm 1 Federated Learning for Predictive Maintenance

Dataset D split into N participants with privacy-sensitive data.
Central server for model aggregation.
Local datasets D_i for $i = 1, 2, \ldots, N$.
Initial global model θ^0.

Trained global model θ^*.

procedure FEDERATEDLEARNING(θ^0)
  for $t = 1, 2, \ldots, T$ **do** {Iterate for T global rounds}
    Distribute global model θ^{t-1} to all clients.
    for each client i in parallel **do**
      Perform local training on D_i using θ^{t-1}.
      Update local model θ_i^t via local optimization.
    end for
    Aggregate local models $\{\theta_i^t\}$ on the server:

$$\theta^t = \frac{1}{N} \sum_{i=1}^{N} \theta_i^t \tag{1}$$

  end for
  return θ^T
end procedure

procedure EVALUATION(θ^*)
  Use θ^* for predictive maintenance:
  Predict failure probabilities over H historical data points.
  Validate results against failure events over prediction window h.
end procedure

Model aggregation at the server:
Once local training is complete, participants return their model updates (weights and biases) to the central server. The server aggregates these updates to compute an improved global model θ^t. The aggregation process uses a weighted average of all the local model updates:

$$\theta^t = \frac{1}{N} \sum_{i=1}^{N} \theta_i^t \tag{2}$$

Here, N represents the total number of participants, and θ_i^t corresponds to the local model from each client. This step ensures that the global model benefits from the collective intelligence of all participants without accessing their raw data. The iterative nature of this process allows the global model to converge toward an optimal solution over multiple rounds T.

Model evaluation and application:
After T training rounds, the final global model θ^T is obtained. This trained model is evaluated for predictive maintenance tasks like failure prediction and condition monitoring. The model takes historical input data (e.g., sensor readings or machine performance metrics) and predicts potential failures within a specified time window. By leveraging federated learning, the model can generalize well across industrial participants, ensuring accurate and reliable predictions while maintaining data privacy.

The proposed deep learning model, illustrated in Algorithm 2, combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) layers to address the challenges of predictive maintenance in IIoT systems. The model is designed to leverage both the spatial and temporal characteristics of the input data, ensuring robust and accurate failure pre-

diction. CNN layers extract meaningful spatial features from the historical input data, while LSTM layers capture temporal dependencies critical for identifying trends and anomalies. These features are integrated and processed through dense layers to predict the probability of failure, enabling proactive maintenance and reducing operational downtime. The step-by-step methodology of the proposed model is detailed in Algorithm 2.

Algorithm 2 Proposed Deep Learning Model for Failure Prediction

Input data $X \in \mathbb{R}^{H \times 9 \times 1}$ where H is the history window size.

Predicted failure probability for the next time window.

procedure FAILUREPREDICTION(X)
  Step 1: Feature Extraction with CNN
  Initialize convolutional channels $C = \{2, 4, 6, 8\}$.
  for $c \in C$ **do**
    Apply **Conv2D**(X, filters=c).
    Apply **Batch Normalization**.
    Apply **ReLU Activation**.
    Apply **Average Pooling**.
  end for
  Obtain feature maps $F \in \mathbb{R}^{x' \times 8}$.

  Step 2: Sequential Feature Extraction with LSTM
  Pass input X to **LSTM**(20) $\rightarrow$ Output L_1.
  Pass F to **Bidirectional LSTM**(16) $\rightarrow$ Output L_2.

  Step 3: Dense Layer Processing
  Flatten $F \rightarrow F_{flat}$.
  Concatenate L_1, L_2, and $F_{flat} \rightarrow C_{concat}$.
  Pass C_{concat} through:
    Dense(16) with ReLU activation.
    Dense(4) with ReLU activation.
    Dense(1) with Sigmoid activation.

  Step 4: Prediction
  Output: Predicted failure probability Y_{pred}.
end procedure

Feature Extraction with CNN:
The first step of the proposed model focuses on extracting features from the input data $X \in \mathbb{R}^{H \times 9 \times 1}$, where H represents the history window size. Multiple convolutional layers are applied iteratively with increasing channel sizes $C = \{2, 4, 6, 8\}$ to achieve this. For each channel $c \in C$, the following operations are performed sequentially:

1) Apply a **2D Convolution** operation with c filters to extract spatial features.
2) Perform **Batch Normalization** to normalize feature maps and improve training stability.
3) Use the **ReLU activation function** to introduce non-linearity into the network.
4) Apply **Average Pooling** to reduce the dimensionality of feature maps.

These operations are repeated for each channel size, and the resulting feature maps $F \in \mathbb{R}^{x' \times 8}$ are obtained for further processing.

Sequential Feature Extraction with LSTM:
The model incorporates LSTM layers to capture temporal dependencies in the input data. This step processes both the original input X and the extracted CNN features F:

- The input X is passed to an **LSTM** layer with 20 units, producing an output L_1.
- The feature maps F are processed using a **Bidirectional LSTM** layer with 16 units, resulting in an output L_2.

The LSTM layers are designed to analyze the sequential nature of the data, enabling the model to capture temporal patterns critical for predictive maintenance.

Dense Layer Processing:
To integrate the extracted spatial and sequential features, the model flattens the feature maps F into a one-dimensional vector F_{flat}. These features are concatenated with the outputs L_1 and L_2 from the LSTM layers:

$$C_{\text{concat}} = \text{Concatenate}(L_1, L_2, F_{\text{flat}}) \tag{3}$$

The concatenated features C_{concat} are passed through a series of fully connected (Dense) layers to refine the predictions:

1) A **Dense** layer with 16 units and ReLU activation.
2) A **Dense** layer with 4 units and ReLU activation.
3) A final **Dense** layer with 1 unit and Sigmoid activation to produce the failure probability.

Prediction :
The output of the final Dense layer represents the predicted failure probability Y_{pred}, which is given as:

$$Y_{\text{pred}} = \text{Sigmoid}(C_{\text{concat}}) \tag{4}$$

This probability indicates the likelihood of a failure occurring within a specified time window based on the input data.

The proposed deep learning model combines CNN for spatial feature extraction, LSTM layers for sequential analysis, and Dense layers for prediction. By leveraging both spatial and temporal features, the model achieves robust and accurate failure prediction, making it suitable for predictive maintenance tasks in Industrial IoT environments.

IV. Experiments and Analysis

In this section, we assess the effectiveness of the proposed collaborative framework by implementing a novel FL-based predictive maintenance approach. As a key application of CCM, predictive maintenance focuses on forecasting potential future defects or failures in equipment to determine the optimal timing for maintenance activities. The proposed framework is evaluated under two scenarios:

- **Scenario I**: Assumes full participant trust, enabling centralized data sharing for model training.
- **Scenario II**: Focuses on collaborative model training using Federated Learning, where participants retain data privacy due to a lack of mutual trust.

We utilized an open predictive maintenance dataset from one of Schwan's factories **55**. The dataset comprises three distinct sections covering a total of 100 machines. The experiment was conducted on a system with the following computational platform specifications, as detailed in Table I.

TABLE I
HARDWARE SPECIFICATIONS

Component	Specification
CPU	Intel(R) Core i7-10750H @ 2.60 GHz
RAM	32 GB
GPU	NVIDIA GeForce RTX 3060
Operating System	Ubuntu 20.04

The results in Figure 5 demonstrate a clear relationship between the history window size and the model's performance metrics, including accuracy, recall, precision, and the corresponding runtime. Increasing the history window size allows the model to access more historical data, enhancing its ability to learn complex patterns and trends associated with equipment behaviour. This results in significant improvements in predictive performance, as reflected in the higher accuracy and precision values. The model's consistently high recall indicates its effectiveness in detecting potential equipment failures and minimizing false negatives—a critical aspect of predictive maintenance where missed failures can lead to costly operational downtimes or system breakdowns. However, the performance gains achieved through larger history windows come with a notable increase in runtimes. The larger the history window, the greater the volume of input data the model must process, which demands more computational resources and time for training. While smaller history windows offer faster runtimes and lower resource consumption, they result in comparatively lower predictive performance due to the limited amount of historical information available to the model. This trade-off between accuracy and computational efficiency becomes a key consideration for practical deployment in industrial IoT environments, where real-time analysis and resource constraints are often critical.

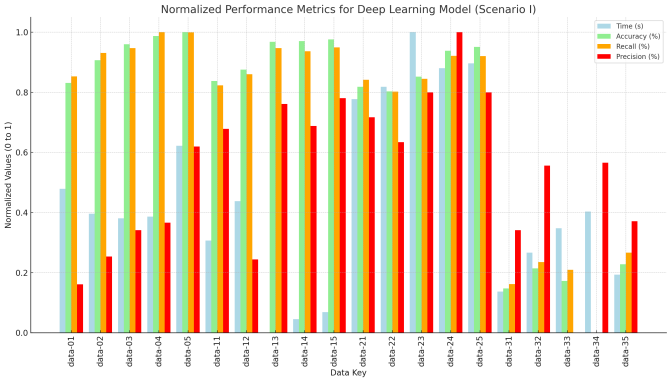

Fig. 5. Results of Scenario 1

The results in Figure 6 demonstrate that in Scenario II, using the Federated Learning method significantly reduces runtime as the number of clients increases. This reduction is achieved by splitting the data among clients, which decreases the computational load for each client, leading to faster training. However, this data distribution slightly decreases accuracy because local models are trained on smaller data portions. While FL provides a scalable framework for distributed learning, its scalability is not without limitations. As the number of participants increases, several challenges arise, including model convergence issues, computational overhead, and trade-offs between privacy and performance.

Increasing the number of clients can lead to slower convergence due to heterogeneous local data distributions (non-IID data), which may introduce inconsistencies in model updates. To mitigate this, advanced aggregation techniques such as personalized FL and clustered FL can be employed, where similar data distributions are grouped to enhance learning effectiveness. Additionally, computational and storage overhead on edge devices can become a limiting factor, particularly for resource-constrained IIoT devices. Techniques such as federated dropout and adaptive learning rates help optimize resource utilization while maintaining model accuracy. Furthermore, with more participants, ensuring privacy while maintaining high model performance becomes challenging. Privacy-enhancing techniques like differential privacy and secure multi-party computation introduce additional computational complexity, requiring a careful balance between security and efficiency. By considering these scalability limitations, FL can be better adapted to large-scale IIoT applications, ensuring robustness and efficiency in industrial deployments.

Despite this, the reduction in accuracy is minimal and remains within acceptable limits, making Federated Learning a viable option when data privacy is a priority. Furthermore, the results show that as the number of clients increases beyond 7 to 19, the rate of change in accuracy becomes shallow. This indicates that the performance stabilizes, even with more clients participating in the training process. This scalability highlights Federated Learning's robustness, as it can efficiently balance privacy, computational efficiency, and model performance. Overall, Federated Learning is an effective solution for collaborative model training, ensuring reduced runtime and acceptable accuracy trade-offs while preserving data confidentiality.

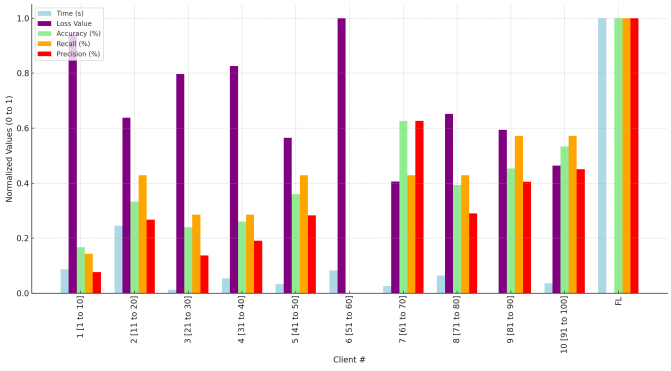

Fig. 6. Results of Scenario 2

The results in Figure 7 represent the scenario where none of the clients participate in the Federated Learning process.

Instead, each client trains their model using only local data. For this analysis, the dataset was split into ten parts, with each client using a distinct subset of the data for training. While the proposed method is flexible and could accommodate a different number of clients, this choice does not affect the general trend or the conclusions drawn from the experiment. As shown in Figure 7, the accuracy of the models trained on local data is lower compared to the global model trained through the FL process. This is primarily because local models lack access to the entire dataset, limiting their ability to generalize and learn comprehensive patterns. Additionally, the variation in performance across clients indicates a dependency on the quality and size of the data available to each client. Some clients may achieve higher accuracy due to better-quality local data, while others may perform poorly if their subsets lack diversity or are insufficiently representative.

Despite these limitations, the results demonstrate that local data training offers a privacy-preserving alternative, as data remains decentralized and never leaves the client's control. However, this approach sacrifices overall model performance in favour of privacy. These findings emphasize the importance of collaborative learning in achieving more robust and accurate models, as Federated Learning enables the aggregation of insights from distributed datasets without compromising privacy.

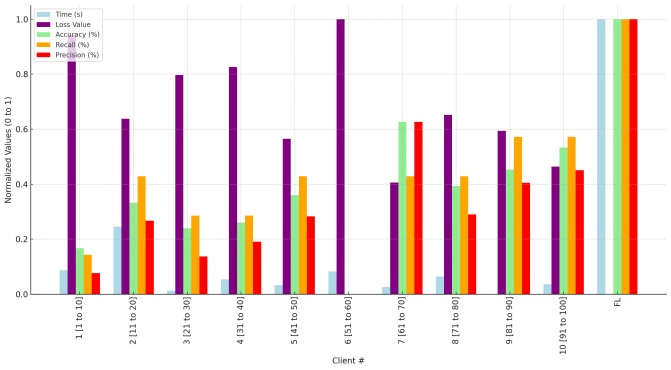

Fig. 7. Results of Clustered Federated Learning with K-Means Clustering for Model Training (No. of Clusters = 1–5

Figure 8 illustrates the impact of the clustering technique and the number of clusters (K) on the performance of the Federated Learning model. The results highlight that incorporating clustering into the FL process allows for model personalization, which leads to improved overall performance compared to traditional FL without clustering. This improvement arises because clustering enables grouping clients with similar data distributions, allowing the model to tailor its training to the specific characteristics of each cluster. As the number of K increases, the model's performance improves due to more precise personalization. By focusing on smaller, more homogeneous subsets of clients, the model can learn more effectively from localized patterns and variations within the data. However, the extent of improvement tends to diminish as K reaches higher values, indicating that there is a balance between achieving better personalization and maintaining sufficient data diversity within each cluster. These findings underscore the effectiveness of clustering techniques in Federated Learning, as they enhance model accuracy and generalization while preserving the benefits of data privacy and decentralization.

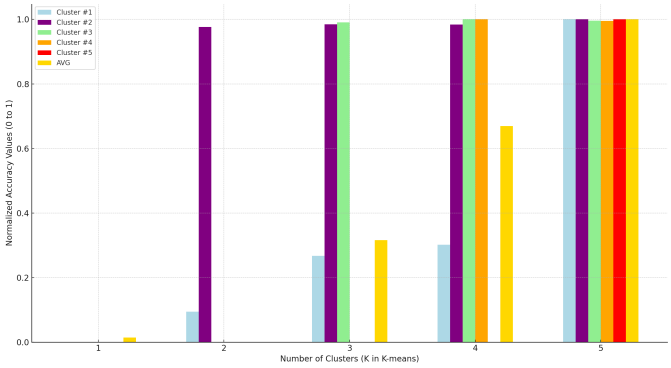

Fig. 8. Results of Federated Learning

V. Conclusion

This paper presents a comprehensive approach to addressing the challenges of data privacy, fragmentation, and interoperability in IIoT environments, particularly in Industry 4.0. By leveraging a privacy-preserving Federated Learning framework, this research demonstrates the feasibility of secure, decentralized data analysis across multiple stakeholders while maintaining data sovereignty and reducing communication overhead. The proposed framework integrates advanced deep learning techniques with edge-fog-cloud architectures, enabling predictive maintenance, anomaly detection, and resource optimization in industrial ecosystems. Experimental results highlight the framework's effectiveness in achieving high predictive accuracy, scalability, and efficiency while ensuring robust data privacy. Using clustering techniques within the FL process further enhances model personalization, enabling localized insights and improved generalization across diverse IIoT environments. This capability underscores FL's potential to foster collaboration and innovation among industrial stakeholders without compromising sensitive data.

Despite its promising results, this study also identifies areas for further research, including addressing challenges related to system heterogeneity, communication bottlenecks, and the computational costs associated with large-scale FL deployments. Future work will focus on refining the framework to enhance its adaptability and robustness, exploring advanced optimization techniques, and incorporating real-time analytics for dynamic IIoT environments. In conclusion, this research establishes Federated Learning as a transformative enabler for secure and collaborative intelligence in Industry 4.0, paving the way for the next generation of smart, sustainable, and efficient industrial systems.

REFERENCES

[1] K. C. Rath, A. Khang, and D. Roy, "The Role of Internet of Things (IoT) Technology in Industry 4.0 Economy," in *Advanced IoT Technologies and Applications in the Industry 4.0 Digital Economy*, CRC Press, pp. 1-28, 2024.

[2] G. Lampropoulos and K. Siakas, "Enhancing and Securing Cyber-Physical Systems and Industry 4.0 Through Digital Twins: A Critical Review," *Journal of Software: Evolution and Process*, vol. 35, no. 7, p. e2494, 2023.

[3] K. Logeswaran, S. Savitha, P. Suresh, K. R. Prasanna Kumar, M. Gunasekar, R. Rajadevi, and A. S. Jayasurya, "Unifying Technologies in Industry 4.0: Harnessing the Synergy of Internet of Things, Big Data, Augmented Reality/Virtual Reality, and Blockchain Technologies," in *Topics in Artificial Intelligence Applied to Industry 4.0*, pp. 127-147, 2024.

[4] J. Vijay Arputharaj, B. N. J. William, A. A. Haruna, and D. D. Prasad, "Exploring the Synergy of IIoT, AI, and Data Analytics in Industry 6.0," in *Industry 6.0: Technology, Practices, Challenges, and Applications*, vol. 1, 2024.

[5] W. Wang, O. Abbasi, H. Yanikomeroglu, C. Liang, L. Tang, and Q. Chen, "A Vertical Heterogeneous Network (VHetNet)–Enabled Asynchronous Federated Learning-Based Anomaly Detection Framework for Ubiquitous IoT," *IEEE Open Journal of the Communications Society*, 2023.

[6] Y. Qi, Y. Feng, X. Wang, H. Li, and J. Tian, "Leveraging Federated Learning and Edge Computing for Recommendation Systems within Cloud Computing Networks," *arXiv preprint*, arXiv:2403.03165, 2024.

[7] P. Boobalan, S. P. Ramu, Q. V. Pham, K. Dev, S. Pandya, P. K. R. Maddikunta, and T. Huynh-The, "Fusion of Federated Learning and Industrial Internet of Things: A Survey," *Computer Networks*, vol. 212, p. 109048, 2022.

[8] D. Anand, I. Mavromatis, P. Carnelli, and A. Khan, "A Federated Learning-Enabled Smart Street Light Monitoring Application: Benefits and Future Challenges," in *Proceedings of the 1st ACM Workshop on AI Empowered Mobile and Wireless Sensing*, pp. 7-12, October 2022.

[9] S. H. Alsamhi, R. Myrzashova, A. Hawbani, S. Kumar, S. Srivastava, L. Zhao, and E. Curry, "Federated Learning Meets Blockchain in Decentralized Data-Sharing: Healthcare Use Case," *IEEE Internet of Things Journal*, 2024.

[10] W. Dai, H. Nishi, V. Vyatkin, V. Huang, Y. Shi, and X. Guan, "Industrial Edge Computing: Enabling Embedded Intelligence," *IEEE Industrial Electronics Magazine*, vol. 13, no. 4, pp. 48-56, 2019.

[11] L. Kong, X.-Y. Liu, H. Sheng, P. Zeng, and G. Chen, "Federated Tensor Mining for Secure Industrial Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 2144–2153, 2019.

[12] T. Huynh-The, C.-H. Hua, Q.-V. Pham, and D.-S. Kim, "MCNet: An Efficient CNN Architecture for Robust Automatic Modulation Classification," *IEEE Communications Letters*, vol. 24, no. 4, pp. 811–815, 2020.

[13] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated Learning in Mobile Edge Networks: A Comprehensive Survey," *IEEE Communications Surveys Tutorials*, vol. 22, no. 3, pp. 2031–2063, 2020.

[14] X. Zhang, X. Chen, J. K. Liu, and Y. Xiang, "DeepPAR and DeepDPA: Privacy Preserving and Asynchronous Deep Learning for Industrial IoT," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 2081–2090, 2019.

[15] H.-D. Cho, P. D. P. Engineer, K. Chung, and T. Kim, "Benefits of the Big.LITTLE Architecture," 2012.

[16] X. Zhang, X. Chen, J. K. Liu, and Y. Xiang, "DeepPAR and DeepDPA: Privacy Preserving and Asynchronous Deep Learning for Industrial IoT," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 2081–2090, 2019.

[17] X. Wang, C. Wang, X. Li, V. C. Leung, and T. Taleb, "Federated Deep Reinforcement Learning for Internet of Things with Decentralized Cooperative Edge Caching," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9441–9455, 2020.

[18] Y. Liu, J. James, J. Kang, D. Niyato, and S. Zhang, "Privacy-Preserving Traffic Flow Prediction: A Federated Learning Approach," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7751–7763, 2020.

[19] Y. Chen, X. Sun, and Y. Jin, "Communication-Efficient Federated Deep Learning with Layerwise Asynchronous Model Update and Temporally Weighted Aggregation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 10, pp. 4229–4238, 2019.